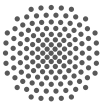


Universität Stuttgart

The Enterprise Data Marketplace: A Platform for Democratizing Company Data

Von der Fakultät für Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von

Rebecca Kay Eichler

aus Waiblingen

Hauptberichter: Prof. Dr.-Ing. habil. Bernhard Mitschang

Mitberichter: Prof. Dr. Willem-Jan van den Heuvel

Tag der mündlichen Prüfung: 14. Juni 2024

Institut für Parallele und Verteilte Systeme
Abteilung Anwendersoftware

2024

Acknowledgements

This dissertation was authored at the Institute for Parallel and Distributed Systems within the Department Application of Parallel and Distributed Systems, at the University of Stuttgart as part of a joint research initiative with an industrial collaborator.

I would like to take this opportunity to thank all the people who have guided and supported me throughout the process of completing this thesis. A special thanks goes to my doctoral supervisor Prof. Dr. Bernhard Mitschang, who made this work possible through his continuous support. I thank him for the ongoing discussions over the years, the scientific guidance, and for initiating my research project. I also thank Prof. Dr. Willem-Jan van den Heuvel for his contribution as co-examiner. My sincere appreciation also goes to my supervisors Holger Schwarz, Christoph Gröger, Eva Hoos, and Alexander Röck. Their insights both in scientific and practical regards contributed to this thesis significantly. Besides numerous fruitful discussions, their personal encouragement and tips have been instrumental in shaping my PhD experience, and for that, I also extend my deepest appreciation.

I thank all my colleagues at the IPVS and fellow researchers for the insightful discussions both in scientific as well as personal matters. A special thanks goes out to Corinna Giebler, who encouraged me to take the leap and give this PhD experience a try. I also thank Christoph Stach, for the tips on writing papers and booking conference hotels with the most glittering reviews, Dennis Treder-Tschechlov, for the discussions on productivity techniques or generally

the mysteries of the universe, Julius Voggesberger, for leveling us up in prickly pranksterhood, Jan Schneider, for teaching me the art of writing e-mails and having my back when teaching the wonders of Data Warehousing, and Christian Weber for being the best office buddy. To Mathias Mormul, Daniel Del Gaudio, Manuel Fritz, and the whole Mensa crew I express my gratitude for ultimately making the lunch hour relaxing through riveting conversations about lizards and candy. I thank Michael Behringer for increasing productivity across several departments by dutifully policing the coffee machine. My gratitude also goes out to Stefan Driessen for the fun collaboration, heated discussions on data products, and for enlightening me that German is a Dutch dialect.

I would also like to thank my colleagues at the industrial partner, including but not limited to Ingo Hollenbeck, Melanie Dold, Bernd Johannes, Matthias Dod, Rebecca Kalter, and Susanne Bitzer. Furthermore, for her support in organizational matters, I thank Eva Strähle, as well as my students including but not limited to Laura Schuiki, Ganna Berezhna, Marvin Fischer, and Franz Sebastian Müller, whose work contributed to the content of this dissertation.

My heartfelt gratitude goes to my parents Krista and Lorenz who have supported me throughout all of the stages of my life. To Gabi and Helmut, my brother Gabriel, and Johanna, I express my utmost appreciation for their unwavering support and understanding throughout my PhD journey. For their patience and relentlessness in upholding my mental happiness, I also thank my friends, even though there was some confusion about whether or not I'm an impoverished student who has yet to start real work. Lastly, but most profoundly, I thank my husband, Hendrik, for being my rock and my greatest cheerleader. Thank you for being my partner in every sense of the word, for celebrating my successes, and for being a source of strength and inspiration. I am endlessly grateful to have you by my side.

Contents

Abstract

In the era of big data, multitudes of data are generated and collected in companies. This data contains a potential value that can be leveraged to gain new insights, e.g., for enhancing business models or reengineering industrial products. Extracting data value requires that this data is available for use. Yet, studies show that significant amounts of the data remain unused in companies. In this regard, data democratization initiatives, with the goal of empowering company employees to find, understand, access, use, and share data, are gaining in importance.

Towards this end, data marketplaces are studied as metadata-based platforms to facilitate the exchange and provisioning of data and data-related services. However, data marketplaces are mainly investigated for the exchange of data and services between organizations or private individuals, i.e., as external data marketplaces. Little research focuses on the use of data marketplaces in the company-internal context, i.e., as an Enterprise Data Marketplace (EDMP). Topics of how the EDMP differs from an external marketplace, the scope of its offerings and functionality, challenges that arise in the company-internal context, or how the EDMP can be embedded in and leverage the existent company system and tool landscape have not been investigated in detail thus far.

In this thesis, the Enterprise Data Marketplace is examined as a platform for democratizing data in companies, and in this context, the above-listed gaps are addressed. To this end, four research goals (RGs) are put forward: (RG1) the identification of the processes and challenges company employees

face in finding, understanding, accessing, and sharing data in the enterprise without an EDMP; (RG2) the identification of the distinctive aspects of an EDMP; (RG3) establishing an architectural foundation for building an EDMP; and (RG4) the goal of leveraging existent metadata in the company tool and system landscape in the EDMP. The research goals are covered through nine research contributions. These entail the current data provider and data consumer journeys and challenges, an EDMP type distinction based on an EDMP definition, as well as a presentation of its distinctive characteristics, requirements, and challenges. An enterprise integration and platform architecture, together with an approach for leveraging existent metadata, yields the foundation for building an EDMP.

The feasibility of the concepts put forward in this thesis is demonstrated through an EDMP prototype and an evaluation based on an experiment and qualitative assessment. The evaluation yields that the EDMP is well-suited for the effective realization of data democratization within companies and that it not only addresses several of the current issues data providers and consumers face but also increases the efficiency and reduces the complexity in accessing data. This thesis therefore introduces the EDMP as a platform for democratizing company data and lays the foundation for establishing the EDMP within a company.

German Abstract

Im Zeitalter von Big Data entstehen in Unternehmen große Mengen von Daten. Diese Daten bieten das Potential neue Erkenntnisse zu gewinnen, um beispielsweise Geschäftsmodelle oder Industrieprodukte zu optimieren. Die Ausschöpfung dieses Potentials setzt voraus, dass Daten für die Nutzung verfügbar sind. Studien zeigen jedoch, dass erhebliche Mengen an Daten in Unternehmen ungenutzt bleiben. In diesem Zusammenhang gewinnen Initiativen zur Datendemokratisierung an Bedeutung. Diese zielen darauf ab, Mitarbeiter in die Lage zu versetzen Daten zu finden, zu verstehen, auf sie zuzugreifen, sie zu nutzen und mit anderen zu teilen.

In diesem Sinne werden Datenmarktplätze, metadatenbasierte Plattformen, die den Austausch und die Bereitstellung von Daten und datenbezogenen Diensten unterstützen, untersucht. Der Fokus liegt bisher auf Datenmarktplätze für den Austausch von Daten und Dienstleistungen zwischen Organisationen oder Privatpersonen, d.h. auf externen Datenmarktplätzen. Der Einsatz von Datenmarktplätzen im unternehmensinternen Kontext ist dagegen wenig erforscht. Es ist unklar wie sich unternehmensinterne und externe Datenmarktplätze unterscheiden. Gleichzeitig sind Angebotsumfang und Funktionalität, Herausforderungen im unternehmensinternen Kontext, sowie Möglichkeiten zur Integration in die bestehende System- und Toollandschaft eines Unternehmens nicht im Detail bekannt.

In dieser Arbeit wird der Datenmarktplatz als Plattform für die Demokratisierung von Daten innerhalb eines Unternehmens erforscht und die genannten

Herausforderungen adressiert. Zu diesem Zweck werden vier Forschungsziele formuliert: (RG1) die Identifikation der Prozesse und Herausforderungen für Mitarbeiter, wenn sie Daten ohne einen unternehmensinternen Datenmarktplatz finden, verstehen, darauf zugreifen und gemeinsam nutzen möchten; (RG2) die Ermittlung der charakteristischen Merkmale eines internen Datenmarktplatzes; (RG3) die Skizzierung einer Architektur als Grundlage für den Aufbau eines internen Datenmarktplatzes; und (RG4) das Ziel, vorhandene Metadaten in der Tool- und Systemlandschaft des Unternehmens für den internen Datenmarktplatz nutzbar zu machen. Die Forschungsziele werden durch neun Forschungsbeiträge abgedeckt. Diese beinhalten die aktuellen Workflows und Herausforderungen von Datenanbietern und -konsumenten, eine Definition des internen Datenmarktplatzes sowie die Darstellung seiner besonderen Merkmale, Anforderungen und Herausforderungen. Dazu kommt eine Architektur für die Marktplatz-Plattform, deren Unternehmensintegration sowie ein Ansatz zur optimalen Nutzung vorhandener Metadaten. Zusammengenommen schaffen die Forschungsbeiträge damit die Grundlage für das Umsetzen eines internen Datenmarktplatzes.

Die Validierung der in dieser Arbeit vorgestellten Konzepte erfolgt anhand eines Prototyps für einen internen Datenmarktplatz. Die Eignung dieser Konzepte wird auf Basis eines Nutzer-Experiments und einer qualitativen Bewertung nachgewiesen. Die Evaluation zeigt, dass der interne Datenmarktplatz für die effektive Umsetzung der Datendemokratisierung in Unternehmen geeignet ist. Der Marktplatz adressiert dabei nicht nur mehrere der Probleme der Datenanbieter und -konsumenten, sondern steigert auch die Effizienz im Datenzugriff und reduziert die Komplexität in diesem Prozess. Diese Arbeit stellt damit den unternehmensinternen Datenmarktplatz als Plattform für die Datendemokratisierung vor und legt den Grundstein für dessen Etablierung im Unternehmen.

Chapter 1

Introduction

In this dissertation, the Enterprise Data Marketplace (EDMP) is introduced as a platform for democratizing company data. Therein, the extent to which an EDMP enables company employees with varying skill-sets to find, understand, access, use, and share data within a company is examined. As data marketplaces have mainly been studied for the exchange of data between organizations, the novelty lies in the usage of a data marketplace in a company-internal context. In this regard, this work identifies the EDMP as a distinctive marketplace type, examines how it deviates from traditionally used data marketplaces, and exhibits how to establish an EDMP within a company.

This chapter serves as an introduction to this dissertation. After providing the motivation to this dissertation in Section 1.1, the central research question, the research goals, and according contributions are presented in Section 1.2. Finally, Section 1.3 outlines the structure of the dissertation.

1.1 Motivation

In this day and age, an enormous amount of data is generated by, for instance, the Internet of Things (IoT), social media networks, transactional processing systems, or wearables and mobile devices [Cao17]. This data has become a

strategic asset for companies and plays an essential role in shaping the development of a range of sectors, from industry over healthcare to research [Cao17]. The importance of the data stems from its inherent value. Data may, for example, hold the potential to gain new insights which may lead to the discovery of new business models or the expansion into new markets. This data value can, however, only be extracted if the data is available for use. Studies show that over half of the data goes unused within companies [Sea20; spl19]. For this reason, data democratization initiatives are growing in importance.

Data democratization has the objective to empower and motivate the majority of company employees to find, understand, access, use, and share data within the company, in consideration of data security and compliance [LLF21; AG20]. This involves the enablement of broader access to data and tools, the development of data-related and analytics skills, collaborative knowledge-sharing, as well as the promotion of data value [LLF21]. Towards this end, tools such as data catalogs are being studied to support data democratization in companies [LLEF20]. These maintain a metadata-based inventory of datasets and enable data consumers to find and understand these for the goal of extracting business value [ZDED17; SML+23]. Related to data catalogs, are so-called data marketplaces (DMPs). These are metadata-based platforms for trading data as well as data-related services [MS19; LSV18; Grö21]. The data marketplace provides infrastructure for the data exchange by acting as a digital intermediary connecting data providers and data consumers [MS19]. Through data marketplaces, data becomes available, which is in turn the basis for extracting data value.

However, data marketplaces have mainly been studied for the exchange of data and services between organizations or private individuals, i.e., the external data marketplace (Ext-DMP). In the company-internal context, the data marketplace is referred to as an Enterprise Data Marketplace [Grö21; Wel17] or an internal data marketplace [FSF20]. Research areas surrounding data marketplaces include the consideration of data as a good, the determination of the value and quality of data, how data becomes a product [HL23], conceptual analysis of different aspects of data marketplaces like business models, participants or marketplace classifications, and the prominent topic of pricing data [LSV18]. Very little research focuses on the EDMP. Amongst others,

Gröger [Grö21] highlights the need for this specific marketplace type, Fernandez et al. [FSF20] consider them to bring down data silos, and Wells [Wel17] defines and presents the EDMP in a report. The questions of how the EDMP differs from an external data marketplace, whether it provides the same offerings and functionality, and what challenges arise when employing a data marketplace within a company, have not been sufficiently addressed. Moreover, how the EDMP integrates with a company's IT system landscape and thus reuses existent infrastructure, has not been examined. This also includes the topic of how the EDMP as a metadata-based platform can leverage existent metadata spread across a company's IT system landscape. Without insights into these topics, it remains challenging for companies to utilize and maximize the value of a data marketplace within the organization so that it supports their goal of democratizing the company data.

1.2 Research Question, Goals, and Contributions

In order to address the above-mentioned challenges, we investigate the following research question (RQ) in the scope of this thesis:

> *How does the Enterprise Data Marketplace support data democratization in terms of enabling employees with varying skill-sets to find, understand, access, use, and share data in a secure and compliant way within a company?*

A set of four research goals (RGs) are derived, on the basis of which the above-mentioned research question will be addressed. Each research goal has an associated set of contributions (Cs).

RG1 –Identification of the processes and challenges that company employees face in finding, understanding, accessing, and sharing data within their enterprise.

In order to examine the extent to which a data marketplace drives data democratization, it is necessary to first identify the current state without the use of a data marketplace. In this regard, the current processes for both data consumers and data providers must be identified. From these, the consumer and

provider challenges can be derived. This provides the basis to identify whether the EDMP addresses the current challenges, improving the data consumer and provider processes, and thereby drives the democratization of data. Although enabling the consumer to use data is part of data democratization, it is not considered in the scope of this research goal as a data marketplace is a platform for sharing yet not the further processing, i.e., usage of data. For the purpose of achieving the goal RG1, this thesis offers the following contributions:

C1.1 Identification of the data provider journey for sharing data and the data consumer journey for finding, understanding, and accessing data.

C1.2 Identification of challenges throughout the data provider and data consumer journey.

RG2 –Identification of the distinctive aspects of an Enterprise Data Marketplace.

The utilization of an EDMP, referring to both its implementation and operation, requires an understanding of which aspects need to be taken into account in the internal context. Before the EDMP can be used effectively, it is necessary to clarify how data marketplaces differ inside and outside a company context. This involves clarifying whether the EDMP is utilized for the same objective as an external data marketplace. Based on whether the objective is the same it may provide different offerings. If the offerings vary, the provided functionality may be affected. Based on the objective, offerings, and functionality the EDMP may also have a distinct set of characteristics. Lastly, these characteristics may elicit specific challenges in a company-internal context that need to be resolved to operate a marketplace effectively. In this regard, this thesis provides three contributions that in conjunction yield the basis for a uniform understanding of the EDMP:

C2.1 Establishing a type distinction by providing an EDMP definition and identifying the EDMP specific characteristics.

C2.2 Identification of EDMP requirements.

C2.3 Identification of EDMP specific challenges.

RG3 –Establishing an architectural foundation for building an Enterprise Data Marketplace.

Merely gaining knowledge about the idiosyncrasies of an EDMP does not yet suffice to realize such a marketplace. Based on insights about the distinctive aspects of an EDMP, resulting from research goal RG2, an architectural foundation for building an EDMP can be derived. This entails how the marketplace platform will integrate and interact with a company's existent tool and system landscape. Depending on the level of desired integration, the components can be derived which are required within the marketplace and which may be reused from other tools and systems. Based thereon, an architecture for an EDMP can be designed, yielding the foundation for building this type of marketplace tailored for use within a company. While an assortment of marketplace architectures have been presented throughout literature, none are explicitly tailored to reflect the EDMP specific characteristics and for building a marketplace that tightly integrates with a company's existent infrastructure. To this end, the following two contributions are presented:

C3.1 Derivation of an enterprise integration architecture for the EDMP.

C3.2 Designing of an EDMP platform architecture.

RG4 –Leveraging existent Metadata in the Enterprise Data Marketplace.

The EDMP is a metadata-based platform [Grö21]. Metadata is required for various purposes in the data marketplace ranging from the indexing of datasets so the consumer can discover existent data, over descriptions of the dataset, so the consumers can understand the dataset, to governance aspects like the documentation of access rights, so consumers only gain access to datasets in a legal and compliant way. In order to share their data the data providers have to supply this metadata for their datasets. A lot of metadata is already maintained within a variety of tools in companies [EGG+21a], in, for instance, data catalogs, as inventories of datasets, business glossaries for managing business term definitions [HES17], or data quality platforms for measuring, maintaining and improving data quality [JF20]. As indicated through the contribution C3.1, an EDMP as opposed to an external data marketplace can tightly integrate

with this existent infrastructure. Regarding metadata, this means it can build on and reuse the existent metadata. This not only reduces redundancy but also alleviates the data consumers and data providers in their workflows. For instance, the consumer does not have to access a variety of tools to find the metadata required to understand a dataset and the data provider does not have to maintain metadata on a dataset within several tools. Leveraging this metadata in the EDMP entails a three-part approach: First, the EDMP has to integrate with the existent tools according to the enterprise integration architecture as outlined in RG3. Secondly, it has to be able to support a set of different metadata per dataset, depending on the dataset itself as well as what metadata is available in the company. Thirdly, it has to be able to visualize this diverse metadata per dataset in an integrated view. Literature does not provide an in-depth discussion on how an EDMP can be integrated in a company tool and system landscape, and consequently also does not provide insights into how the metadata therein can be leveraged within an EDMP. Therefore, this thesis offers the following contribution towards leveraging existent metadata:

C4.1 An approach for integrating with different tools.

C4.2 An approach for supporting diverse metadata per dataset and visualizing this metadata in an integrated view.

1.3 Outline of this Thesis

This thesis is structured as follows:

Chapter 2 – Background: This chapter covers the fundamentals required for understanding the topics discussed throughout the rest of this work. This entails basics on data marketplaces as well as data democratization, paired with insights into the topic of metadata management as data marketplaces are metadata-based platforms.

Chapter 3 – The Data Consumer and Data Provider Perspective - Processes and Challenges: In order to gain insights into the current state of data democratization according to research goal RG1, a use case of an industrial manufacturer is presented in this chapter. Based on this use case and on a literature study the data provider and data consumer processes for sharing

and attaining data are derived and discussed, yielding the research contribution C1.1. These give rise to a number of challenges in both the process for the data provider and data consumer, yielding the research contribution C1.2. The following chapters consider how these processes can be improved and how the challenges are addressed through an EDMP platform.

Chapter 4 – Introducing the EDMP: In alignment with research goal RG2, this chapter introduces the EDMP as a distinct type of data marketplace for use within the company-internal context. Besides providing a definition and the objective of the EDMP, the EDMP is placed in a data marketplace classification framework, highlighting the EDMP's characteristics, as required per research contribution C2.1. Furthermore, requirements concerning the EDMP's offerings, functionality, and enterprise integration are listed, providing research contribution C2.2. Lastly, a set of challenges is identified based on the idiosyncrasies of this marketplace type in accordance with research contribution C2.3. This chapter thereby provides a comprehensive understanding of the EDMP and lays the foundation for building this marketplace type, as is discussed in the following chapter.

Chapter 5 – Enterprise Integration and Platform Architecture: Towards the goal of implementing an EDMP, this chapter establishes the foundation therefore, by discussing how an EDMP can integrate with a company's tool and system landscape, and provides an according platform architecture that supports this integration. Additionally, this chapter introduces the differentiation between data assets and data products, as this enables the EDMP to handle a wider scope of data within the enterprise. Thereby, research goal RG3 is addressed with both the contributions C3.1 and C3.2 concerning the EDMP's enterprise integration and platform architectures.

Chapter 6 – Leveraging Distributed Metadata in the Enterprise Data Marketplace: Having identified how the EDMP should integrate with the existing company tools and systems in the previous chapter, this chapter focuses on how the metadata contained therein can be leveraged within the marketplace to support both the data consumers and data providers. In this context, an approach is presented that enables an EDMP to connect to different tools and systems, supports diverse sets of metadata per dataset based on metadata templates, and enables displaying an integrated view on these diverse

metadata sets in the marketplace frontend. This approach yields research contributions C4.1 and C4.2, thus, addressing research goal RG4.

Chapter 7 – Prototypical Implementation and Evaluation: Having introduced a variety of concepts in this thesis, this chapter describes their conceptual and physical implementation through a prototype. Furthermore, the significance of introducing an EDMP is evaluated based on an experiment in which the EDMP's impact on the efficiency, effectiveness, and complexity of the data consumer process is highlighted. Lastly, the extent to which the EDMP addresses the data consumer and provider challenges as presented in Chapter 3, and the extent to which it supports data democratization is assessed.

Chapter 8 – Conclusion and Future Work: This chapter concludes this thesis and provides an outlook on future work.

Chapter 2
Background

In this chapter, we introduce the fundamental concepts on which this thesis is based. To begin with, the basics of data marketplaces are presented in Section 2.1. As we examine data marketplaces in the context of data democratization, we provide a brief data democratization definition in Section 2.2. Lastly, data marketplaces are metadata-based platforms [Grö21] and we discuss the handling of metadata within these, therefore we provide insights into the topic of metadata management in Section 2.3.

2.1 The Data Marketplace

After providing a brief definition of data marketplaces, this section covers roles in the data exchange context, different types of data marketplaces, and how the data marketplace relates to similar research fields like the data mesh, in Sections 2.1.1 to 2.1.3. As data marketplace functionality is examined as part of research contribution C2.2, information on this topic is supplied in the course of this thesis in Chapter 4.

Data marketplaces are electronic, metadata-based self-service platforms for trading data and providing data-related services [MS19; LSV18; Grö21]. A marketplace provides infrastructure for the data exchange by acting as a

digital intermediary connecting market participants and promoting their interactions [Spi19]. It therefore composes a central pivotal component in a data ecosystem [Grö21]. The utilization of data marketplaces may yield several advantageous outcomes. For instance, they stimulate innovation as consumers can acquire data that would have been unavailable, and available data can initiate the improvement of products, services, processes, and the development of new business models [JCZ12].

2.1.1 Roles in the Data Exchange Context

Several roles are involved in the process of exchanging data. To clarify the semantics of these roles, we highlight the main participants on both the data provisioning and data consumption side as well as roles surrounding the marketplace platform. Figure 2.1 depicts an overview of the roles.

Roles in the Provisioning Context: Initially, data is created. This may be done by a person, an application, or something else like sensors. The creating instance is referred to as the *data producer*. Having created the data, a *data owner* is assigned who is accountable for the data, e.g., that it fulfills legal or technical requirements [Ott11; HES17]. The data owners are often line-of-business executives and can delegate the responsibility for realizing data management tasks to so called *data stewards*, who have a detailed knowledge about the business and data requirements [Ott11; HES17; ASvB19]. To this point, the data has been created by the producer, the owner is accountable for it and the steward is responsible for maintaining it. The person who is tasked with making the data available is called the *data provider* [Spi19]. The roles are not mutually exclusive, meaning, a data provider may, for instance, also be the data producer, data steward, or data owner [DMV22].

Roles in the Consumption Context: On the consumption side, the *data consumers* gain access to and use the data that has been made available by the data provider [DMV22]. The data consumers are not limited to people, departments, or whole organizations, but can, for instance, also be applications [Wel18].

Roles in the Data Marketplace Platform Context: Driessen, Monsieur, and Van Den Heuvel [DMV22] identify and summarize a set of additional roles in their literature review on data markets. These include the *platform*

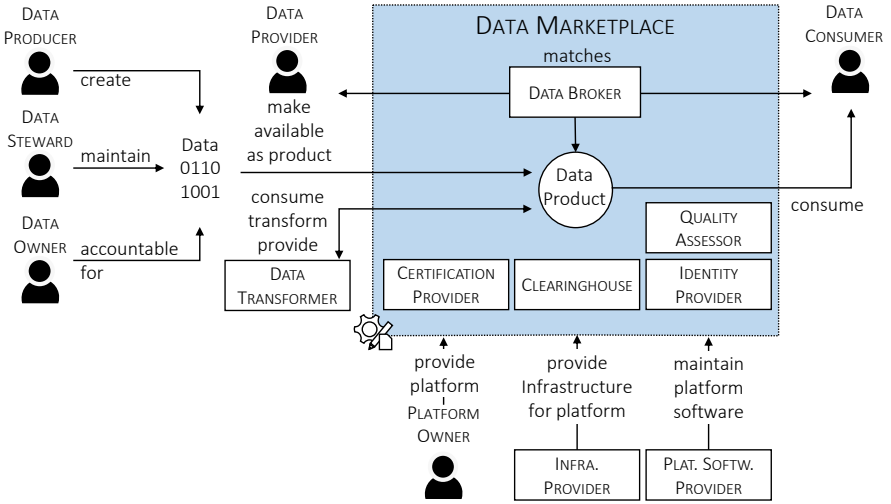


Figure 2.1: Roles in the Data Marketplace.

owner who provides the market platform [SSVV16; MSLV13], the *data broker* matching the providers and consumers based on their needs [AOA17; RRK18], the *clearinghouse* in authority of validating and finalizing transactions [Gan23; OSTL+19], the *identity provider* delivering identity information [WLYH19; CCB19] and the *certification provider* in extension also delivering certifications to actors, e.g., to enable trustworthiness or compliance to quality standards [RD18; SLS+18]. This set of roles also entails the *data transformer* responsible for further processing of the data like cleaning or aggregating it or creating new data products [FSF20; AOA17]. Furthermore, Driessen, Monsieur, and Van Den Heuvel [DMV22] identify the *infrastructure provider* who assists, e.g., the data broker and transformer with required infrastructure [FSS20], the *quality assessor* evaluating data quality based on a set of guidelines [HL19; DMvdHT21], and lastly, the *platform software provider* who designs and maintains the software for operating the data market, which is challenging in decentralized marketplace environments where actors have to agree on the software used to exchange data [JYJS20].

2.1.2 Data Marketplace Types

Data marketplaces can be classified into different types based on a set of criteria. These criteria may, for instance, include the user groups [LSV18], e.g., private individuals or organizations, platform architectures [KLT17], e.g., central or decentral, or the platform ownership [SSVV17], e.g., dependent, independent or consortium based. Data marketplaces are also differentiated based on their application domain, which refers to the data content [DMV22]. For instance, data marketplaces are considered for exchanging personal data [Ish20], data of the automotive and mobility sector [GWF17; PWZ+17] as well as industrial sector [RD18; WZJT21], smart city data [RRK18; Bar18], healthcare data [AS19; AN20], and governmental data [AOA17].

In this sense, a variety of data marketplace types have been defined throughout literature. For instance, Driessen, Monsieur, and Van Den Heuvel [DMV22] identify the following five types of markets in their literature review. *The generalist* offers heterogeneous data across multiple domains and can also be used within a single large company in contrast to *the specialist*, which focuses on homogenized data from a single domain. *The industry data exchange* in which the participants are (large) companies, often from the same domain and with a lot of data, contrasts *the enabler*, which has a large number of individual data providers, often with little data and is typically used in the Internet of Things (IoT) domain. In the fifth type, *the aggregator*, the platform provider acquires a lot of data, transforms and aggregates it, and then proceeds to offer this data to a variety of other data consumers.

Similarly, Azcoitia and Laoutaris [AL22] differentiate four types of data marketplaces in their marketplace survey. These are *general-purpose* and *niche* data marketplaces, wherein niche marketplaces target specific industries and address specific innovative purposes like supplying machine learning algorithms with data. The third type, the *embedded* data marketplace, is integrated in data-management systems or other digital solutions, and lastly, the *personal information management systems* constitutes the fourth type, through which individuals can manage and sell their data. This indicates that marketplace types can be identified on the basis of different criteria.

This thesis focuses on the distinction between *internal* and *external marketplaces*. This differentiation refers to the use of a marketplace for the exchange

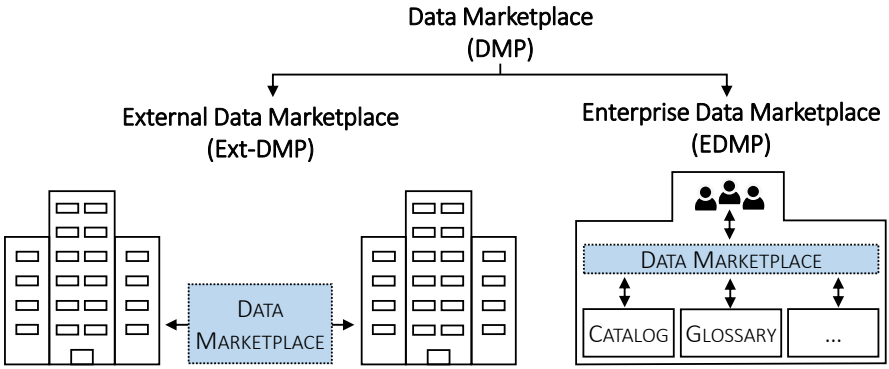


Figure 2.2: Relevant Data Marketplace Types in this Thesis (Based on [EGG+21a].)

of data within an organization or between organizations [FSF20]. The internal marketplace is also referred to as an *Enterprise Data Marketplace (EDMP)* [Grö21]. How the internal and external marketplace types differ and what the distinctive features of the internal type are have not been studied in detail in literature thus far, and are amongst other topics subject in this thesis. Figure 2.2 depicts the topology of the terms data marketplace (DMP), external data marketplace (Ext-DMP), and Enterprise Data Marketplace (EDMP).

2.1.3 Relation and Differentiation to Associated Research Areas

In the context of data exchanges, there are a few concepts and research areas that overlap with the data marketplace. In this section, we aim to illustrate the relation of the data marketplace to these topics to ensure a clear differentiation and understanding of what the data marketplace platform encompasses and what concepts go beyond it. Namely, these topics involve data spaces, the data mesh, and data fabric.

Data Mesh: As explained by the data mesh founder Dehghani [Deh22], the data mesh is a new organizational paradigm for handling analytical data. It is a decentralized approach for managing, sharing, and accessing analytical data at scale, mainly inside but also beyond enterprises. It operates on four principles: firstly, domain ownership, i.e., an organization is divided into

domains that are responsible for managing and sharing their data. Secondly, data-as-a-product, in which data is packaged and made available based on certain requirements. The third principle indicates that there is a self-service platform that offers services to manage a data product throughout its lifecycle, from creation to consumption. The fourth, federated computational governance, should enable both federated decision-making and accountability and interoperability throughout the mesh. In this organizational paradigm, we see the data marketplace as a component of the self-service platform, offering services such as product registration in the mesh, discovery, and access.

Data Fabric: Gartner defines a data fabric as "an integrated layer of connected data" [Gar22]. Like the data mesh, the data fabric is an approach to facilitate managing, sharing, and accessing of corporate data across a hybrid data landscape, supporting both centralized and decentralized systems [Mez23]. As opposed to the data mesh, which is an organizational approach, e.g., by dividing an organization into domains or establishing product-thinking, the data fabric is technology-driven, e.g., by combining interchangeable tools and technologies to enable the above-mentioned goals in an automated, use case agnostic fashion. Its three building blocks entail data governance, data integration, and self-service [Mez23]. These can amongst others be respectively based on metadata, a knowledge graph that relates data across a distributed system, and tools like a data marketplace. The data fabric and data mesh approaches can be combined as the data fabric can be implemented in a way that it meets the data mesh's principles [Mez23; HWW23]. In both, the topic of self-service is addressed through some kind of data platform, in which a data marketplace may be a component.

Data Space: 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 [OSTL+19]. 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. The data space also entails several roles such as data providers, data consumers, broker service providers, clearing house, an app store provider, or certification body. In the data space context, data marketplaces can constitute data providers that

contribute data into the data space community [OJS+16]. 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 [OSTL+19; OJS+16]. Yet, as opposed to the data marketplaces, the broker service provider is not involved in the exchange of data [OSTL+19]. Furthermore, there may be several of these brokers in a data space, whereas the data marketplace as a broker would constitute one central component.

2.2 Data Democratization

In this work, we explore the data marketplace as a potential approach for facilitating data democratization within an enterprise. Therefore, we briefly illuminate what data democratization encompasses. 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 [LLF21; AG20]. Lefebvre et al. [LLF21] identify five dimensions that in conjunction enable data democratization:

Dim 1 –Data Access:
The first dimension describes the enablement of broader access to data for users with varying skill-sets, i.e., also non-specialist users.

Dim 2 –Tool Access:
Similarly, the second dimension entails the enablement of broader access to self-service analytics tools, again in consideration of users with varying skill-sets.

Dim 3 –Data Skills:
The third dimension signifies the development of data and analytic skills, in order for employees to become data-literate and be able to manipulate and analyze data, e.g., starting with skills such as data cleaning.

Dim 4 –Knowledge-sharing:
The fourth dimension covers collective empowerment through collaborative knowledge-sharing between employees.

Dim 5 –Data Value:
The fifth dimension entails the promotion of data value like communicating the importance of data assets.

In the course of this thesis, we discuss whether and to what extent a data marketplace addresses these five data democratization dimensions.

The data democratization initiative is related to the so-called FAIR principles for managing data [LLEF20]. The FAIR principles intend for data, algorithms, tools, and workflows related to data, to become findable, accessible, interoperable, and reusable in both human-driven and machine-driven activities [WDA+16]. While the FAIR principles are rooted in an academic perspective, data democratization can be viewed as the interpretation of the FAIR principles in an enterprise context [LLEF20]. In this sense, Labadie et al. [LLEF20] have highlighted differences between the academic environment and the enterprise context such as the motivations of people to actually access and use data, which may be given in the academic context, yet requires additional incentivation in the enterprise. Based on the data democratization definition of Lefebvre, Legner, and Fadler [LLF21], the FAIR principles are reflected through the democratization inititative's intention to make data, findable, understandable, accessible, and usable. In this sense, the FAIR principles constitute a central aspect in the democratization of data in an enterprise.

2.3 Metadata Management

The data marketplace is a metadata-driven tool [Grö21], which is why this thesis focuses on how a data marketplace handles metadata and can integrate with the existing metadata management structures within an enterprise. For this reason, this section provides a brief definition of metadata, insights into how it is managed, an excerpt of tools and systems that store and provide metadata, and a brief introduction to metadata exchange standards in Sections 2.3.1 to 2.3.4 respectively.

2.3.1 Metadata

One of the most common definitions of metadata states that metadata is data about data [Fur20]. It provides contextual information, yet this is not only limited to data, metadata can also provide insights into processes, data rules and constraints, as well as systems and workflows [HES17]. For example, a dataset's storage location, its size, information on its data owner or producer, or a description of its contents all constitute metadata. Metadata is also used for modeling and meta-modeling, and by following the Meta Object Facility (MOF) international standard for defining metamodels, can be used to, e.g., enable interoperability between metadata-driven systems [Obj19]. Metadata enables processing, maintaining, securing, auditing, and governing data, and generally contributes to managing organizational knowledge about data [HES17]. Without metadata, an organization might not know what data it has, what this data represents, who has accessed it and so on.

Like the marketplaces, metadata can be categorized into types based on certain criteria. A common type distinction based on the metadata content involves the division into *business, technical* and *operational* metadata [HES17]. Business metadata describes the data's content, e.g., a description that this data contains customer demographics. As the term implies, technical metadata provides technical information, e.g., on the format, structure, and the database. Lastly, operational metadata covers information related to the processing and accessing of data, such as the data owner or access rules. There are also other categorizations, for instance, by Gröger and Hoos [GH19], yet, as these are not of explicit relevance for this work, we will not go into further detail.

2.3.2 Metadata Management - Definition and Tasks

As metadata is also data, it has to be managed like data, which is referred to as metadata management [HES17]. In our work [EGG+21b], we reason that metadata management is a type of data management and thus, provide the following definition: metadata management is data management for metadata.

Based on this definition, there are different variants of data management depending on the type of data, e.g., metadata, master data, and transactional data. The basic set of data management tasks, however, remains the same for all

data types, including metadata. These are depicted in Figure 2.3, and involve the three main blocks: data governance, lifecycle management, and foundational activities [HES17]. The blocks and contained tasks are mainly based on DAMA's data management function framework [HES17]. Data governance presents the basis for data management as it involves the planning, monitoring, and enforcement of both the data lifecycle management and foundational activities. This includes creating policies that specify what has to be done and standards that specify how to do things [HES17]. Lifecycle management involves all processes related to the design, creation, obtaining, storing, using, maintaining, enhancing as well as archiving and deleting of data [HES17; YW10]. The foundational tasks are performed throughout all of the above-described lifecycle steps and include the realization of security, privacy, compliance, and data quality management.

Metadata management and data management are generally viewed as two separate activities as there is a dependency between metadata management and the other data management types, e.g., master data management. In order to conduct data management, metadata is required. For example, retaining data privacy and security involves introducing and enforcing access rights, which are essentially metadata, as these describe who is allowed to access the data. Each data management task may require the collection of metadata, for instance, the "Usage" task may contain the tracking of data lineage, data access, and more. Hence, in order to manage data, metadata is required throughout the data management tasks and this metadata in turn is managed based on the same set of tasks.

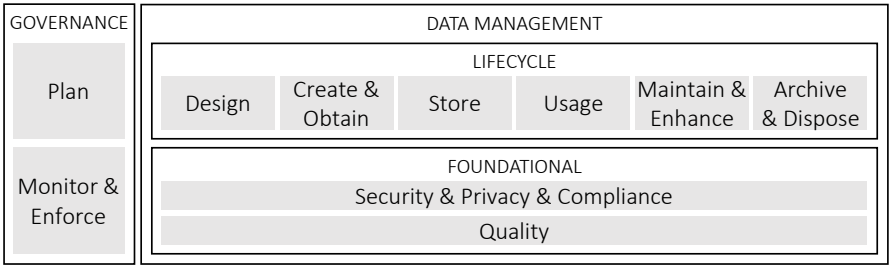


Figure 2.3: Data Management Activities (Based on [HES17], Adjusted from [EGG+21b]).

In our work on enterprise-wide metadata management [EGG+21a], we highlight current goals, challenges, and research gaps in metadata management, amongst which we expose the use of Enterprise Data Marketplaces (EDMPs) as metadata-driven exchange platforms within an enterprise as a research gap. An in-depth discussion of all these aspects would however exceed the scope of this thesis and will therefore not be discussed in detail.

2.3.3 Metadata Sources

Metadata is generated and collected through a multitude of tools and systems within companies. This encompasses both data and metadata management tools which have capabilities for managing metadata. Depending on the tools' and systems' focus, a wide variety of metadata is collected on different topics. Amongst others, these include tools like business glossaries, data dictionaries, data catalogs, data quality tools, business intelligence tools, data integration tools, event messaging tools, ETL tools, modeling tools and repositories, service registries or application metadata repositories of, e.g., access rights management tools [HES17]. In addition, systems that collect metadata encompass analytical systems like data warehouses and data lakes. This list of tools and systems highlights the variety of metadata sources. In the following, we briefly present a few of these, namely the data catalog, business glossary, data quality tool, and the data lake in more detail, as these will be used repeatedly as exemplary metadata sources throughout this thesis.

Data Catalog: A data catalog, such as Alation[1], is a tool for maintaining a data inventory with documentation, i.e., metadata on the registered datasets, and amongst others, offers discovery, administration, and governance functionality [LLEF20; ZDED17]. As the catalog provides an interface for an enterprise-wide search of data and the documentation ranges over different types of metadata, from business metadata such as the content descriptions, over technical metadata, to operational metadata, it also fosters enterprise-wide metadata management [JSR22]. Through its search, documentation, and other functionality, the catalog facilitates the findability and understandability of data within an enterprise.

[1]Alation: https://www.alation.com

Business Glossary: A business glossary specifies business terms and their definitions together with term relations for all business relevant concepts, such as a customer-to-product relation [HES17]. This constitutes business metadata. A business glossary tool, such as Erwin Data Literacy[2], can be used and embedded into the overall application landscape, so the terms do not merely serve as documentation but can be reused in other applications, such as an enterprise-knowledge graph or data models.

Data Quality Tool: Generally speaking, these tools aim to improve and maintain data quality by identifying, understanding, and correlating flaws in data [JF20]. To do this, they offer functionality such as profiling, parsing, standardization, and cleansing [JF20]. Talend[3], for instance, offers this type of data quality functionality. A dataset's quality can be expressed through so-called data quality dimensions like completeness, timeliness, validity, or consistency [Seb13], which constitute metadata.

Data Lake: The data lake is a data management platform designed to incorporate data at scale, of varying structure, from heterogeneous sources, both in its raw and pre-processed formats [GGH+20a; GGH+20b]. It is the goal of the data lake to enable all kinds of analytics, ranging from reporting, over Online Analytical Processing (OLAP), to advanced analytics, to exploit the data's value [GGH+20a; Mat17]. In order to support the handling of a dataset in various processing degrees, a data lake can be divided into so-called zones [GGH+20b]. These define the processing degree such as a raw or cleansed state, as well as how the data is governed concerning topics like access rights or data quality. Metadata is required to manage the data in a data lake, for instance, to track the data's movement through the zones [EGG+21b; GGH+21]. In this regard, metadata models have been designed to facilitate metadata management for data lakes, such as MEDAL [SSF+19] or our metadata model HANDLE [EGG+21b].

Based on these descriptions, it is clear that a wide variety of metadata is collected in enterprises throughout the scope of these various tools and systems.

[2]Erwin Data Literacy: https://www.erwin.com/products/erwin-data-literacy
[3]Talend: https://talend.com/products/data-quality

2.3.4 Metadata Standards

As discussed in the previous section, metadata is created and collected on all kinds of different topics in a variety of tools and systems. To facilitate consistency in the way metadata is collected, standards are required. A metadata standard enables understandability, so people or machines can consume metadata, and also facilitates interoperability between tools and systems that support the same standards [BG00]. Often metadata standards are designed for a specific domain, such as the Data Catalog Vocabulary (DCAT)[4] [ABC+20]. It is a metadata vocabulary that enables interoperability between data catalogs. For example, it defines classes such as the dcat:Catalog, dcat:Dataset, and dcat:Distribution, which in turn have a set of recommended properties like dct:Title, dct:Description, or dcat:ContactPoint. Often standards can be combined to establish a more comprehensive understanding of objects, for instance, DCAT can be complemented with the provenance ontology PROV-O[5] to integrate provenance information in data catalogs [ABC+20]. An understanding of what a metadata standard constitutes is of value throughout this thesis, as a data marketplace could also support or build upon such metadata standards as a metadata-based platform [Grö21].

2.4 Summary

Within this chapter, we established a fundamental understanding of topics discussed throughout the rest of this work. This entailed insights into data marketplaces, concretely the roles in the data exchange context, marketplace type distinctions like the differentiation of internal and external data marketplaces, as well as a delineation of the data marketplace to related research areas. Furthermore, five dimensions of data democratization and definitions of metadata and metadata management were presented. Lastly, sources of metadata within companies like data catalog and business glossary tools, and data lakes were introduced, followed by a brief introduction to metadata standards.

[4] DCAT: https://w3.org/TR/vocab-dcat-2
[5] PROV-O: https://w3.org/TR/prov-o

Chapter 3

The Data Consumer and Data Provider Perspective – Processes and Challenges

To address the research question of how employees are enabled to find, understand, access, use, and share data in a company, we must first examine the current processes and challenges that employees are faced with (RG 1). In this chapter, we therefore examine the processes and challenges for both the data providers and data consumers. We specifically focus on how the data becomes available within an enterprise, i.e., the first data democratization dimension of enabling broader access to data.

The research contributions provided in this chapter cover the data provider journey for publishing and provisioning data and the data consumer journey for finding, understanding, and gaining access to data (C1.1), as well as the provider and consumer challenges (C1.2).

In order to incorporate a practical perspective in addition to insights derived from literature, we interviewed a globally active industrial manufacturer, whose case we present in Section 3.1. Thereafter, the data provider perspective

is examined in Section 3.2, complemented by the data consumer perspective in Section 3.3. This chapter is a revised and composite version of the author publications [EGH+22a] and [EGHS22].

3.1 An Industrial Use Case

The use case described in this section is based on a globally active industrial manufacturer and will serve as the basis for continuous examples throughout the rest of this thesis. The manufacturer is engaged in various sectors, such as the industrial and consumer goods sector, and operates a global manufacturing network. It is one of their main goals to become data-driven and transition into an Industry 4.0 company, whereby they aim to improve their processes and increase their competitive advantages. To this end, the manufacturer is focusing on the adoption of sensors in production and end products, on collecting customer and product-related data from social networks, integrating data from different systems, as well as launching data analytics projects, covering both traditional reporting and advanced analytics. In this regard, the company collects user-generated data, IoT data, and web content in addition to the enterprise data like master data or transactional data. In order to drive more innovative utilization of this data and leverage more data value, the manufacturer has set out to establish an environment in which data can be shared freely and efficiently within the company and has thus launched a data democratization initiative.

The data in question is stored and managed through a diverse system and tool landscape, as depicted in Figure 3.1. The manufacturer employs a variety of operational systems from different vendors, ranging from Enterprise Resource Planning (ERP) systems over Product Lifecycle Management (PLM) and Customer Relationship Management (CRM) systems to Manufacturing Execution System (MES). In addition to the operational systems, the manufacturer also operates multiple instances of analytical systems, such as data warehouses and data lakes. Additionally, there are a number of tools that support the administration and management of this data throughout its lifecycle. These include ETL tools, which support extract, transform, and load-processes, tools for maintaining data quality, business glossaries to clarify semantics across the

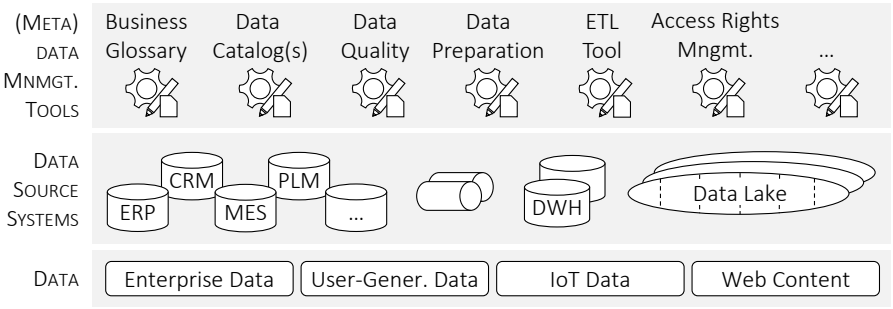


Figure 3.1: A Heterogeneous Enterprise Data, System and Tool Landscape.

systems, and multiple data catalogs for inventorying the accumulated data and maintaining basic metadata. In extension to data catalogs, the manufacturer is investigating data marketplaces. Since such a heterogeneous system and data landscape is typical for large industrial companies this use case provides a representative view on industrial enterprises [Grö18; KJAJ23].

In the following, this representative case will be used to gain practical insights into the current state and challenges in data democratization in larger companies with heterogeneous data, system, and tool landscapes.

3.2 The Data Provider Perspective

As explained in Section 2.2, data democratization aims at empowering employees to find, access, use, and share data [LLF21]. Finding, accessing, and using data are activities of data consumers. Yet, before these activities can be carried out, data providers must first make their data available by publishing it and providing provisioning options so it can then be shared within and potentially beyond the enterprise. In this section, we therefore begin by investigating how data is currently made available by data providers and what challenges they face throughout this process. In a more general sense, we thereby investigate the current state and challenges data providers face in data democratization, specifically within the first democratization dimension on enabling broader access to data within the enterprise.

In order to identify the data provider's process for making data available within an enterprise, we conducted a literature study, amongst others including the works [Grö21; FSF20; GH19; ZG18; LLEF20]. We found that many articles focus on the consumer perspective and not the provider perspective, or only give abstract insight into the provider's processes. To gain a more detailed and practical perspective, we conducted expert interviews with employees of the industrial manufacturer in our use case introduced in Section 3.1. The exchange with experts from various key data-related roles in an industrial enterprise, including enterprise and solution architects, data scientists, and business analysts, gives us a representative view of current processes for making data available in industrial enterprises from different perspectives.

Based on the conducted literature study and expert interviews, we merged the data provider's processes for making data available within industrial enterprises into an overarching data provider journey, which is presented in Section 3.2.1. The journey presents the required steps, the parties that are involved as well as the tools that are used throughout these steps. It also yields the foundation for deriving challenges the data providers face within the scope of the first data democratization dimension of providing broader access to data. The challenges are discussed in Section 3.2.2.

3.2.1 The Data Provider Journey

The data provider's journey of making data available in the company, as illustrated in Figure 3.2, consists of three processes: firstly, *documenting* the data so the data is understandable, secondly, *publishing* the data within the company so it can be found, and thirdly, *preparing the provisioning* of the data, i.e. making it available to consumers so these can access and work with it. These three processes contain a set of steps that are carried out by different roles in the company, namely the data provider, the data owner, IT or operations, legal experts, and management.

To illustrate both the provider journey as well as the consumer journey, presented in the following section, we demonstrate both workflows through an exemplary scenario. In this scenario a data steward employed at our industrial manufacturer wants to provide machine sensor data from running production lines pro-actively. It is their goal to support machine maintenance use cases,

specifically predictive maintenance [AFSC23] of manufacturing machines. For this, machine sensor data on, e.g., temperatures, humidity, vibrations, pressure, and torque are of relevance. Some of this sensor data is stored in a database for up to 15 years for machine warranty cases.

Part 1 - Document: To begin with, the data provider *assembles the data*. In the second step, the according *documentation is assembled* so other employees can understand the data. This is essentially metadata on various aspects of the data such as descriptions of the content, the data's quality or lineage, the underlying data model, technical descriptions, and lifecycle specifications. Ba-

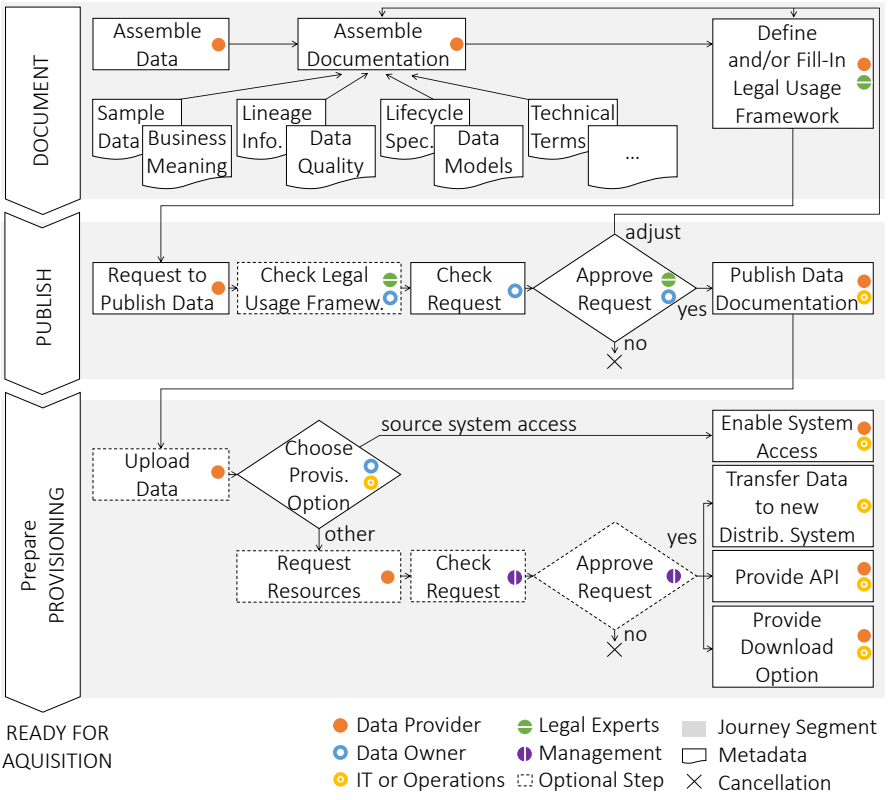


Figure 3.2: The Steps and Parties Involved for Publishing and Preparing the Provisioning of Data Within an Enterprise (Based on [EGHS22].)

sically, this constitutes all information a data consumer will require to evaluate whether the data fits their need and work with the data. In our example, this metadata could be descriptions of the machines that provide the sensor data, their semantics, e.g., machine temperatures or torque, and lifecycle information indicating that this sensor data is stored up to 15 years. These metadata can be maintained through a variety of tools such as business glossaries, data quality tools, or data catalogs.

Thereafter, the data provider *defines and fills in the legal usage framework*. If an appropriate legal usage framework has been specified previously the provider only has to fill in the according information. This entails topics such as specifying the access rights, the allowed usage, and the data's security class that defines the data's sensitivity, e.g., whether it is ranked as public, internal, or confidential. Specifications such as these are relevant for ensuring data privacy and compliance with legal regulations such as the General Data Protection Regulation (GDPR) [Eur16]. In this step, the provider may seek the assistance and guidance of legal experts. If a legal usage framework already exists for similar data this may be reused. In our example, the sensor data is not personal data and is ranked as internal, so all employees can access it, and there are no limitations to the usage.

Part 2 - Publish: After documenting and specifying the legal usage of the data, it must be published, i.e., made known so the data can be found by company employees. To begin with, the data provider must issue a *request to publish data*. If a new legal usage framework was defined, it has to be verified by the data owner and legal experts. If the data provider has attained the consent of the data owner to publish the data, they also have to verify the authorization of the requester to release the data and whether the publication of this data is compliant with legal regulations. If the legal requirements or authorization are not sufficient, the request is either rejected or the legal framework must be adjusted. If it is approved, the provider may *publish the data documentation*.

In practice, there are tools for publishing data such as data catalogs. Catalogs support finding and understanding data, however, are not built to access data. For this reason, there are further publishing tools such as Enterprise Data Marketplaces (EDMPs) through which the data can also be requested

and accessed. Examples are Snowflake[1] or the Dawex Data Exchange Platform[2]. Companies, including the industrial manufacturer, are in the process of building a tool landscape for finding, understanding, and accessing data using tools such as these [EGG+21a]. Publishing data therefore entails entering the data into these tools so the data becomes discoverable. For this, a basic set of metadata must be provided, such as the name and location of the data source, a short description what it contains, and who owns the data. The providers may require assistance from IT for integrating their data source into the inventories as this may require technical expertise.

In continuation of our example, the steward would contact the data owner and legal experts, e.g., by email, asking for the permission to publish the data together with the defined usage framework. Given approval, the steward registers the data in a data catalog as well as the data marketplace adding metadata on the data source in both tools. In this example, an Oracle database with the machine sensor data is registered, with metadata like a description that this is data from the production lines and the name of the data owner.

Part 3 - Prepare Provisioning: At this point, employees can find and understand data through the inventory and provided metadata. The data provider now enters the third part of the journey and must provide a provisioning option for the data in the event that there is an access request. If the data is currently hosted on a local machine, the provider must first *upload the data* to some system through which it can be made accessible, e.g., a data lake. If the data is uploaded to a different system, the inventory must be updated. Next, the *provisioning option must be chosen*. The responsible person for the data, i.e. the owner, has to decide, possibly with help of the IT department, whether to *enable source system access*, e.g., via a user account. This might not be desirable, for instance, due to a potential system overload or the risk of data manipulation in operational systems. In our example, direct access to the sensor data is not possible because of the risk that it may be manipulated and jeopardize the machine warranty. Based on this decision, the provider can either enable system access for the data consumers or implement an alternative access method. Providing another access method may be resource intensive,

[1] Snowflake: https://snowflake.com/workloads/data-sharing
[2] Dawex Data Exchange Platform: https://www.dawex.com/en/data-exchange-platform

e.g., by requiring a team of developers. If this is the case the provider has to *request the resources* to implement the provisioning option. Given permission by management for these resources, access methods such as the *transfer of data into another system* like a data lake, the implementation of an *API* for access without a specific user account, or *download options* can be carried out by IT. By way of example, the data steward requests resources to provide an API through which the machine sensor data can be accessed. As machine maintenance is of high relevance, the resources are granted and the API development approved. If management rejects the request and no provisioning option can be guaranteed, it would be useful to indicate this circumstance in the inventory or to remove the data from it accordingly. Subsequent to performing these steps, employees can find, understand, and receive access to the data through tools such as data catalogs or a data marketplace.

3.2.2 The Data Provider Challenges

To enable broader access to the data within the enterprise, as intended through the first democratization dimension, the data providers must carry out the previously described journey and first, document, publish, and then provide provisioning options to the data. The role of the data provider is, however, not an explicit job role such as that of a data scientist or enterprise architect. Therefore, employees in varying key-data related roles must take on this additional role and become data providers to drive the democratization initiative. As the task of providing data is often not part of their job role description, they may lack resources like time for this additional task. Therefore, to enhance the willingness to execute the journey, it must be efficient and involve little effort. From this point of view, the following challenges arise in the provider journey:

The *assembly of metadata* (Prov-C1) constitutes the first challenge for the data provider. Although documentation is a best practice in many processes, it is often neglected. To ensure the usability of the data, however, a certain degree of documentation is indispensable. If no metadata is present, the provider will first have to gather this information and potentially add it to a variety of tools. Since the provider is not necessarily an expert on the data, they may have to rely on other employees, such as the data producer or a data steward, to supply this metadata.

Besides assembling documentation, *supplying provisioning options* (Prov-C2) apart from direct system access, is potentially costly and time-consuming. This task may require an IT project, e.g., for the implementation and realization of pipelines for moving data or developing an API. While it is the goal to enable broad access to as much data in the enterprise as possible, there are many cases in which it is unknown whether the data is of interest to other employees beyond the current consumers. Therefore, adding provisioning options may be an expense that does not provide any benefit as these may not be required.

As described in the industrial use case and journey, companies are currently building diverse tool landscapes that may contain several tools suited for publishing data. In consequence, this means the provider might have to register the data in several tools such as the data catalog as well as a data marketplace. Therefore, challenge three refers to the effort of *registering data in several publishing tools* (Prov-C3) which partly require the same metadata.

Finally, *the process involves several parties* (Prov-C4) that need to be found, contacted, and coordinated. With each new party, the process becomes more complex and time-consuming. Furthermore, there is no integrated tool support for handling communication with all the involved parties throughout the provider journey.

3.3 The Data Consumer Perspective

Broad and easy access to data, as part of the first democratization dimension, enables a more extensive and innovative utilization of data and, consequently, a more extensive extraction of this data's value. It is based on the ability of company employees to find, understand, and gain access to data that is relevant for their use case. This section complements the previous section on the data provider perspective by investigating the current state how data consumers gain access to data and which challenges they face throughout this process.

To identify the relevant aspects in the data consumer's process, we conducted a literature study, amongst others including the works [LLF21; AG20; Grö21; GH19; FSF20; LLEF20; ZG18]. Besides discussing data democratization, several of these research articles yield rough insights into the data consumers' workflows. In order to gain a practical and more detailed perspective, we also

conducted expert interviews with employees of our industrial manufacturer. The exchange with more than ten experts from various key data-related roles, including enterprise and solution architects as well as data scientists and business analysts, gives us a representative view on the current workflows for gaining access to data from different perspectives.

The findings on the data consumers' processes from both the literature study and expert interviews were merged into an overarching data consumer journey, which we present in Section 3.3.1. Like the provider journey, the consumer journey renders an overview of the steps, involved parties, and tools used in the process to attain data. It also yields a number of data consumer challenges which are discussed in Section 3.3.2.

3.3.1 The Data Consumer Journey

The representative data consumer journey for industrial enterprises, illustrated in Figure 3.3, consists of three segments: *finding*, gaining *access* to, and *preparing* data. As the preparation of data is strongly dependent on the individual use case, we do not discuss this step in the following. It may, for instance, involve a simple cleansing of data in one use case and a complex integration with other data in another use case. As companies are still in the process of realizing data democratization, only fragments of the company data if at all, are prepared for consumption according to the previously presented provider journey. Therefore, the following data consumer journey does not strictly assume that the data under consideration has been published and prepared according to the provider journey. To illustrate the consumer journey by way of an example, we demonstrate it based on the scenario of a manufacturing engineer working for our industrial manufacturer. The engineer wishes to access data from the sensors of running production lines to create a machine maintenance dashboard to promote the realization of predictive maintenance.

Part 1 - Find: To begin with, data consumers must find relevant data for their use case. Finding data involves *searching for data*, *understanding* it, and *evaluating its relevance*, i.e., whether it suits the use case. If data has not been registered in an inventory, such as a data catalog tool, the consumer likely does

not know of its existence and is dependent on tribal knowledge. The consumer must contact various employees and rely on this knowledge spreading mouth to mouth. If, in contrast, it has been published through a tool like a catalog, then a consumer can use this tool's explicit search and browse function to search for specific data. In our example, the data has been registered in a data catalog so the manufacturing engineer can enter a search string such as "sensor data production line P1" into the catalog's search interface to find

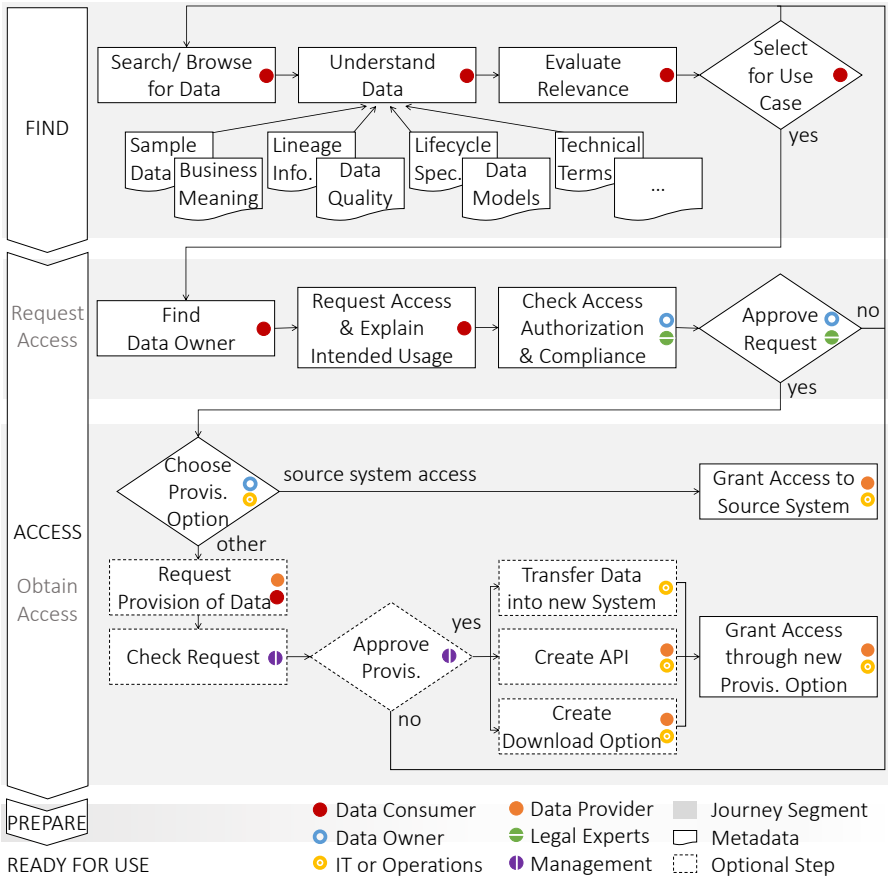


Figure 3.3: The Steps and Parties Involved in the Journey for Finding and Accessing Data Within an Enterprise (Based on [EGH+22a].)

search results on various sensor data. Once potentially relevant data has been discovered, the consumer evaluates whether it fits their use case based on metadata, or if not existent, explanations provided by, e.g., the data owner or a domain expert. While the data catalog may provide some metadata, other tools are specialized on specific metadata and provide further information such as business glossaries, data quality platforms, or model repositories with semantic data models that may be integrated with the business glossary [EGG+21a]. For more information on metadata sources refer to Section 2.3.3. Having determined the data's suitability, the consumer either continues searching for more suitable data or moves on to pursue data access. In continuation of our example, the engineer consults the metadata offered through the data catalog on the search results, e.g., explaining that a dataset contains sensor data on machine temperatures and vibrations. The manufacturing engineer knows that these measures reflect machine lifetimes and thus, are relevant for the maintenance dashboard. The engineer gathers additional information on the dataset's completeness, i.e., quality measures from a data quality tool and gains insight into the dataset structure from a data model repository and is now ready to select the dataset for further use.

Part 2 - Access: Gaining access to data entails *requesting access* and *obtaining access*. Within the first step, the consumer must *find the data owner* and send a *request for access including the intended usage*. This is relevant for compliance with legal regulations such as the GDPR. The request is either sent manually, e.g., via email, or tools that regulate the access to certain resources and enforce access policies such as SafeNet Trusted Access[3]. Having received an access request, the owner or a legal entity *checks the consumer's authorization* based on the data's confidentiality and the employee's clearance level. Even with an appropriate clearance level, the consumer may not be allowed to process the data in the intended manner. For example, GDPR allows people to take influence on how their personal data is processed [Eur16]. Hence, if either the authorization or intended usage is insufficient, the request is declined and the consumer must resume their search. In our example, the engineer wants to access a small section of data that cannot be requested

[3]SafeNet Trusted Access: https://cpl.thalesgroup.com/en-gb/access-management/safenet-trusted-access

through the used authorization tool, however, the owner is registered in the data catalog and the engineer can send this person a request through other channels like email. As the sensor data is not personal data, the engineer may process it for the intended use case so the request is approved by the owner.

Having acquired access approval, the data owner or IT specialists determine how to grant access to the data. Depending on factors such as the intended usage or the source system's capacity, they might directly *grant source system access* or choose to *grant access through a new provisioning option*, e.g., by *transferring data into a new system, creating an API*, or *creating a download option*. If the other provisioning options do not exist yet, the data consumer or provider might have to *request the provision of the data*. This requires resources, like an IT project to transfer the data into a new system, and, therefore, the *request is checked* and has to be approved through management. If denied, the consumer resumes the search, if approved, the new option is implemented and access is granted to the data. In our example, the engineer is not granted access to the source system as it already has limited capacities. Hence, they contact management to approve replication of the data into a data lake. As the design of the dashboard is a high-profile project, management approves and IT experts replicate the data and finally grant the engineer access permission.

3.3.2 The Data Consumer Challenges

Like the provider journey, the consumer journey also yields a number of challenges that hinder the democratization of data. Firstly, this *process also involves several parties* (Cons-C1) ranging from the data consumer over the data owner, IT specialists, legal experts, and management. With each additional party, the process becomes more complex and time consuming as the responsible people have to be found and contacted. There may be no alloted responsibilities and according documentation concerning the data, which means this information has to be gathered and found out laboriously by the data consumer. Additionally, there will inevitably be delays until each party processes their tasks. In our example scenario, the manufacturing engineer has to locate various persons such as the data owner, domain expert, and people from management, if not defined in the catalog.

The second challenge lies therein that *the metadata for understanding the data is spread across a variety of tools* (Cons-C2). It is inconvenient and challenging to maintain an overview of both the tools and metadata viewed on each dataset. The manufacturing engineer, for instance, had to collect metadata from at least the data catalog, the business glossary, the data quality tool, and model repository.

Lastly, *the tools are not integrated across the access process* (Cons-C3), so several tools are required, and not all data that can be found in the metadata management tools like the catalog may be requested in the authorization tool. The engineer, for example, had to manually contact the data owner through channels like email to receive access as the authorization tool did not offer the same entries as the catalog.

3.4 Summary

To lay the foundation for improving data democratization in companies, this chapter covered the current processes and challenges regarding the democratization dimension of broader access to data. Based on a representative use case of an industrial manufacturer and literature study, a generalized data provider and data consumer journey were introduced for making data available and gaining access to data respectively. The journeys yielded a number of challenges both the data providers and consumers face in their processes.

The data provider challenges entail difficulties in the assembly of metadata (Prov-C1), the effort of supplying provisioning options (Prov-C2) and registering data in various publishing tools (Prov-C3), and that the process involves several parties which need to be found, contacted and coordinated (Prov-C4). Like the provider, the consumer has to find, contact, and coordinate several parties to gain access to data (Cons-C1). Understanding and consequently selecting data is difficult as the according metadata is spread across a variety of tools (Cons-C2). Lastly, there is no integrated tool support to assist the data consumer across this process (Cons-C3).

In this regard, companies are investigating data marketplaces to improve their data democratization initiatives [EGG+21a; Wel18; Grö21]. In the following Chapters 4 -6, we will examine how a data marketplace can be used in the

company-internal context. How data marketplaces address the data provider and consumer challenges presented in this chapter and to what extent these support the data democratization initiatives is assessed in Chapter 7.

Chapter 4

Introducing the Enterprise Data Marketplace (EDMP)

With the multitudes of data generated and collected in companies today, it is one major ambition of these companies to leverage their data's value, e.g., to gain new insights for enhancing business models. The data value can, however, only be extracted if the data is available for use. Studies show that over half of the data goes unused within companies [Sea20; spl19]. Therefore, companies are implementing data democratization initiatives to empower their employees to find, understand, access, use, and share their data within the company. However, as presented in the previous chapter, they are faced with various challenges for both the data consumers and providers. To further their data democratization cause and address the consumer and provider challenges, companies are investigating data marketplaces.

In literature the data marketplace is proposed as a solution for ensuring data access, which is one key component of data democratization [TA23]. Yet, data marketplaces are mainly considered for the exchange of data and services between organizations or private individuals, i.e. as external data marketplaces. In the company-internal context, the data marketplace is referred to as an Enterprise Data Marketplace (EDMP) [Grö21; Wel18]. 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 [JO23]. Topics in this regard also include how it differs from other types of marketplaces, which specific requirements it demands, or what challenges arise in this marketplace context.

In this chapter, we close this gap by introducing the Enterprise Data Marketplace as a marketplace type and identify the distinctive aspects of this marketplace, thus, addressing research goal two (RG2). The research contributions of this chapter encompass the following: Firstly, a definition of this marketplace type is given and we position the EDMP in a classification framework differentiating it from other marketplaces and hence, identify its characteristics. This corresponds to the research contribution C2.1. Secondly, we present requirements for data marketplaces and highlight which of these are specific to an EDMP, providing research contribution C2.2. Lastly, we illustrate EDMP challenges based on the idiosyncrasies of this marketplace type, addressing research contribution C2.3.

This chapter is structured as follows: Related work is presented in Section 4.1 followed by the definition and goals of the EDMP in Section 4.2. The EDMP is classified based on a classification framework in Section 4.3 and requirements are specified in Section 4.4. Thereafter, the EDMP is differentiated from the related data catalog tool in Section 4.5. Challenges for EDMPs are listed in Section 4.6 and finally, Section 4.7 summarizes and concludes this chapter.

This chapter is a revised and composite version of the author publications [EGH+23; EGH+22b] and [EGHS22].

4.1 Related Work

The EDMP is addressed in only a few research articles. Amongst others, Gröger [Grö21] highlights the need for this specific marketplace type, Fernandez, Subramaniam, and Franklin [FSF20] consider them to bring down data silos, and Wells [Wel17] defines and presents the EDMP in a report. Driessen, Monsieur, and Van Den Heuvel [DMV22] present 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 [EGG+21a; EGH+22a; EGHS22]. Azcoitia and Laoutaris [AL22] 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 that these are often limited in their functionality. Hence, they are similar but not equivalent to the company-wide EDMP. Jahnke and Otto [JO23] identify the EDMP as one class of data catalog application. They highlight the EDMP as a research gap for which further details on its capabilities and value-adds have yet to be determined. Lastly, Zasadzinski et al. [ZTTR21] present how they built 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.

Data marketplace characteristics, relevant in the context of classifying data marketplaces, are studied in various research articles such as [FRP20; SSVV17; SSVV16; SSV13; TL18; MS19; LSV18; Spi19; KLT17; AL22]. For instance, Schomm, Stahl, and Vossen [SSV13] provide an initial set of classification dimensions which are extended by Stahl et al. [SSVV17]. Meisel and Spiekermann [MS19] derive five classification characteristics and Spiekermann [Spi19] provides economic and technological characteristics of marketplaces. Täuscher and Laudien [TL18] list key business model attributes of marketplaces, which are however not exclusive to data marketplaces, and Azcoitia and Laoutaris [AL22] classify data exchange entities including data marketplaces through business model attributes. Fruhwirth, Rachinger, and Prlja [FRP20] provide a list of characteristics that are assigned to dimensions such as value capture, delivery, proposition, and creation. Yet so far, the EDMP has not been classified based on any of these frameworks, hence, we provide this placement in such a framework.

Requirements for data marketplaces are listed in a range of research articles. Fernandez, Subramaniam, and Franklin [FSF20] 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 [DMvdH23], trustworthiness through,

e.g., blockchain [LSR19], or marketplaces in the Internet of Things (IoT) context [SBK+17; KPKS18]. 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 [ALL20]. Requirements for the EDMP could be derived from this marketplace type's descriptions as supplied in, e.g., [Grö21; Wel17; EGH+22a], 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.

Lastly, challenges within data marketplaces are discussed in many research articles. Amongst others, these include challenges concerning the valuation and pricing of data [Pei22; LLL+21; ADS19; SSV13], providing fair compensation for data providers [AL22], the derivation and assurance of data quality [KLT20; FSF20; DMV22], the ability to combine datasets to satisfy buyers' needs and establishing trust amongst the participants [FSF20; DMV22], and issues of data procurement [ZMWC19]. Driessen, Monsieur, and Van Den Heuvel [DMV22] provide a taxonomy for problem categories specific for data markets and map these to different types of marketplaces. We highlight challenges related to providing and accessing data within the enterprise, as discussed in the previous Chapter 3 and in our work [EGH+22a; EGHS22], yet, these are role-specific to the data consumer and data provider. Hence, as none of the research articles discuss challenges specific to the EDMP we address this gap.

4.2 Definition and Goal of the Enterprise Data Marketplace

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 [Grö21; Wel18] or an internal data marketplace [FSF20]. In extension of Wells [Wel18] definition, we propose the following:

The EDMP 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 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 marketplace can also work with existing tools in a company such as data catalogs. These provide an inventory of data over the above-mentioned storage systems and facilitate finding and understanding the contained data [ZDED17]. The EDMP complements tools like the data catalogs with additional functionality such as features for requesting and managing access to data.

In an example usage scenario in line with the scenario presented in the last chapter, a manufacturing engineer is looking for data collected from the sensors of running production lines in order to realize a predictive maintenance use case. The manufacturing engineer can use the EDMP to search for such data and find that there is data according to their requirements stored in a data lake. The manufacturing engineer requests access to this data through the marketplace and receives access to available provisioning options for this data. Optionally, the manufacturing engineer can request the data with additional software or infrastructure like a virtual machine (VM), so they are directly supported throughout their use case or also according to their skill-set.

4.3 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 Section 4.3.1 and the identified characteristics are discussed in the following Section 4.3.2. By highlighting its specific features, we introduce the EDMP as a distinct marketplace type.

4.3.1 The Marketplace Classification Framework

In this section, we present a classification framework that we 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 (DMP) and on the other hand, it can be used to determine whether DMP solutions are suited for the use within an enterprise as an EDMP or if they are more suited as an external data marketplace (Ext-DMP) for use between enterprises.

To create the framework, we studied data marketplace characteristics provided throughout the research articles listed in Section 4.1. The characteristics range from aspects like marketplace ownership over the value proposition, data access methods, monetization aspects, to the underlying architecture. Some characteristics such as the marketplace ownership [SSVV17] are relevant for the distinction of the EDMP, whereas other characteristics like the offering of pre-purchase testability [SSVV17] are not. Meisel and Spiekermann [MS19] provide a classification framework by combining characteristics identified through various research articles including [LSV18; SSVV16; KLT17; SSV13]. Spiekermann [Spi19] also provides a data marketplace classification framework based on a taxonomy developed for classifying data marketplaces based on their business models. By combining these frameworks an overview covering various dimensions of data marketplace characteristics can be obtained.

Hence, we developed the data marketplace classification framework to distinctly classify the EDMP, as displayed in Figure 4.1, by combining the two frameworks provided in [MS19; Spi19]. By grouping the attributes from these two frameworks, 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*. These dimensions are depicted on the left in Figure 4.1. The according dimension attributes, like the *provider* or the *consumer* for the market participants dimension, are listed in the middle, and attribute characteristics on the right. The characteristics represent the alternative options in which the attributes may manifest. For instance, *private individuals* may be a DMP's data providers and thus signify a characteristic. The characteristics that apply to the EDMP are highlighted in shades of blue.

Besides combining the two 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 to company which complements the characteristics *private individual* and *public institution* in this section. Also, the attribute "market positioning" [Spi19] is replaced through the more expansive attribute *ownership* of [MS19] and the attribute

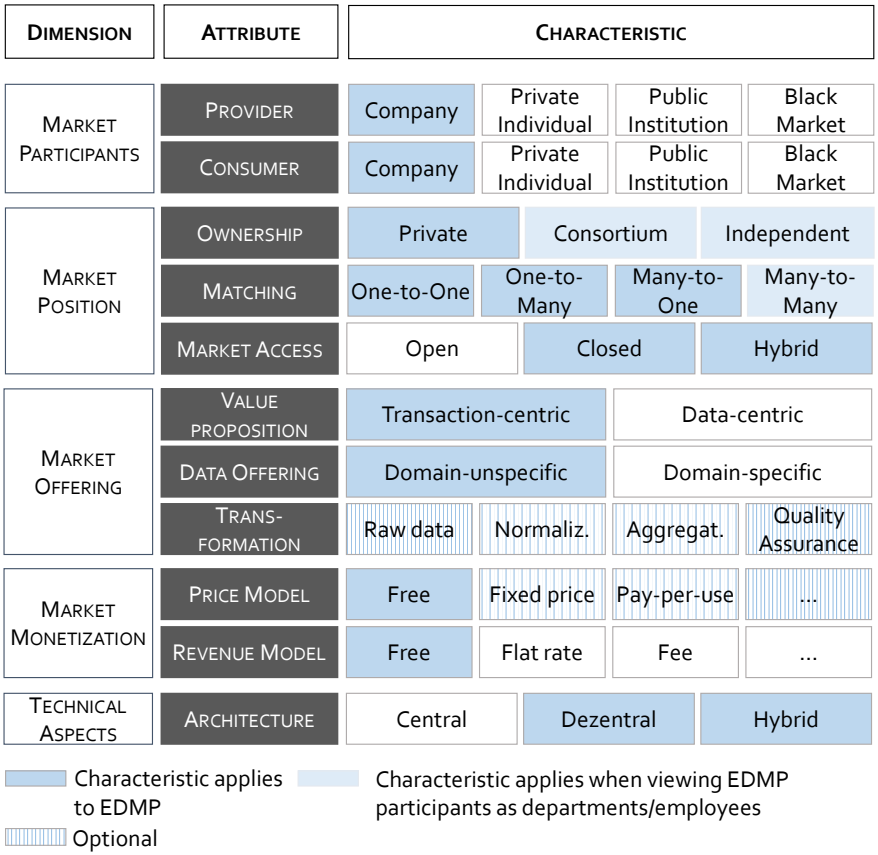


Figure 4.1: Data Marketplace Classification Framework Highlighting the Characteristic-Profile of the EDMP in Blue [EGH+23].

"integration" [Spi19] is renamed to *data offering* for the sake of a more precise naming.

4.3.2 Enterprise Data Marketplace Characteristics

For each attribute 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 marketplace. In the case of the EDMP, the participants in both categories are employees within the same *company*, which is not immediately apparent through the classification framework. In some cases, an enterprise may choose to open its marketplace to selected business partners [Wel18], who also classify as companies.

The dimension *market position* signifies who *owns or operates* the marketplace, the *matching*, i.e., the number of parties involved, together with the service orientation among these, as well as the *accessibility* of the 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* marketplace depending on whether the department operating the marketplace 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 [Spi19]. The scope of offered data spans across all company data, hence, the data offering is *domain-unspecific*. According to Spiekermann [Spi19], transformation refers to the marketplace's ability to transform raw data into a normalized or aggregated state or assure data quality. While we argued in [EGHS22] that a marketplace does not offer functionality to process data, e.g., aggregate it, the marketplace can offer data in various transformation states, e.g., data stored in data lake zones in varying processing degrees. Therefore, 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 marketplace. As a revenue model would be a barrier for employees to use the marketplace, 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 marketplace repository. Therefore, it 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.

4.4 Enterprise Data Marketplace Requirements

Having identified that an EDMP has delimiting characteristics, we now examine whether it also has an independent set of requirements that go beyond

those of data marketplaces in general. Firstly, the required marketplace offerings, in terms of what data consumers can acquire through it are discussed in Section 4.4.1. Secondly, the required marketplace functionality is presented in Section 4.4.2. Thirdly, as this marketplace is operated within a company, requirements on how the marketplace should integrate with the existent enterprise system landscape are illustrated in Section 4.4.3. The requirements are derived from existing literature on data marketplaces and data democratization, complemented by practical insights provided through the use case of the industrial manufacturer, introduced in Section 3.1. Some requirements for the EDMP overlap with those for data marketplaces in general but we highlight the requirements that are specific to the EDMP. A subsumed version of the requirements and their relevance for the EDMP are listed in Table 4.1.

4.4.1 Required Marketplace Service Offerings

The term offerings refers to the items, or in this case services, which a consumer can acquire in the 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.

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

Table 4.1: Relevance of Requirements in the DMP and EDMP.

In order to facilitate the first data democratization dimension of broader access to data [LLF21], all kinds of data have to be made available within the company [Grö21]. Therefore, the data marketplace's main offer must be *Data-as-a-Service* [Wel17]. Ultimately, the marketplace should make all corporate data available. It should be noted that availability does not mean that any user can gain access to all data, it means the data can be found, understood and access can at least be requested. The scope of data 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. The scope also covers, raw data, data in various processing degrees as well as ready-to-use data, and data insights such as machine learning models or reports. As explained in Section 4.3.2, 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 [Grö21; AG20]. This type of user 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 marketplace must 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 VM. The marketplace could provide the VM so 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 marketplace 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 second data democratization dimension.

The development and sharing of data skills is part of the third data democratization dimension [LLF21; AG20]. Hence, the marketplace 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.

[1]Tableau: https://tableau.com
[2]Microsoft Power BI: https://powerbi.microsoft.com

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

4.4.2 Required Marketplace Functionality

Literature and reports yield a range of functionality for data marketplaces such as functionality for selling and buying data, governance topics like license management, monetization aspects like pricing, revenue allocation and sharing functions [FSF20; SLT+20], functionality for data cleansing and preparation [Wel17; MS19; KLT17; RPT+17; SLT+20] or integration [MS19] and analytics [Sax18]. Throughout these lists, the extent of functionality differs and is described at different levels of detail. Furthermore, the structure differs as well, with some articles listing functionality by role, e.g., [RPT+17], and others by functional groups like marketplace infrastructure, interfaces, and security, e.g., [MS19]. We notice that some lists go beyond the scope of the marketplace as we understand it. From our point of view, a data marketplace is purely a broker for data and data related services. It is a platform on which data providers can publish data and services and data consumers can find, understand, and gain access to these. How the data is, for instance, prepared or integrated with other data is, in our opinion, beyond the scope of the marketplace as a broker. Finally, we noticed that literature devotes little detail to the topic of metadata in the context of data marketplaces. Since finding and understanding data is a crucial feature of the marketplace, and this is dependent on metadata supplied by the data provider, we consider metadata management to be a relevant underlying and role independent functionality in the marketplace. Therefore, we have created a functionality framework that takes these three aspects into account: the division by role, the delineation of functionality that lies within and outside the marketplace, and the metadata management that is the basis for the role-specific features.

The Functionality Framework: To create the functionality framework, we incorporated the common parts from the different functionality lists derived from the literature study and allocated them to the roles of the data consumer, data provider, and administrator who is associated with the role of the platform owner. The functionality that was only partially represented, such

as rating or data cleaning, was examined whether it belonged to the functional scope of a broker and was accordingly included or excluded. We also had to find a common level of abstraction that subsumes the more detailed tasks. Additionally, we extended functionality such as the necessary metadata features based on input from the expert interviews and the role-based functionality.

An overview of the resulting functionality framework is illustrated in Figure 4.2. More detailed views on the functionality per topic-area are provided throughout the following passages. The framework provides insight on the required functionality and can therefore be used as a guideline for implementing a data marketplace. It also provides a basis for comparing commercial tools, as well as a basis for evaluating which functionality is already offered by other tools within a company and which functionality is missing. While most of the functionality listed is also required in other data marketplaces, we point out that specific aspects like the metadata management as well as privacy and security may take on a broader scope in the EDMP.

The marketplace's functionality is displayed in the blue box and other functionality outside of it. Data governance and data management including data quality or data lifecycle management take place outside its functional scope as

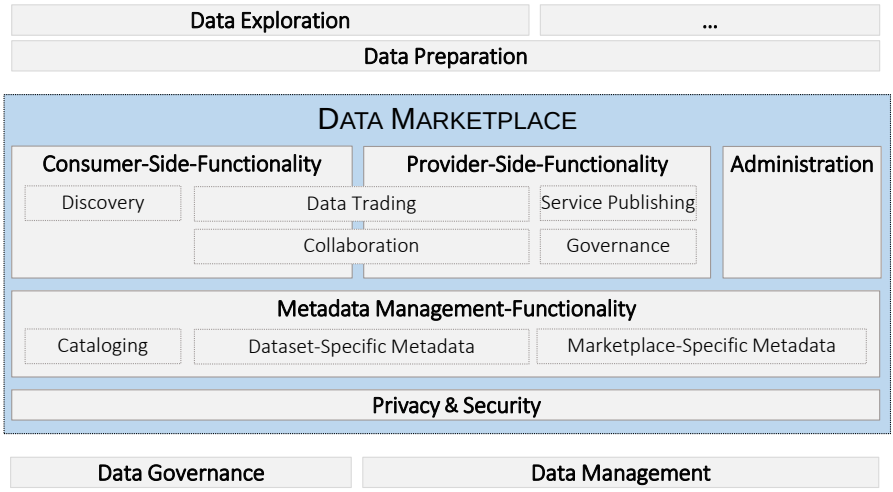


Figure 4.2: The Data Marketplace Functionality Framework (Summarized and Adapted from [EGHS22]) [EGH+23].

these concern the management as opposed to the exchange of data. Equally, activities that follow the acquisition of data such as data preparation or exploration are out of scope as these involve data processing which goes beyond data sharing. To enable an integrated data processing toolchain, the marketplace nevertheless provides interfaces to tools that perform tasks outside the marketplace context, such as preparation tools. Within the marketplace, we distinguish between consumer-side functionality, provider-side functionality, and administration functionality. The metadata management functionality, privacy, and security extend across these areas.

Consumer-Side-Functionality: The role-based functionality as illustrated in Figure 4.3 is not necessarily specific to EDMPs, as it is required both in the company-internal as well as company external context. The consumer-side functionality includes discovery, data trading, and collaboration features. The consumer can browse or search for data and services in the marketplace, like the machine sensor data provided by the data steward in the provider journey example. For each search result, there is a detailed description with an integrated view of all available metadata. For example, the detailed description could contain a description of the machine and the according production line with technical details how and where the data is stored or operational information such as the data's lineage. The marketplace can also offer service recommendations based on the conducted search, previous acquisitions, and those of similar users.

In order to support the fourth data democratization dimension regarding collaboration and knowledge sharing [LLF21], the marketplace also offers functionality such as commenting to both the consumer and producer. Furthermore, consumers can rate data and document their use case, thereby enabling other users to see if the data has been used for similar use cases. In our example, a user could specify how they used the data in a machine maintenance use case. These functionalities are only available to the consumer and, thus, placed in the consumer-side box in Figure 4.3.

The data trading functionality like the collaboration functionality is overarching in Figure 4.3, as it is available to both the consumer and producer. On the consumer-side, service-access-management signifies the ability to request and receive access details and a license to use data, e.g., by ordering it through

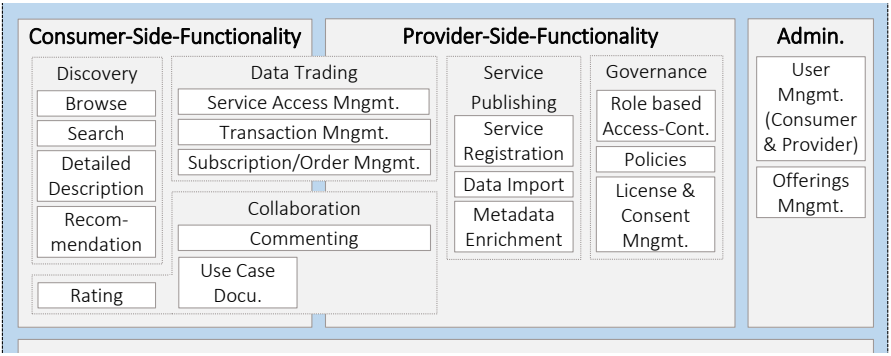


Figure 4.3: Role Specific Functionality in the Data Marketplace (Summarized from [EGHS22].)

a shopping cart. Additionally, the consumer can manage transactions related to reimbursements for services, as well as active, expired, and pending orders through the subscription and order management. In continuation of the example described in Chapter 3, the user can order the machine sensor data through the shopping cart and request access, then pay for this data through the transaction management and after receiving access, view the active subscription on this data with according details.

Provider-Side-Functionality: The provider-side-functionality involves publishing, governance, and data trading functionality. For publishing services such as data-as-a-service or professional services like courses for data preparation, the provider uses the service registration. In this step, the marketplace adds the service, mostly a data source or a specific dataset, to the marketplace service inventory so it can be found via the search. In our example, the data steward registers the machine sensor data in the marketplace which is thereby added to the marketplace inventory. Although the marketplace will reference most data as opposed to storing it, it does have a data import feature for cases like the upload of a locally stored singular csv file. In our example, the database in which the machine sensor data is stored is registered instead of uploading the data directly. The metadata enrichment allows the provider to add additional metadata to ensure a better understanding of the data. This may include a variety of metadata such as a content description, technical details

as well as information about the data provenance. These could for instance be descriptions of the machine and production line, the database, and the lineage showing how the data originates in the machine and is moved to the database by a specific script.

Besides service registration, the marketplace offers governance functionality that supports the specification as well as compliance to aspects defined within the data's legal usage framework explained in the provisioning journey in Section 3.2. A provider can define role-based access control, usage rights within policies and package these in licenses for specific services. This functionality does not replace the underlying data governance, it merely enables the implementation of the marketplace-specific governance aspects. For instance, the decision who is allowed to do what with the machine sensor data is part of the governance outside the marketplace. Within the marketplace, the steward in our example merely specifies that only people from department x are allowed to use it for maintenance use cases.

Like the consumer, the provider has functionality to support them in the trading of data. For instance, the provider can manage access requests, i.e., receive notifications, consult an overview, and accept or decline these requests. If monetization or other forms of reimbursement are included in the marketplace, the provider can monitor transactions for the offerings. For example, the steward can view closed and outstanding invoices for the sensor data. Having provided access to the data, a provider can then handle the subscriptions and orders on the offered services. This includes an overview of who is subscribing to which data, options to contact all subscribers or functionality to terminate subscriptions and revoke access rights. In terms of collaboration, the provider can also enter into a comment dialog with the data consumers.

Administration Functionality: In addition to supporting consumers and providers with their tasks, a marketplace must also offer functionality for administrative purposes such as user and offerings management. As companies usually maintain a multitude of tools [EGG+21a], it would make sense in the internal context if the user administration is not handled separately per tool, but jointly through, e.g., a user rights management system. Hence, the administrative tasks may be realized differently internally and externally.

Metadata Management Functionality: In contrast to the role-specific functionality, the metadata management functionality is distinctive in EDMPs. Data marketplaces are metadata-driven platforms, and, therefore, the handling of metadata is a central aspect within these. A data marketplace requires a variety of metadata to support the role-specific functionality. As depicted in Figure 4.4, this comprises general metadata for cataloging, as known from a data catalog tool like an inventory, dataset-specific metadata such as a business description, and marketplace-specific metadata like the purchase history.

Metadata for cataloging can refer to a range of datasets and helps to provide an overview of existing offerings. It includes a data inventory, e.g., a list of contained datasets, data links that indicate whether data sets are related as well as data similarity information, which reveals replicated and similar datasets.

Dataset-specific metadata refers to a specific dataset and helps users to understand and trust this data. Amongst others, this covers data quality, lineage, and versions. It is important to understand that the maintenance of the dataset-specific metadata is not part of the marketplace, merely, that it is relevant for the consumer in the sense of finding and understanding data. Therefore, this metadata has to be supplied by the provider and the marketplace must support some form of indexing and integrated processing and presentation of these. For instance, data quality metrics like completeness or accuracy can be extracted from tools that maintain data quality.

The marketplace-specific metadata comprises a product registry, purchase history, transaction history, search history and metadata statistics. The statis-

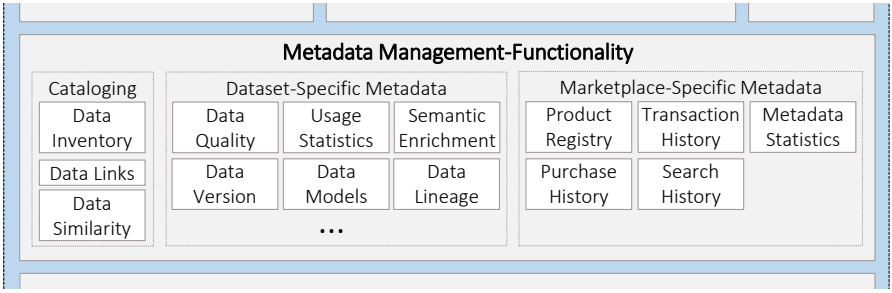


Figure 4.4: Metadata Management Functionality in the Data Marketplace (Summarized from [EGHS22].)

tics indicate to what extent the metadata is complete or contains user statistics such as how often a service has been viewed.

As companies already have infrastructure that collects and manages a wide variety of metadata with tools such as data catalogs or business glossaries [EGG+21a], the EDMP has significantly more metadata at its disposal than external data marketplaces. Furthermore, the EDMP can be tailored to reflect enterprise idiosyncrasies. For example, companies often have a company-internal "language," i.e., specific vocabulary that is maintained through tools like business glossaries. By way of example, a company may refer to an end product as "material." Yet, usually, 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.

Privacy and Security: 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 may be registered in the EDMP, which includes both highly confidential and sensitive data. The ISO/IEC 27000 series [ISO18] 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 [GSB+22], digital signatures to realize data integrity [ZYC+19], proof of retrieveability to address availability [Gri20], and attribute-based signatures to ensure authenticity [GÖM19].

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 is requested for a variety of use cases. For in-

stance, 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 [Eur16]. 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 dataset, other parties may access an anonymized or distorted version of the data, and some may not be allowed to know that this data exists. However, by distorting the data the data quality may be affected [ASZ+19], 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 [Sta23]. 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 [ACD+07; ABF+18], images [Fan19; YZK+17], and time series data [Pou89; RP19].

However, the topic of privacy and security constitutes an extensive research area and is not the focus of this thesis and hence, not considered in detail throughout the rest of this thesis. In contrast, metadata aspects will be the focus of this paper and are discussed in more detail. Nonetheless, these topics are related, as the 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 [CKD+04].

Based on this framework, we have gained insight into which explicit functionality a marketplace should offer to the different roles and which implicit functionality like the metadata management or privacy and security is required. Furthermore, it has been pointed out which of this functionality is more or less relevant within the company-internal context.

4.4.3 Enterprise Integration Requirements

External marketplaces for trading data between organizations are often standalone marketplaces like Advaneo[3] or part of an entire platform such as the AWS Data Exchange[4] 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 lake's customized zone architecture such as that by Giebler et al. [GGH+20b] and reference data in the according zones.

As mentioned previously, there are a variety of data and metadata management tools that are used to manage data within a company. These tools include data catalogs, business glossaries, and model repositories. Some of these tools provide functionality that is required in a 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 that is relevant for finding, understanding, and, consequently, choosing data for use. This information can be reused within a marketplace. Therefore, the EDMP should tightly *integrate with the existent data and 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 rights. By *integrating with administrative tools*, single sign-on and authorization management across source systems, including

[3]Advaneo: https://advaneo-datamarketplace.de
[4]AWS Data Exchange: https://aws.amazon.com/data-exchange

the EDMP is possible. The marketplace 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.

In this Section 4.4, we have presented a series of requirements for the EDMP concerning the offerings, the functionality, and the way it should integrate into the company system landscape. As emphasized repeatedly, some of the requirements apply to data marketplaces in general whereas others are particularly relevant in the company-internal context and need to be addressed with certain aspects in mind.

4.5 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 [ZDED17; Grö21] 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.

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 [ZDED17]. 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 [ZDED17; LLEF20]. 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 marketplace also has functionality to register and discover data as well as collaborative features [EGHS22]. As shown in Figure 4.5, the catalog and EDMP both provide this functionality. Having understood and selected the data, a user will want to gain access to it. In this regard, the marketplace 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 marketplace 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 [LLEF20]. Thus, the discernment between the catalog and data marketplace becomes unclear. Jahnke and Otto [JO23] 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 Figure 4.5, wherein, the data catalog is depicted as a component of an EDMP. However, Jahnke and Otto [JO23] 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.

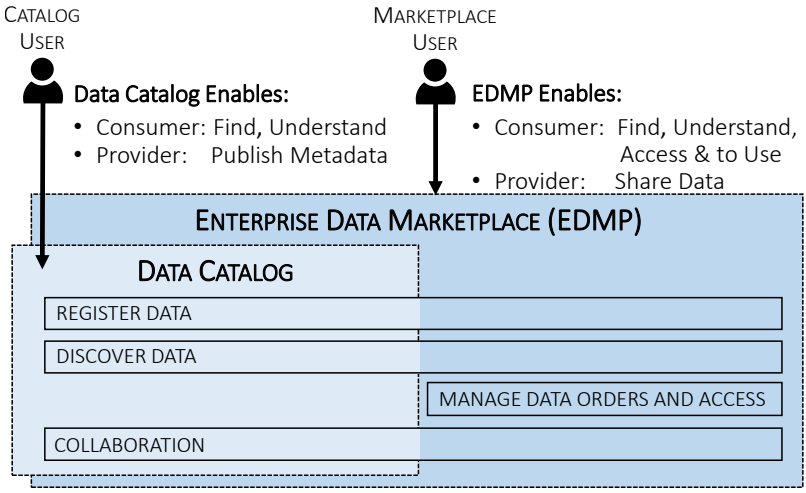


Figure 4.5: Differentiation of the EDMP and the Data Catalog [EGH+23].

Besides the offered functionality, the data catalog and EDMP also differ in terms of services offered to users. As explained in Section 4.4.1, 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 VM or software like Tableau, and with a course, e.g., to learn how to integrate data. The data catalogs, if they support data access functionality, offer data-as-a-service. In this regard, the scope of offered services 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 provide a data inventory and enable connecting data supply and demand through data discovery [JO23], whereas 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 marketplace serves as an extension through data access management functionality, as depicted in Figure 4.5. How the catalog and the marketplace can be combined to complement each other is discussed in more detail in the following Chapter 5.

4.6 Challenges in the Enterprise Data Marketplace

Based on an in-depth knowledge exchange with the industrial manufacturer who is investigating EDMPs, we discovered technical and organizational challenges surrounding this type of marketplace.

One challenge is the *lack of incentives* for providers to share the data. By omitting monetization in the company-internal context to support data democratization, the main incentive for data providers to share data is removed. As monetization is the main incentive in the company-external context, this

challenge is specific to the EDMP. Initially, the provider has effort that is not compensated. Consequently, other forms of incentives are required for providing data in the EDMP. Driessen, Monsieur, and Van Den Heuvel [DMV22] also emphasize that an exchange in a data marketplace requires that both the data consumer and provider give and gain something [DMV22]. Researchers have suggested bonus points [FSF20] or a gain in new insights or improved data-driven processes [DMV22] in this context. Possibly, gamification in the marketplace, publicity through visibility, coupons for coffee, or the awarding of data sharing titles might also be interesting ideas to add incentivization. This topic calls for an investigation on what drives providers and their executives to promote data sharing or on what would prevent them from doing so. With this knowledge, possible impediments could be removed. For example, we have discussed in our work [EGHS22] how the marketplace can support the provider in data sharing by minimizing the sharing effort.

How data ownership can be retained or passed to acquiring consumers constitutes another challenge. As opposed to the issue of data ownership in the company-external context as outlined by Azcoitia and Laoutaris [AL22], this does not concern the issue of privacy and theft. It rather concerns the question of who will be accountable for the data so it is handled and processed in accordance with legal regulations. If data is acquired, processed, and then made available again as a product, the question arises who will be the data owner. It might be the original data owner, who must then keep track of all processed versions of their data. Yet, how far they can feasibly track their data down the chain of processed versions of this data is questionable. Also, there may be a situation in which several datasets from different owners are processed together, providing the question, which of the data owners would be responsible in this case. Alternatively, the person processing the data or their supervisor may be the new data owner. In any case, there must be a data owner for every processed data instance to ensure compliance with legal regulations. Having said this, merely finding and assigning a data owner to each dataset also constitutes an issue. The individuals in question have to be in a position to take the responsibility, it also constitutes another additional task, and when the data-driven culture is not yet fully established, employees often do not understand what this responsibility entails and are therefore

hesitant to assume this role. Concerned with establishing and maintaining data ownership and control over data throughout the exchange processes, this challenge is related to the data governance problem in data marketplaces, identified by Driessen, Monsieur, and Van Den Heuvel [DMV22].

Furthermore, *preventing the flooding of the EDMP with unusable data* also represents a challenge. This issue is especially pronounced in the EDMP as it should register the overall scope of data available in a company. Externally, data is predominantly registered in data marketplaces with the intention of generating profit. For this to succeed, the selected datasets usually meet a certain data quality or are relevant for some known use case. Company-internally two scenarios emanate from the data democratization goal of publishing as much data for as many users as possible. For one, there is the targeted provision of data for known use cases. In this case, the relevance of the data, as well as the processing state in which this data is required, are known. In the second scenario, data is provisioned without knowing if it is relevant for other participants and in what form they would need it. This bears the risk of flooding the marketplace with data that nobody needs or that is unusable for further processing. This challenge is closely related to the topic of providing data in a way that increases the consumer's utility by considering the consumer's needs, as addressed in, e.g., [FSF20], and the challenge of providing high quality data as described in [DMV22].

Finally, *integrating the EDMP into the existing system landscape* can be challenging. Different tools may support different metadata exchange standards that must be supported or an alternative standardization for the inclusion of metadata must be provided. In addition, the marketplace must be able to display metadata dynamically since the tools may provide a variety of different metadata per dataset. In this context, implementing an EDMP is more complex than those that function as stand-alone marketplaces.

These challenges are particularly pronounced in the company-internal context. Nevertheless, general challenges identified for designing data marketplaces which are not specifically related to the company-internal context, like transaction enforcement, achieving trust, or data transformation [DMV22], may also apply to the EDMP.

4.7 Summary

In this chapter, we have established that the Enterprise Data Marketplace 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 and challenges that are specific to the EDMP. We have thereby provided a comprehensive understanding of this marketplace type which in turn presents the basis for designing an architecture and concepts for integrating an EDMP within a company, as discussed in the following Chapter 5.

Chapter 5

Enterprise Integration and Platform Architecture

One of the main defining qualities of the Enterprise Data Marketplace (EDMP), as opposed to external data marketplaces, is that it can leverage the company's existing system and tool landscape. For this, it must be embedded in this landscape and be able to interact with the tools and systems therein. Literature thus far does not provide an in-depth discussion on how an EDMP has to be built to incorporate these aspects. In this chapter, we therefore discuss how to design an EDMP so that it embeds itself in this landscape. To begin with, an overview of the intended tool and system interactions is presented in Section 5.1. What needs to be considered to enable this integration is discussed in Section 5.2, and an according platform architecture that takes these aspects into account is presented in Section 5.3. Section 5.4 summarizes this chapter.

By addressing these topics, we establish the architectural basis for building an EDMP, according to research goal three (RG3). The research contributions provided through this chapter thus constitute an enterprise integration architecture (C3.1) and the platform architecture for an EDMP (C3.2).

This chapter is a revised and composite version of the author publications [EGH+23; EGH+22b] and [EGHS22].

5.1 Enterprise Integration Architecture

In this section, we explain how the EDMP can integrate into a company's existent system and tool landscape, as depicted in Figure 5.1, and how this integration can be advantageous for the company. This is distinctive for the EDMP, as stand-alone marketplaces, such as those that are used for trading data between companies, are usually not tightly 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 heterogeneous system landscapes that the marketplace would have to be able to reflect. The typical enterprise integration scenarios are derived from our previous work [EGG+21a; EGH+22a; EGHS22].

Only a few architectures presented in literature consider the marketplace in the context of a company's internal system or tool landscape. Gröger [Grö21] presents the core elements of a data ecosystem with an EDMP but states that implementation and integration aspects have yet to be investigated. Wells [Wel18] roughly highlights which technologies are needed within the marketplace components, according to him these involve components for data lake management, data pipeline management, data catalogs, and data preparation. How the marketplaces can interact with existing tools that already implement these technologies is not discussed. Therefore, we address this topic in this section. First, we discuss the integration with data sources in Section 5.1.1, secondly, the integration with various existent tools in Section 5.1.2, and lastly, we highlight the integration advantages in Section 5.1.3.

5.1.1 Integration with Enterprise Data Sources

To begin with, we would like to illustrate how the marketplace will integrate with the enterprise source systems and reference the data therein. This does not refer to the integration of data in these source systems, but rather the connection between the marketplace and these systems. According to the industrial use case presented in Section 3.1, companies may employ a wide variety of data sources, such as operational systems, e.g., Enterprise Resource Planning (ERP) systems, and analytical systems, such as data lakes. These source systems are illustrated in Figure 5.1 on the bottom. The marketplace

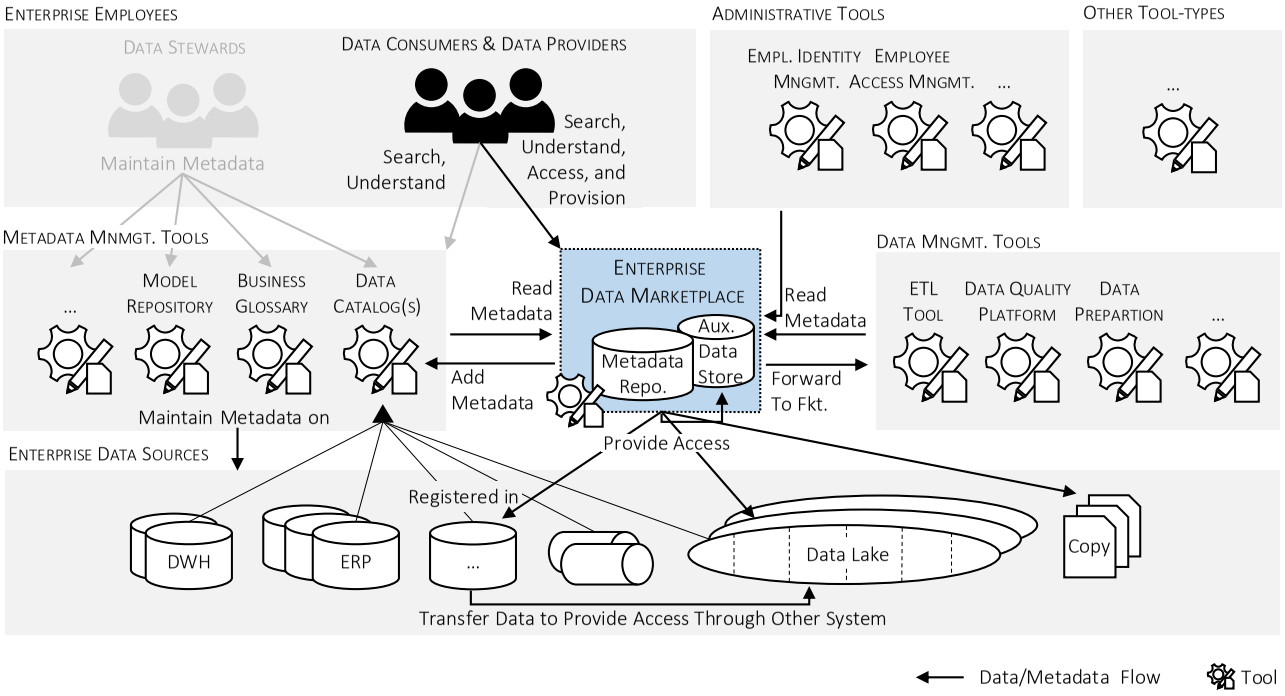


Figure 5.1: Integration of the Enterprise Data Marketplace in a Company's Existent System Landscape [EGH+23].

should register these source systems and reference the data therein so that the data they host can be offered through the marketplace. As companies are in the process of setting up and maintaining data catalogs [EGG+21a], we illustrate in Figure 5.1 that these source systems are already registered in one or more data catalogs. If data catalogs exist within the company, then the marketplace can reuse the existent data inventories and metadata on the data sources and thereby reference the source systems via the data catalog. As discussed previously, only data that cannot be referenced in any external system should be loaded and stored in the marketplace's auxiliary storage to avoid redundancy. The marketplace must either maintain an additional data inventory for the data hosted in the auxiliary storage as an extension of the existent data catalogs or register this data in an existing data catalog. In cases in which access to data in the source system is not possible, the data can be transferred into another system such as a data lake through which it can then be accessed. The catalogs or marketplace inventories have to be updated to register the data's other locations in addition to its original source. The marketplace then knows both systems and grants access to the data in whichever system is better suited to both the needs of the data consumer and data provider. Hence, the marketplace knows the data sources via an inventory which can be provided by existing data catalogs.

5.1.2 Integration with Administrative and (Meta)Data Management Tools

As discussed in Section 2.3.3 on Metadata Sources, a company has a variety of tools and systems that provide metadata, and as discussed in Section 4.5, some of these tools provide functionality that is partly required in a data marketplace. This includes tools for managing data and metadata and administrative tools. Figure 5.1 indicates how the marketplace can interact with these tool groups.

As the EDMP is a metadata-driven tool [Grö21], most of its functionality is based on metadata. An example of this is the data inventory, which consists of metadata that describes the available datasets with information such as the storage location. Apart from the auxiliary data store, the marketplace does not interact with the actual data, only with the according metadata. As can be seen in Figure 5.1, metadata is collected and maintained in the company through *metadata management tools* such as data catalogs, business glossaries, and

model repositories. This metadata is relevant in the selection process for a dataset as demonstrated in the consumer journey in Section 3.3.1. As also discussed in Section 3.3.2, the distribution of metadata across a wide range of tools constitutes a challenge for data consumers in the process of finding relevant data. For this reason, the marketplace requests the metadata from these tools and provides it in an integrated view. This is a read-only process on these tools, data catalogs are an exception in this context. Since an inventory of data records is already maintained in the catalog, the marketplace builds on this inventory, i.e. when new data is registered in the marketplace, it creates an entry in the existing data catalog for the new dataset, and thus performs a write operation. Although the marketplace reads or 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 data catalogs, with their metadata maintained through both the marketplace and catalog. Therefore, the introduction of the marketplace does not change the entire metadata management workflow and the marketplace 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 marketplace (more details on this in Chapter 6), it is still possible to view this metadata in the individual tools.

There are also *data management tools* that collect metadata. These include ETL tools that can reflect data lineage, or quality management platforms that, amongst other things, collect quality metrics such as a dataset's completeness. As with the other tools, the marketplace can extract metadata from these tools and provide it as part of an integrated view if these are of interest in the data selection process. Furthermore, as explained in Section 4.4, 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[1], or refer the consumer to tools with required functionality like data preparation after data acquisition.

In addition to the data and metadata management tools, the marketplace is integrated with *administrative tools* for, e.g., identity management to enable features like single sign-on. Thereby, employees only need to acquire the rights to access the marketplace and the marketplace can then extract employee

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

information from these tools. Based on the extracted information, it can, for instance, display only those records that match the employee's clearance level.

5.1.3 Enterprise Integration Advantages

Integrating the EDMP in the enterprise system and tool landscape as proposed in the previous sections has several advantages. For one, *existent functionality is reused*. By building on the existent tools, the marketplace does not duplicate functionality such as access rights management, which also avoids the marketplace 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 marketplace this provides holistic information on the data. It is, however, important to note that integrating the marketplace with metadata management tools, as well as the integration of the metadata itself, are a complex topic that elicits a variety of challenges including the classic data integration problems, as described by Leser and Naumann [LN06]. Another advantage of integrating the marketplace in the enterprise is a *reduction in metadata management effort and errors* for data providers. By reusing metadata that has already been collected within other tools, there is no additional effort for maintaining a redundant set of metadata in the marketplace. 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 [EGHS22]. Finally, there is *less redundant data*. When referencing data within the data sources as opposed to uploading the data redundantly in the marketplace, there is less effort on behalf of the providers, reduced storage cost, and no synchronization efforts.

5.2 A Distinction of Data Assets and Data Products

By integrating with the enterprise system and tool landscape, the EDMP provides a variety of metadata for the registered datasets. Ultimately, it is the goal to make these datasets accessible through the marketplace and that these can be traded like a product. However, merely collecting metadata does not suffice to turn a dataset into a data product. Building data products entails a

range of activities for data providers. As described in the data provider journey in Section 3.2.1, this also involves preparing provisioning options and in extension thereof the specification of metadata on these provisioning options, as well as on contractual information, and the terms of use. This means that integrating the marketplace with data sources and tools does not automatically turn data into a data product. Furthermore, turning all the data in the source systems into data products would present a considerable effort for the data provider and is unfeasible. As it is nonetheless the goal to democratize the full scope of a company's data, an approach is required which supports data consumers in finding and understanding all data without the data providers having to build the full scope of data products. For this reason, we introduce a differentiation of data assets and data products in Section 5.2.1, reveal how the marketplace deals with these in Section 5.2.2 and how data assets are turned into data products in Section 5.2.3. Finally, we discuss the benefits gained through this distinction in Section 5.2.4.

5.2.1 Defining Data Assets and Data Products

Figure 5.2 illustrates how data, data assets, and data products are related. At the center is *data* representing the dataset in question. As the term suggests, data becomes an asset when it has a potential financial value for a company [Spi19]. The *data assets* comprise the data as well as according metadata which enables finding and understanding the data, such as the content description, lineage, or data owner. While the data is maintained within the source system, the data assets are registered in tools like the data catalog [LLEF20; ZDED17]. Additional asset metadata like quality metrics can also be maintained throughout a variety of (meta)data management tools. Data assets are usually only available to very few employees, e.g., those who work in the team hosting this data asset and work on a similar topic. *Data products* are data assets that have been prepared for access and provisioning and have been enriched through the data marketplace with according product-specific metadata. This definition and understanding of data products is similar to that of Driessen, Monsieur, and Van Den Heuvel [DMV22], who define data products as data assets which have been prepared for consumption through a data marketplace. As mentioned previously, this involves, amongst other

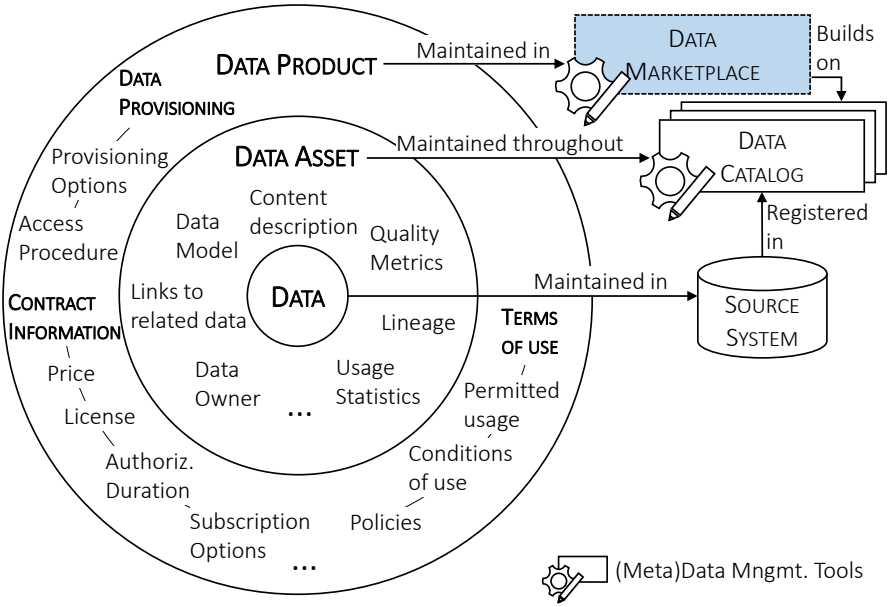


Figure 5.2: The figure illustrates the distinction of data assets and data products with exemplary metadata, as well as the systems in which these are maintained. Metadata which are connected through dashes belong to a specific topic that is portrayed though capital letters. (Based on [EGHS22])

steps, the preparation of provisioning options and enrichment with according metadata. As shown in Figure 5.2 this metadata includes, for instance, details on these *provisioning options* such as an API, download or source system access, and the according access procedures that go with these options. Also, *contractual information* such as the price, if data is monetized, the license, or subscription options, as well as the *terms of use* such as the permitted usage or conditions of use constitute product metadata. Metadata to both the data asset and data product belong to the dataset-specific metadata in the functionality framework as presented in Section 4.4.2 (see Figures 4.2 and 4.4.) While the data assets are registered in a data catalog, the data products are registered and maintained through the data marketplace.

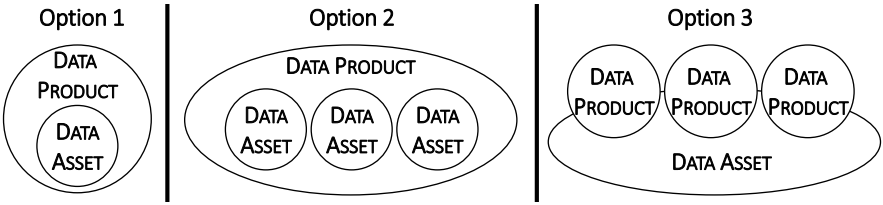


Figure 5.3: Options for Scoping Data Products.

How a data product is scoped can vary depending on the data. Figure 5.3 depicts three options for scoping data products: Firstly, a data asset could become one data product, as also illustrated in Figure 5.2. Secondly, several data assets could be bundled into one data product if, e.g., these are often requested and used together. Lastly, a large data asset might be split into several data products, each containing a part of the asset.

5.2.2 Supporting Data Assets and Data Products in the EDMP

The marketplace should support both data assets and data products. If the marketplace integrates with the existent data catalogs as suggested in Section 5.1 and uses these as its inventory, all registered data assets can also be found through the marketplace. Data consumers will have the option to request access to data assets through the marketplace. For this, the provider must, however, first turn them into a data product. The marketplace stores the product metadata in its metadata repository and therefore recognizes data assets that have already been turned into data products. Like with the data assets, data consumers can request access to these, yet in this case no additional effort on the data provider side is required. For requested data products, either the data owner, or another responsible person such as the data provider, only have to deny or accept access requests so the consumers can gain access and start working with the data. Concluding, data assets can be found and maintained both through the data catalogs as well as the data marketplace. The marketplace, however, provides a comprehensive search over all data assets registered throughout all catalogs in addition to the registered data products.

5.2.3 Turning a Data Asset into a Data Product

Data assets can be transformed into data products in different ways, as depicted in Figure 5.4. In this sense, we illustrate three main transformation scenarios. Within the first scenario, the provider has already prepared the asset for consumption, and therefore *explicitly registers a data product in the marketplace* and directly specifies all the product metadata. By implication, if this asset is new and has not yet been registered in the data catalog, the marketplace then also registers the according data asset in a data catalog. In the second scenario, the data provider *registers the data asset within the catalog* and does not deal with the data marketplace. Some employee, e.g., a data scientist, can at this point search for data in the marketplace and would also find this data asset. The employee can then send a request to access this data to the provider who is prompted to turn it into a data product and specify the additionally required product metadata. Having turned the data asset into a data product, access can be granted to this data. The third scenario assumes that provisioning options already exist and *another employee can fill in the required product metadata* and send a request for asset-product transition to the provider. For instance, a data steward may know that the data asset already has provisioning options like an API and can fill in the product metadata for the data provider. The provider or data owner are notified and can accept or reject the proposed asset-product transition. If accepted, the data asset is

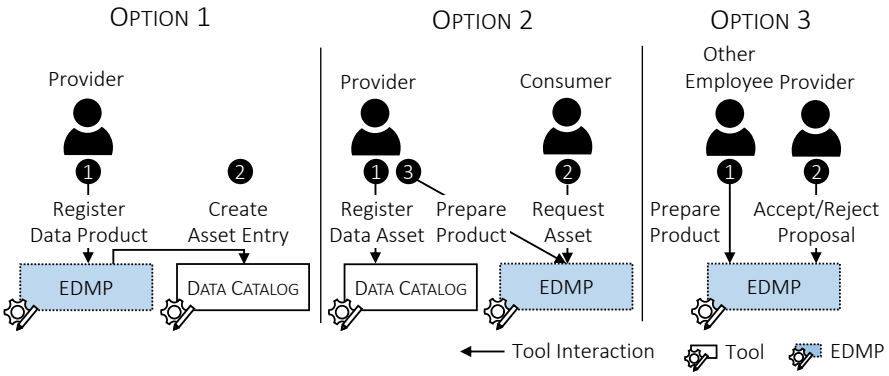


Figure 5.4: Options for Turning Assets into Products.

turned into a product, otherwise, it remains a data asset that cannot yet be accessed.

5.2.4 Advantages of the Asset Product Distinction

The first advantage is that the *marketplace can offer a company-wide scope of data* if it supports data assets in addition to data products. Not all data assets will be turned into a data product, yet it is still the goal to also democratize the data assets. Secondly, by integrating in the enterprise tool and system landscape, the data marketplace can *offer data assets even if these have only been registered in other tools* like a data catalog, and not in the marketplace directly. This yields the third advantage, that *providers merely have to register a data asset* so it can already be found and understood. This is initially less effort for the provider since it is not necessary to build or register a complete product on the marketplace, which requires a few more steps in contrast to registering a data asset. Similarly, *the provider only has to register the data asset or data product in one tool*, i.e., in the catalog or marketplace, as opposed to both. Lastly, this distinction enables data providers to register data assets and wait to see if these are actually requested through the marketplace. The provider, thus, *only has additional effort in building a data product if this data asset is actually relevant* to other data consumers. Besides alleviating the data providers in their task, this distinction also addresses the EDMP challenge of flooding the marketplace with unusable data, as presented in Section 4.6. For the scenario of a targeted provision of data for a known use case, an according product is built. For the scenario in which data should be made available so other consumers know it exists, it suffices to create a data asset until this asset is actually requested and further consumer requirements on the data are clarified.

So far, we have shown how the marketplace can fit into and interact with the corporate tool and system landscape. To facilitate this, the distinction between data assets and products was introduced. In the following, we will discuss how these aspects affect the platform architecture of the EDMP.

5.3 Platform Architecture

Having discussed how the EDMP can integrate with a company system landscape leaves the question of how this influences this type of marketplace's platform architecture.

In literature, there are a variety of architectural proposals, most of which are, however, tailored to a specific context. There are marketplace architectures specific to the use of blockchain [RS16; SLR20], the Internet of Things (IoT) context [KPKS18; ALL20; SBK+17], multilateral marketplace design [KLT17], elements in decentralized marketplaces [RRK18], personal data valuation [KGC+16], or also specific marketplace aspects like a market management system or mashup builder [FSF20]. None of these architectures reflect the specific components of the EDMP. In contrast, Wells [Wel17] does provide a component overview for the EDMP, nonetheless, it is not apparent which aspects are special to the internal setting and also how the components interact. Similarly, components of an EDMP's underlying data platform are illustrated in the report by Zasadzinski et al. [ZTTR21], yet the distinction between the marketplace and data platform components is not clear, nor how it leverages and is embedded within a company's existent system landscape. Therefore, marketplace architectures presented in literature thus far provide various perspectives on required components and the component interactions. Yet, the architectures do not comprehensively consider the specific characteristics and requirements of an EDMP as described in Sections 4.3 and 4.4.

We present a platform architecture that reflects the components of an EDMP, displayed in Figure 5.5. Components that are potentially distinctive in the EDMP, e.g., in regard to implementation aspects, are highlighted in gray. The architecture distinguishes *frontend* and *backend* components. The frontend is responsible for offering functionality to the marketplace 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 exploited by tightly integrating the EDMP with

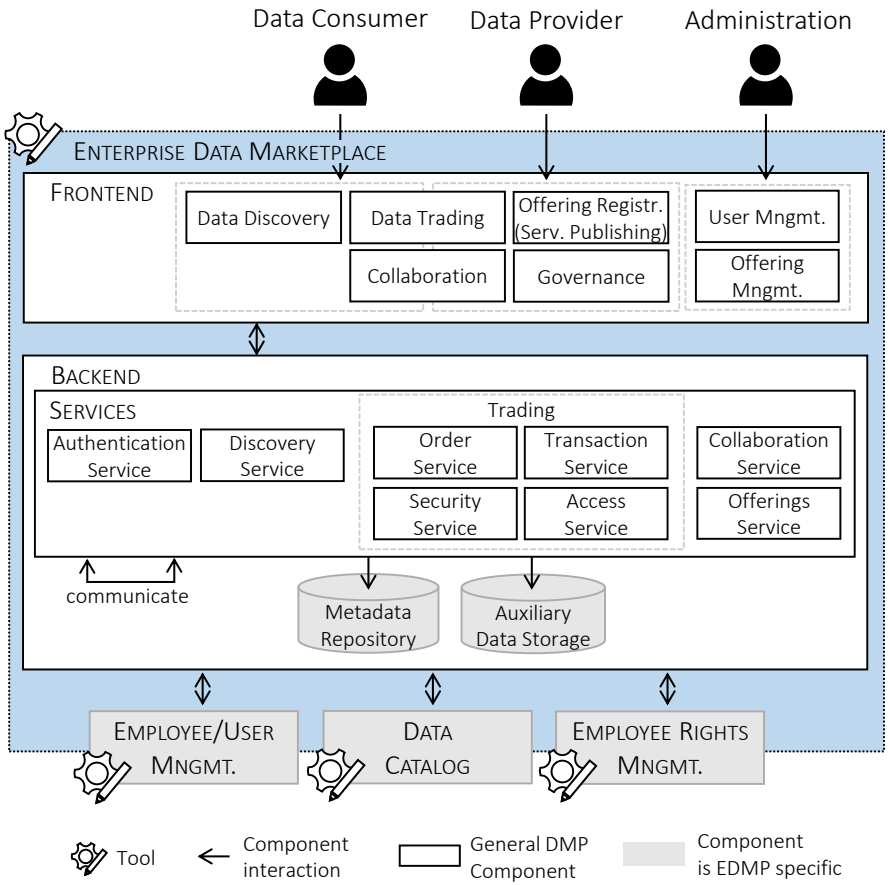


Figure 5.5: Enterprise Data Marketplace Architecture Featuring a Component Overview [EGH+23].

the existent solutions as discussed in Section 5.1. Alternatively, the features of these components can also be implemented within the according backend services, yielding a self-sufficient data marketplace.

5.3.1 Frontend

The marketplace functionality is available to the roles data consumer, data provider, and administrator in the frontend through, e.g., a graphical user in-

terface and/or an API. It includes the functionality as described in Section 4.4.2. 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 or privacy and security, is not directly accessible to users, it is not represented in the frontend. These are addressed indirectly throughout the backend services.

5.3.2 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 marketplace 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 tasks 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 database access, or access-links to data. The offerings service is responsible for the registration of any kind of service as described in Section 4.4.1, i.e., data-as-a-service, infrastructure-as-a-service, software-as-a-service, and professional services like training courses. It adds the data offerings to the data catalog, which maintains a data inventory, and stores additional metadata like the product metadata in the metadata repository. An inventory of the other offerings, i.e., infrastructure, software, or professional services, can also be stored in the metadata repository if the data catalog does not support these kinds of entries. Lastly, the collaboration

service takes care of any form of interaction on the offerings, such as comments, use-case-documentations, or ratings.

5.3.3 Enterprise Data Marketplace Specific Components

The components highlighted in gray in Figure 5.5 are required in all marketplaces, but can be specifically adapted to the enterprise setting, and are therefore termed as EDMP specific components. For instance, the components marked as tools can be implemented as part of the marketplace, producing a stand-alone solution that could be used in an external context. However, these components can already exist within an enterprise, and could therefore also be reused and integrated with the marketplace.

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 which the user information is managed. In terms of the data democratization dimension two, getting access to tools 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[2] which offers a directory of employee profiles and functionality to structure and secure employee data, including personal information, qualifications, and skills. As the marketplace requires 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 rights management*, which handles the users authorization, i.e., rights for various tools and platforms and potentially specific actions therein. Through it, users can apply for, attain, and manage these rights. Like before, there are tools for this on the market that are already used within enterprises, such as Access Rights Manager[3] and could be integrated into the marketplace. This also enables facilitating single-sign-on so the users can access the marketplace through their employee user account.

As mentioned before, the marketplace requires an inventory of its offerings which would reflect, e.g., the offered data or services like training courses,

[2]Employee DB Software: https://scnsoft.com/software-development/databases/employee
[3]Access Rights Manager: https://solarwinds.com/de/access-rights-manager

with according metadata like a content description, the owner, and access policies. This inventory can be maintained as part of the marketplace's metadata repository or could be maintained within an external tool like a *data catalog*. As companies are in the process of building and maintaining data catalogs [EGG+21a], the stored information could be reused within the marketplace as opposed to duplicating the inventory with collected metadata and functionality. As explained in the previous Section 5.2, a data catalog mainly stores information on data assets, so the marketplace must either extend the stored information through the product metadata or store additional product metadata in its metadata repository.

The *metadata repository* stores metadata that is relevant for operating the data marketplace. As data marketplaces are metadata-driven platforms [Grö21], this is an essential component. What metadata is maintained in the EDMPs varies depending on whether the above-mentioned tools are integrated in the marketplace, 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 Section 4.3.2, an EDMP may have a hybrid architecture with both a centralized and decentralized data storage. Most of the offered data should be referenced in the according storage systems, in order to support the whole scope of 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 into the integrated *auxiliary data storage* of the marketplace. 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 marketplace's distinctive components constitute an independent tool or have to be implemented as part of the marketplace also depends on the existing system and tool landscape in the company. According to Jahnke and Otto [JO23], an EDMP of this sort is not represented throughout the current commercial or open-source tools.

5.4 Summary

This chapter illustrated how an Enterprise Data Marketplace can be embedded in the existent corporate tool and system landscape. Specifically, the architectural view on how it integrates with the data source systems, data and metadata management tools, as well as administrative tools was discussed. Amongst others, the advantages of embedding the marketplace in this landscape entail the reuse of existent functionality and metadata, the enabling of a comprehensive view on existent metadata, and the reduction of metadata management effort and errors. However, having metadata for a dataset is not enough to turn it into a data product in the marketplace. For this reason, the differentiation between data assets and data products was introduced, which are both supported in the EDMP. This differentiation yields a number of advantages such as a reduced effort for data providers in sharing their data and also the ability of the data marketplace to offer a company-wide scope of data. Finally, a platform architecture for the EDMP was introduced that reflects these aspects and specifically identifies components that may be implemented differently within the enterprise. These include components for managing user data and access rights, a data catalog, the metadata repository, and lastly, an auxiliary data storage. The presented enterprise integration architecture and platform architecture provide the basis for building an EDMP. Additionally, these architectures also provide the basis for the design of an explicit approach to leverage existent metadata as presented in the following Chapter 6.

Chapter 6

Leveraging Distributed Metadata in the Enterprise Data Marketplace

Metadata, as outlined in Chapter 4, is an essential constituent in the data marketplace and is required to provide a lot of the data marketplace functionality. For instance, the registry of data assets and data products offered through the marketplace is composed of metadata, and also the assembly of dataset-specific metadata on each asset and product provides the basis for data consumers to identify whether this data is relevant for their use case. In the preceding chapters, it has been established that a company already has a lot of relevant metadata for the marketplace distributed across a tool and system landscape. This gave rise to the requirement presented in Section 4.4.3 that the Enterprise Data Marketplace (EDMP) should build on and reuse the existing functionality and metadata within this tool and system landscape. In Section 5.1, we highlighted the architectural perspective of integrating the EDMP in the company system landscape. This decreases redundancy and reduces the complexity and workload for both data consumers and producers, as these only have to

access one tool as opposed to several in order to identify relevant datasets or to maintain metadata on these datasets respectively. With the introduction of new concepts like the data lake or data mesh, the tool and system landscapes, and thus the collected metadata, are continuously evolving. Besides the ability to flexibly integrate with this evolving tool and system landscape, the EDMP must, consequently, also have the ability to handle newly acquired metadata through these tools and systems.

While the need for metadata management in data marketplaces is emphasized repeatedly in literature [MS19; RRK18; KLT20; SLR20], most research articles only briefly address these metadata management aspects. In Section 5.1, we addressed the architectural metadata perspective, yet, it still remains unclear how to explicitly realize the embedding of an EDMP in the company system landscape to reuse the existent metadata and how to realize the metadata management. Therefore, we close this gap by presenting a three-part approach that enables to leverage metadata that is distributed across a company in the EDMP. To begin with, we present the first approach for integrating different tools with the marketplace, addressing research contribution C4.1. Thereafter, we introduce an approach for flexibly supporting a diverse set of metadata per data asset in the marketplace, based on adaptable metadata templates. Lastly, we illustrate an approach for visualizing an integrated view on this diverse metadata in the marketplace. In conjunction, the last two approaches yield research contribution C4.2.

The remainder of this chapter is structured as follows: To begin with, the diversity in metadata collected for a single dataset and also the diversity in metadata collected throughout the metadata sources is underlined in Section 6.1. Thereafter, metadata management requirements for integrating the EDMP in a company system and tool landscape are specified in Section 6.2, considering the diversity of metadata and the metadata sources. Related work is discussed in Section 6.3, followed by our approach for leveraging distributed and diverse metadata in the EDMP in Section 6.4. Thereafter, Section 6.5 yields an assessment of how the presented concepts address the specified requirements and how classic (meta)data integration challenges are addressed. Section 6.6 rounds this chapter off with a summary and conclusion. In parts, this chapter is based on the author publications [EGG+20] and [EGG+21b].

6.1 Diversity in Metadata

In Section 5.2 it is explained that data product metadata like provisioning or contractual information is maintained within the data marketplace, whereas the data asset metadata like content descriptions or quality metrics can be spread across a variety of tools. Given the focus of this chapter to leverage this distributed metadata, we will mainly discuss how the data marketplace deals with data assets in the following. We implicitly assume that it is possible that the marketplace hosts product-specific metadata for these data assets and that these could therefore also be data products. In this section, we outline how diverse a collection of metadata for a single data asset can be, and also how tools and systems can be different in collecting and exchanging metadata. Since it is a requirement that the EDMP must take this metadata into account, the marketplace must be able to accommodate this diversity.

6.1.1 Diverse Metadata per Data Asset

A diverse suite of metadata attributes can be collected for each data asset. This entails different *types of metadata* as described in Section 2.3.1, such as business, technical, and operational metadata [HES17]. These may be a content description, information on the storage, and the data owner, respectively. Furthermore, different metadata attributes can be collected per data asset depending on the *data type*, e.g, if it is a video, text, or image, or also depending on the *storage system*, e.g., if it is a data warehouse or data lake. Moreover, *different attributes may be collected for similar datasets*, e.g., quality metrics are collected for one data asset yet not for the other. For instance, a video collection could be annotated with business metadata that indicates that these are traffic recordings, with technical metadata that describes the storage location of the data on the company's data lake, and operational metadata stating that the user "Ann Li" is the data owner and that the videos may only be used by the division "HX" for research on autonomous driving. It could also have video-specific technical metadata indicating the video format, the resolution, and the duration of the individual video clips. Another video collection asset could have additional video metadata such as the bit and frame rate, display aspect ratio, and audio information.

Handling diverse metadata per data asset thus encompasses the collection of varying metadata according to the data type, e.g., video, different metadata types, e.g., business metadata, and also the accumulation of different metadata among similar data assets. When searching for a dataset, all of these different metadata can play a role in the consumer's choice and are thus potentially relevant in the marketplace. The marketplace consequently has to gather and integrate these diverse metadata from a variety of systems which, as emphasized in the following section, also poses a challenge.

6.1.2 Diversity in Metadata Sources: Exemplified through the Data Lake

As explained in Section 2.3.3, a variety of tools and systems constitute metadata sources. These range over data and metadata management tools like data catalogs and business glossaries to different data storage systems like data warehouses or data lakes. Each of these provides different metadata and, as also pointed out in Section 2.3.4, might support a variety of different standards for exchanging metadata. To highlight how distinct the single tools and systems can be, we demonstrate how metadata can be collected within a data lake.

A data lake, as briefly introduced in Section 2.3.3, is a highly scalable storage repository which enables data analytics in order to exploit the business value of data [GGH+20a]. Metadata management is required to retain the data's context and thereby prevent a data lake from turning into a data swamp, a state in which the data therein is no longer fit for use [SD20; CSN+14]. Several metadata models are discussed in literature which support the modeling of metadata for a wide range of metadata management use cases in data lakes including MEDAL [SSF+19], GEMMS [QHV16], and our metadata model HANDLE, short for "Handling metAdata maNagement in Data LakEs" [EGG+20].

Figure 6.1 depicts an exemplary metadata management use case. In this use case, the data, represented by a *customer table*, resides in a data lake that is divided into several zones in accordance with the zone model by Giebler et al. [GGH+20b]. The metadata collected in this use case is centered around access information on the data in the data lake. As can be seen, the access information illustrated as yellow/dashed metadata objects is collected on different granularity levels in the two different zones, i.e., once per row in the *raw zone* and once per table in the *trusted zone*. This distinction of granularity

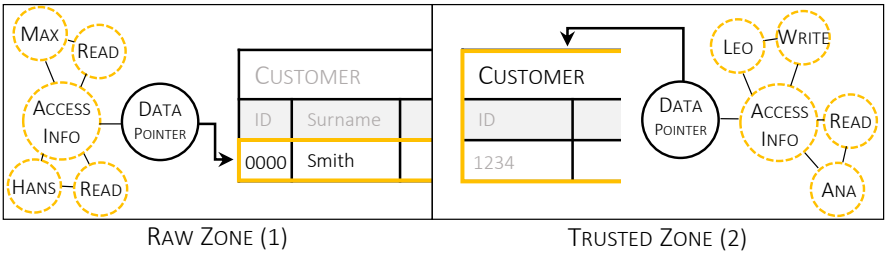


Figure 6.1: Exemplary Extract of a Metadata Management Access Use Case [EGG+20].

levels shows that even within the data lake the same information might be collected on varying parts of the data. For instance, access on the overall table might suffice if it has been anonymized, and the row based access might be required if it is raw personal data, such as the customer data in this example.

The metadata model HANDLE provides a foundation for modeling the metadata in a large variety of such metadata management use cases, collecting and also retroactively extending the according metadata if specific information was missing. It consists of two parts: firstly, the core model, shown in Figure 6.2, which defines all the elements and relations required for modeling a metadata management use case. Secondly, it consists of the three core model extensions, which need to be adapted for each data lake implementation and address the topics of granularity, data lake zones, and metadata categorizations.

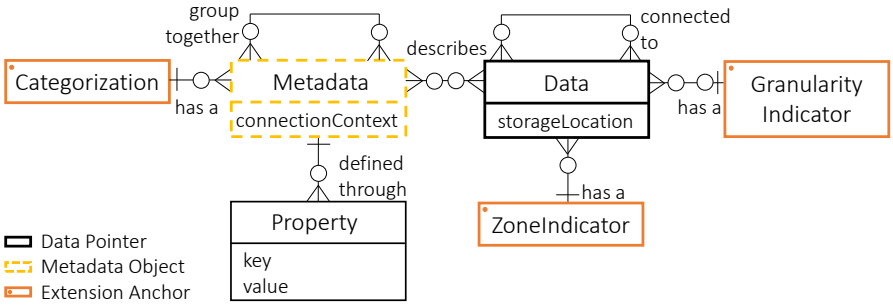


Figure 6.2: The HANDLE Core Model [EGG+21b].

The two central elements in the core model are the *data* entity, illustrated in solid/black, which represents a pointer to the data in the lake, and the *metadata* entity which can represent any number of metadata attributes, illustrated in dashed/yellow. The *zoneIndicator* is part of the zone extension and can be viewed as a label attached to the data element, indicating the zone the data is stored in. Any zone model, such as the one proposed by Giebler et al. [GGH+20b] can be appended here. The *granularityIndicator* points out to what level the metadata belongs in the dataset, e.g., whether the metadata is associated with the overall table, or more specifically with a row, column, or field. Similar to the zoneIndicator, any granularity level can be used in this model. The *categorization* indicates the type of metadata, e.g., business, technical, or operational, and can also be adjusted to use any other metadata type distinction. The model is designed on a high abstraction level so that it can flexibly model any use case and also allows to model the metadata according to the intended employment. Hence, it is possible to model the same content in various ways, as depicted in Figure 6.3, through the three options a), b), and c). More detailed information on the model and a comparison with other metadata models can be viewed in our works [EGG+21b; EGG+20].

A data lake will have a variety of metadata management use cases, which can be reflected through HANDLE in various ways. Some of this metadata might

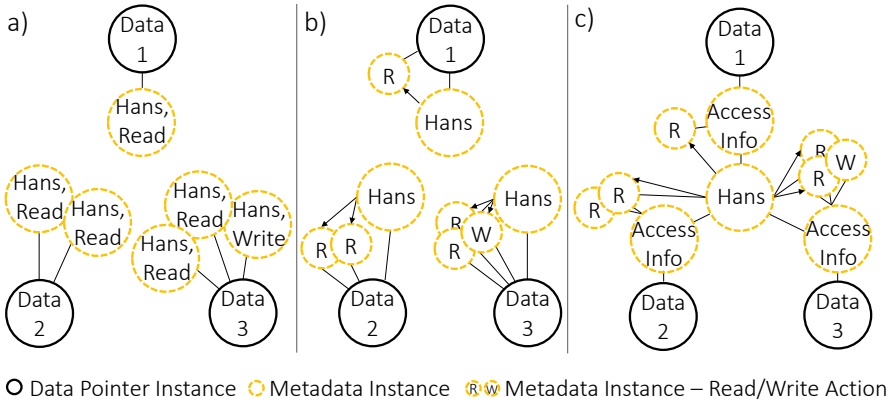


Figure 6.3: Various Modeling Options for the Same Use Case - Logging all Access Calls on Various Datasets (Based on [EGG+21b].)

be of relevance when looking for and selecting data for analytics use cases and hence is also relevant for the data marketplace. Consequently, the data marketplace must be adaptable to incorporate the metadata as provided through the data lake management system according to its model, in this case, HANDLE. By discussing metadata management in data lakes based on the model HANDLE, we have demonstrated not only that different metadata management use cases may be dealt with in diverse ways, but also how distinct different tools and systems may be in collecting metadata. Since the data lake is only one of many metadata sources, this means that the marketplace needs a sophisticated approach to connect to different tools and systems and incorporate a wide variety of metadata. To this end, requirements are established in the following section which must take into account this diversity in the metadata, as well as the diversity of how the tools and systems collect and provide this metadata.

6.2 Requirements for Integrating with Data and Metadata Management Tools

As the previous sections have emphasized, there are a few aspects that the data marketplace needs to consider when integrating in the enterprise system and tool landscape to leverage the distributed metadata. In order to determine metadata-related requirements for integrating the EDMP in this landscape, we first studied related challenges as presented in literature [DMvdH23; ALL20; RRK18; JCZ12]. Thereafter, we also conducted more than ten expert interviews with employees of the globally active manufacturer introduced in Section 3.1 in various data-related roles, such as data scientists as well as enterprise and solution architects, to gain a broad and representative perspective on current issues and practical insights. The manufacturer's case and the requirements are illustrated in Figure 6.4.

As explained in Section 3.1, the manufacturer pursues the goal of becoming data-driven, and is currently in the process of democratizing their data assets by building a tool landscape with tools like data catalogs, business glossaries, and a data quality platform. As pointed out in Section 6.1, the metadata collected by these tools may strongly vary. The company is also investigating EDMPs. In accordance with the challenges discussed in Sections 3.2.2 and 3.3.2, the

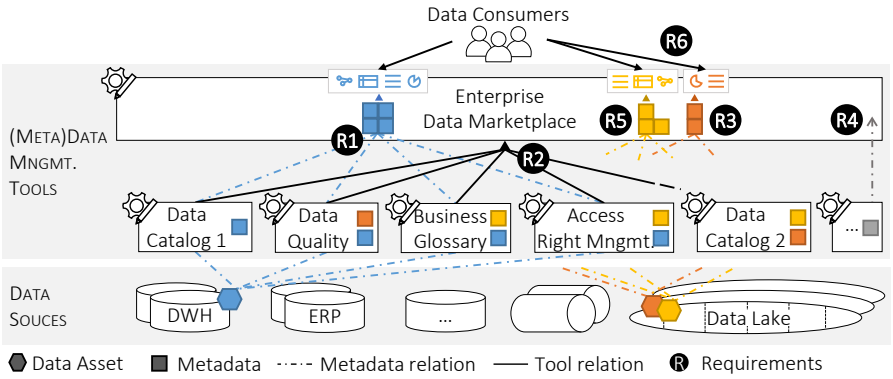


Figure 6.4: Use Case System Overview and Requirements.

multitude of tools and distributed metadata constitute a challenge as they make the process for providers to publish data assets and the process for consumers to understand and access data assets increasingly complex and time-consuming. For instance, the provider must publish their data in several tools like the marketplace, a data catalog, and a quality tool, and the consumer must access all these tools to identify a relevant data asset for their use case.

Based on the literature study as mentioned before, the practical insights from the manufacturer's case, and the provider and consumer challenges, we derive the following requirements for handling metadata in an EDMP. In accordance with the enterprise integration requirements discussed in Section 4.4.3, the first requirement (R1) signifies that the marketplace must *reuse existent metadata that is maintained within other tools*. I.e., if information about a data asset's owner is maintained in a data catalog and quality metrics in another tool, then the marketplace must leverage this metadata as opposed to collecting it twice. This reduces the workload of the provider as they do not have to redundantly maintain metadata in the marketplace and it supports the consumer as they can use the marketplace as a single point of reference as opposed to several tools. In an evolving tool landscape, it is necessary that *tools containing metadata can be connected to and disconnected* from the marketplace (R2). The marketplace thus has to be able to support different interfaces of different tools. As these tools may collect different metadata

attributes for data assets, the marketplace must also *support a diverse set of metadata per data asset* (R3), e.g., a lineage graph with quality metrics for one asset, and a set of descriptive attributes and a data model for another asset. Furthermore, it must be possible to *extend the set of supported metadata* in the marketplace (R4). To support the consumer in finding relevant data assets, the marketplace should *provide an integrated view* on the metadata (R5) and this view should be able to *visualize an individual set of metadata per data asset* (R6). Hence, if present, metadata like the lineage, quality metrics, or descriptive attributes from the different tools should be unified and integrated, and depending on what metadata is available, the metadata views of similar assets may differ.

6.3 Related Work

In the following, we present how metadata and metadata management functionality are considered in data marketplace research and which aspects are covered specifically in EDMPs and in other metadata-intensive systems outside the marketplace context.

The importance of metadata in the data marketplace is highlighted in several research articles. Some articles illustrate the necessity of metadata on data products so that data consumers can find, understand, and trust the offered data [RRK18; KLT20; SLR20], or to enable interoperability and reuse [DMvdH23]. Spiekermann et al. [STWO18] provide a metadata model for data goods to facilitate the selection and trading of data. While this metadata model is generic, other articles propose domain-specific metadata templates, e.g., for data products in the smart city context [RRK18], so-called model cards [MWZ+19] for describing machine learning models, or datasheets [GMV+21] for describing datasets for training these models. These articles, however, lack an in-depth discussion on how the presented concepts impact the data marketplace and how it may manage this metadata and the metadata templates.

Required metadata management functionality in data marketplaces is discussed by, e.g., Meisel and Spiekermann [MS19], who introduce a reference model for marketplace functionality that lists metadata management amongst

the data integration functionality. The necessity for automatic metadata generation is briefly discussed in [SLR20] and Fernandez, Subramaniam, and Franklin [FSF20] describe a metadata engine in the data market context for reading and maintaining the lifecycle of their datasets. Yet these articles do not consider the implications on the metadata management when using the marketplace in the company's internal context.

Several research articles consider data marketplaces for use within a company [Wel18; Grö21; FSF20; EGH+22a], however, they do not go into detail on how to realize metadata management aspects. In contrast, Driessen, Monsieur, and van den Heuvel [DMvdH23] discuss metadata management for achieving interoperability of data products in an internal data exchange and propose an ontology-based solution approach. Yet, they do not provide insights how their solution and the existent metadata management tools are leveraged in a marketplace platform. In our previous work, we outline that the EDMPs should integrate with metadata management tools [EGH+22b] and discuss how they can build on a data catalog [EGHS22]. In both our works we do not go into detail on how the required metadata management is realized.

Research from other areas also discuss topics surrounding the integration of metadata, e.g., in the context of genomic metadata [BCMC22] or metadata on a buildings lifecycle [FPM+20]. Besides the metadata templates mentioned above, we adopt aspects from these articles resulting in similarities in our integration process to [BCMC22], and we adopt components similar to the drivers used in [FPM+20] for connecting the EDMP to the metadata sources. Some research articles use ontologies to provide clarity on the semantics and to enable interoperability amongst the (meta)data [BCMC22; DMvdH23]. The concepts we propose may also build on ontologies. While these concepts address metadata management aspects, they need to be adapted to fit the marketplace use case in an enterprise. Hence, we close this gap and present an approach for reusing and building upon existent metadata in the EDMP, in which the topics like the integration with a company's tool and system landscape as discussed in the previous Chapter 5 are considered.

6.4 An Approach for Leveraging Distributed Metadata in the Enterprise Data Marketplace

In this section, we present three approaches that in conjunction enable an EDMP to leverage and handle existent company metadata that is distributed across a variety of tools. The first approach addresses the challenge of integrating a variety of tools and systems with the EDMP that contain relevant metadata for understanding and selecting data assets. The second concept enables collecting a diverse set of metadata per data asset and the third concept reveals how this diverse set of metadata can be visualized in a comprehensive view for the data consumers.

6.4.1 An Approach for Integrating Different Tools with the Marketplace

Figure 6.5 illustrates the process for integrating new tools with the EDMP and leveraging the contained metadata therein. As it is the goal to provide a comprehensive and integrated view on the metadata, this process contains aspects of information integration. How the following concepts address typical (meta)data integration challenges is assessed in Section 6.5.2. Moreover, several of the steps we describe hereafter are carried out manually. How these can be automated or supported by tools is subject to future work.

In order to leverage the metadata that is distributed across a variety of tools and systems, the EDMP must be able to integrate with these. Referring to step one in Figure 6.5, we propose a connection to these tools and systems through plugins: The plugins are small software programs that must be configured specifically for each tool or system. Each tool that contains relevant metadata

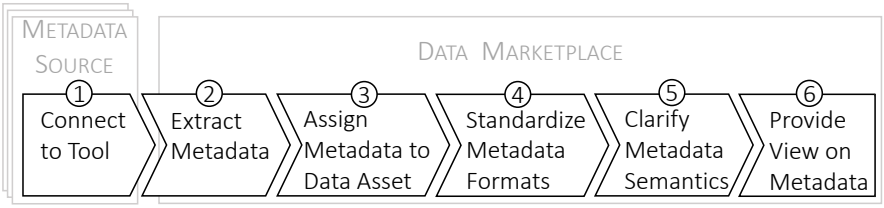


Figure 6.5: Metadata Assimilation Process in the Enterprise Data Marketplace.

for understanding and selecting data assets can be connected via a separate plugin. The benefits of using plugins entail the ability to independently connect and disconnect a variety of different tools to the marketplace. Through the individual configuration of the plugins per tool, the different metadata and specific characteristics of the tools can be addressed. Besides connecting a variety of different tools, it is also possible to connect several instances of the same tool, for example, two instances of the Apache Atlas catalog[1]. In this case, a plugin written for Apache Atlas can be duplicated, so each tool instance is connected through its own plugin instance. Thereby, different metadata attributes can be collected from these two tools. Hence, as the selection and handling of metadata is configured individually per tool, the marketplace can flexibly assimilate a diverse spectrum of metadata.

Thus, to realize the process depicted in Figure 6.5, a plugin is created for each tool instance in step one, *connection to a tool*. Therein, credentials are specified for accessing this tool. As depicted in Figure 6.6, the tools can contain a variety of metadata ranging from simple attributes such as a content description or the owner of a data asset, to more complex metadata like the data's lineage. Not all the contained metadata is necessarily relevant in the data marketplace. Hence, the relevant interfaces for extracting the metadata, as well as an option to enable or disable the collection of specific parts of the metadata, are configured in the plugin. By way of example, Figure 6.6 illustrates the selection of specific metadata through checkboxes for each attribute in the tools. The data marketplace can use the configuration provided through the plugin to connect to the tool and *extract the specified metadata* in step two.

In step three, the marketplace must *assign the extracted metadata to a data asset*. As explained previously in Section 5.2, data assets are datasets that are of value to a company and have been enriched with metadata. The marketplace only references the dataset but does collect the associated metadata. In the following, we refer to each asset's metadata collection as an Integrated Data Asset Metadata (IDAM) entity. Thus in step three, the marketplace assigns the extracted metadata to the according IDAM entity. This is necessary as the marketplace may collect a variety of metadata from, e.g., a data catalog, a quality, and a glossary tool, which may belong to different data assets, as

[1]Apache Atlas: https://atlas.apache.org

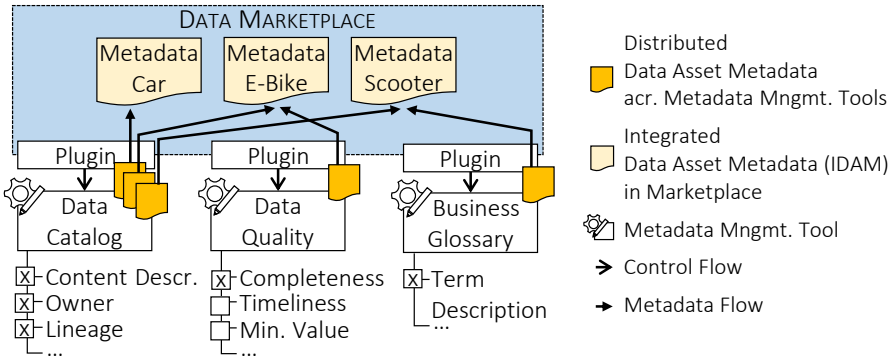


Figure 6.6: This figure depicts a data marketplace that is connected to a variety of tools which contain diverse metadata. Selected metadata attributes are extracted from these tools and grouped in integrated data asset metadata entities in the marketplace.

shown in Figure 6.6. When connecting a new tool, it must be specified in the marketplace configuration how the metadata of the new tool is related to the metadata contained in the other tools. For example, the quality tool may maintain an identifier per data asset that corresponds to the identifier in the data catalog. It is therefore a prerequisite that there is a known relation for the metadata in the new tool to the metadata of at least one of the other connected tools, so this relation can be leveraged by the marketplace to organize the extracted metadata into IDAM entities per data asset. The contents of the different tools could also be semantically associated with each other via an ontology upon which the marketplace could build.

To enable the marketplace to handle the metadata from the different tools in a uniform manner, it must be *transformed into a standardized format*. This can be done by defining a set of both primitive and specialized value types that the marketplace supports, such as *string* or *graph*, respectively. Configuring the plugin therefore also entails matching the tool's metadata to these defined value types and specifying transformation rules. If, for instance, the data catalog's lineage metadata is extracted, this may be mapped to a graph value type and has to be transformed to match the input for graphs as expected by the data marketplace.

To this point, metadata has been selected and extracted from the tool, mapped to a data asset, and standardized. It is, however, still unknown what the metadata represent. It may be known that a metadata field yields a string, but if it represents the content description or perhaps the data owner has yet to be determined. Thus, step five involves *clarifying the semantic meaning* of the extracted metadata. A strategy for dealing with the assignment of semantics to the metadata, how the marketplace can support a variety of semantic input, and how it can visualize this *metadata in a comprehensive view* per data asset in step six, is discussed in the following Sections 6.4.2 and 6.4.3, respectively.

6.4.2 An Approach for Supporting Diverse Metadata per Data Asset

Metadata is the basis for users to find and understand data in the marketplace. As explained in Section 6.1.1, a diverse suite of metadata can be collected per data asset. This entails different metadata depending on the data type, e.g., image or text, or aspects such as its storage, e.g., data lake or data warehouse, different metadata types, i.e., business, technical, and operational, and also the accumulation of different metadata among similar data assets. In order to reflect this variety of metadata in the data marketplace, we present metadata templates, based on which diverse metadata can be collected for each data asset. To begin with, we introduce the template principles and then discuss how these can be applied in the data marketplace context.

6.4.2.1 Introducing Metadata Templates

In our context, a metadata template refers to a generic list of metadata attributes that can be applied to a large number of datasets. A template can focus on a specific topic, such as metadata related to the data storage system, or metadata related to the data type, such as video, image, or text. Based on these templates, an individual collection of supported attributes is created for each data asset that is then populated by the marketplace. At first, we introduce the structure of such a template followed by an introduction of template bundles. The benefit of using metadata templates is discussed in the following section.

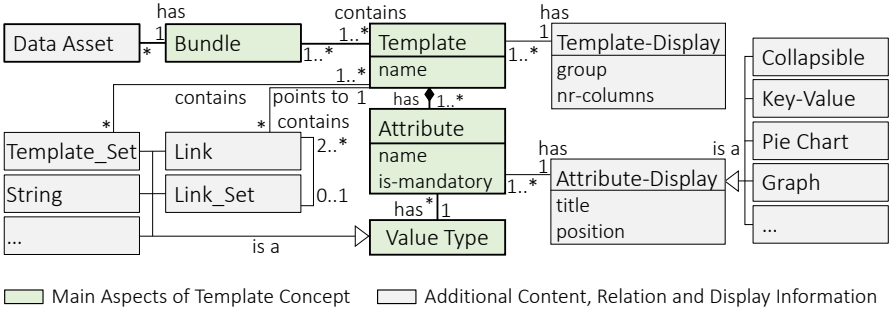


Figure 6.7: Metamodel for Metadata Templates.

Metadata Template Structure: As illustrated in Figure 6.7, a metadata template has mandatory and optional attributes that are associated with one of the value types that are supported by the marketplace, such as string, number, or graph, as described in the previous Section 6.4.1. A template can also contain a set of templates, providing more metadata on the same data asset, as well as a single or set of links to templates of another data asset. By way of example, Figure 6.8 depicts three templates according to the specifications

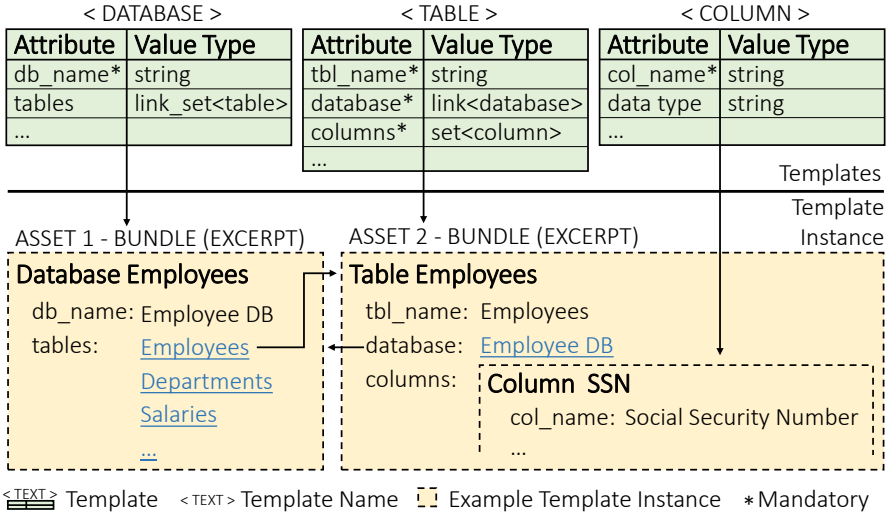


Figure 6.8: Exemplary Metadata Templates with Template Instantiations.

in Figure 6.7 for two types of data assets, a database asset and a table asset. To collect metadata on the database and table asset, there are templates for a *database*, a *table* and *column*. Each template contains a set of attributes, such as *db_name*, together with the value type, e.g., *string*. As a database can contain several tables and a table several columns, the according attributes are grouped in reusable separate templates to avoid redundancy. Whether another template is supposed to be linked or embedded depends on how granularly data assets are defined, e.g., if the whole database represents the asset, or each table is an asset, as in this example. Hence, in Figure 6.8, the table template contains a set of column templates, yet the database links to its according table templates which are separate assets. How granularly data assets are defined and the according template design can be customized to fit a company's needs.

Template Bundles: A data asset is described through a bundle of templates, as defined in the metamodel in Figure 6.7. For example, a data asset stored in a data lake could be associated with a data lake template as well as templates on other aspects like the data quality, its lineage, or the data type, e.g., image. A template can be part of several bundles. Also, as exemplified in Figure 6.8, these can contain links to templates in other bundles, e.g., there is a link from the table asset's bundle to the database asset's bundle. Hence, an individual template bundle can be compiled for different data assets. The marketplace later instantiates and populates the assets' individual template bundles in the form of IDAM entities.

6.4.2.2 Benefits of Metadata Templates

The templates are reusable and can be freely combined with other templates, enabling the creation of custom template bundles for data assets. It is therefore possible to flexibly collect different metadata per data asset. If a new tool is connected to the EDMP which contains metadata that is not reflected by any template, an existing template can be extended or a new template can be created. The flexibility of the template principles also enables the support of various metadata standards, such as Data Catalog Vocabulary (DCAT)[2], or

[2] DCAT: https://w3.org/TR/vocab-dcat-2

metadata models like HANDLE [EGG+21b], as according templates can be created which then enable the collection of the required metadata. In addition, the templates can be implemented in a variety of ways, for example by using semantic technologies like ontologies, or more simply through an assortment of XML or JSON files.

6.4.2.3 Utilizing Metadata Templates in the Enterprise Data Marketplace

Primarily metadata that assists in finding, understanding, selecting, and ordering a data asset is of relevance in the marketplace. Figure 6.9 depicts an exemplary set of templates that may support these steps. The *data-asset-base* template contains a basic set of attributes that describe the data asset and links to a variety of templates that provide further information, such as the *lineage* or *quality* templates. Some attributes reference a distinct set of potential templates, such as the *access-methods* attribute within the *provisioning* template. In this example, we also added templates that provide a set of metadata that is tailored to specific types of data, e.g., images or textual data. Furthermore, templates for the *storage-system* are envisioned. This set of templates lays no claim to completeness. Other metadata on, e.g., the terms of use or reviews may be added to an existent or as part of a separate template.

In the previous Section 6.4.1, it is stipulated that a semantic meaning has to be assigned to the metadata which is extracted from the tools. This refers to a mapping of the tools' metadata to corresponding template attributes. Through

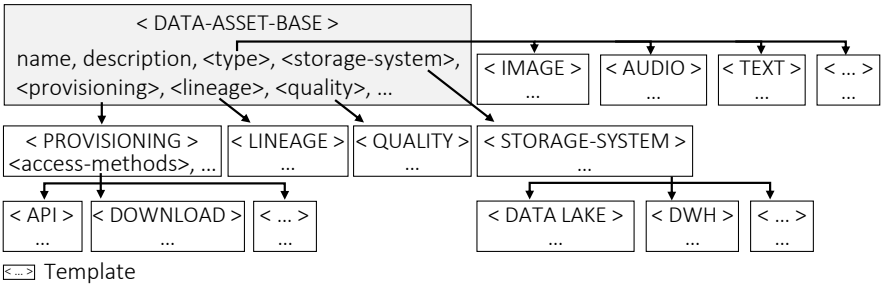


Figure 6.9: Exemplary Set of Metadata Templates for the Enterprise Data Marketplace.

this mapping, the semantics and expected input format are clarified. It can occur that the tool provides the metadata in a different format as defined through the template, e.g., the owner's name as a single value as opposed to two separate values. For this, the data marketplace has a set of predefined population strategies to convert the input into the required format. In this case, a strategy to separate the values is specified with the mapping to the template attribute. Also, if two tools provide input for the same attribute, e.g., if information on the data owner is provided through two catalogs, there are strategies which specify which tool takes prevalence or strategies that record a collection of all supplied values. The topic of metadata integration challenges is discussed in Section 6.5.2 in more detail. Besides the set of supported metadata, the templates also specify how the metadata is visualized, which is discussed in the following Section 6.4.3.

6.4.3 An Approach for Visualizing Diverse Metadata per Data Asset

In this section, we present our approach for visualizing diverse sets of metadata in a uniform manner to support the data consumers through a standardized user interface. The flexible and adaptable visualization of the diverse metadata is achieved by adding visualization specifications to the metadata templates via the template-display and attribute-display options, as defined through Figure 6.7.

Figure 6.10 depicts the general structure how templates are visualized in the marketplace frontend. To achieve recognizability and a concise overview, contextual metadata groups are defined, such as *general information, term of use, data quality and statistics, or lineage*. Like the templates, the groups can be configured individually per data marketplace. As specified through the metamodel in Figure 6.7, templates are allocated to a group. A group can contain several templates. Additionally, the number of columns is specified distinctly for each template. The template attributes are then associated to specific columns. If no content is provided for an attribute, this field is not displayed, enabling a clean design and the use of template excerpts. Exemplified in Figure 6.11 on the top left, a template is allocated to the group *term of use* and one column is specified for this template. The template's attribute *owner* is allocated to this column through the *position* field.

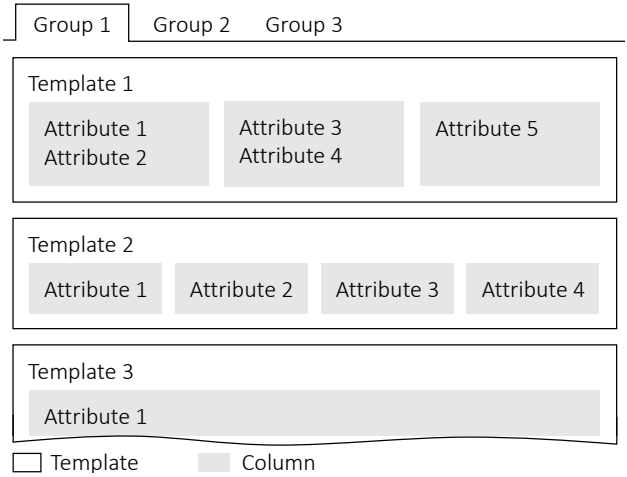


Figure 6.10: Template Visualization Structure.

How each attribute is visualized is also specified in the templates. Amongst others, the visualization options include a simple key-value display, a pie chart, a graph, or a collapsible element, e.g., for embedded templates. For example, a data asset's owner could be configured through an embedded *employee* template that specifies this employee's *name, department* and so on. Figure 6.11 illustrates that the data owner can be visualized as a *collapsible* element, which provides all of the employee data when expanded. An exemplary extract of a populated IDAM entity is shown on the top right and how it could be visualized in the frontend on the bottom. Thus, the marketplace can visualize diverse metadata for each data asset, yet it is presented in a uniform layout as the groupings and visualization types stay the same so the data consumer knows where to look for which information in the user interface.

In short, the three presented concepts enable a) a data marketplace to integrate with a company's tool landscape by connecting with a variety of existent tools which already maintain relevant metadata for the marketplace, b) to support different sets of metadata depending on what is provided through these tools, and c) to visualize an integrated view on this diverse metadata in the marketplace frontend.

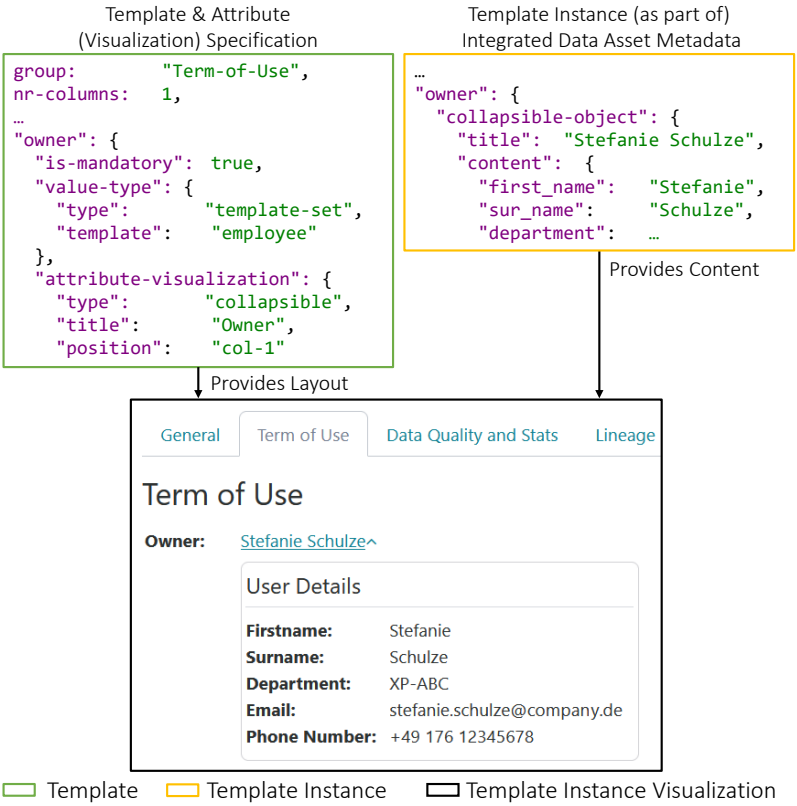


Figure 6.11: This image depicts an exemplary excerpt of a template on the top left, with template and attribute display specifications. An excerpt of an according template instance is shown on the top right. The visualization of the template instance is illustrated on the bottom.

6.5 Assessment of the Metadata Management Concepts

In Section 6.5.1, we assess how the previously presented concepts for leveraging distributed metadata meet the specified requirements, while in Section 6.5.2 we discuss how the concepts deal with the standard (meta)data integration challenges. How the presented concepts can be realized is demonstrated in the next Chapter 7 by means of an EDMP prototype.

6.5.1 Fulfillment of Requirements

Requirement R1 specifies that the EDMP must reuse existent metadata that is maintained within other tools. The concepts discussed in Section 6.4 enable the marketplace to connect to these tools and extract and visualize this metadata. Thereby, the provider does not have to maintain the same metadata in the marketplace and the consumer can use the marketplace as a single point of reference for the distributed metadata when searching for data. Thus, requirement R1 is met. The plugin concept, introduced in Section 6.4.1, enables connecting and disconnecting tools to the marketplace, as well as adding new tools as required, thereby fulfilling requirement R2. By building IDAM entities based on flexibly combinable metadata templates, introduced in Section 6.4.2, each data asset can have its own diverse set of metadata, thus addressing requirement R3. Since it is possible to insert new metadata templates or extend existing ones, new metadata can be added to the marketplace. Hence, R4 is supported. Requirement R5 signifies that the marketplace should provide an integrated view on the metadata, which is met by integrating the metadata in the IDAM entities which are then visualized in the frontend, as described in Section 6.4.3. Similarly, R6 expects the ability to visualize an individual set of metadata per data asset. As the visualization specifications are read and built from the templates and the data assets can have different templates, it is possible to visualize a different set of metadata depending on what metadata is available for each data asset. In short, all six requirements are covered through the three concepts presented in Section 6.4, demonstrating that the concepts enable a marketplace to integrate with a company's existent tool landscape and use the existent metadata therein to the users' advantage.

6.5.2 Addressing Metadata Integration Challenges

The process for providing a consolidated view on the existent metadata entails metadata integration. Therefore, we briefly explain how our process addresses the typical (meta)data integration challenges.

To begin with, we have to consider the heterogeneity of information systems as explained by Leser and Naumann [LN06]. They differentiate technical, syntactic, structural, and semantic heterogeneity, as well as the heterogeneity in data models. The technical heterogeneity refers to the diverse options for data access a system might offer. In the context of the marketplace, this is addressed through the plugins that are written and configured to match the idiosyncrasies of the individual tools and systems. Syntactic heterogeneity, i.e., differences in the representation of the same information, is addressed through the transformation strategies that can map the extracted metadata to a set of predefined formats that are supported by the marketplace. Heterogeneity in data model types signifies different variants for modeling data, such as ER and UML in the context of relational databases or XML in the context of data exchange. The structural heterogeneity refers to different schemata for modeling the same real-world scenario. Both of these, as well as the semantic heterogeneity, are considered by the plugin author. They have to understand the system output and manually map the output to the template attributes which yield the data model as well as content structure and semantics.

Besides these systemic integration challenges, there are further challenges that might occur when integrating data [LN06] or metadata, as in our case. The metadata can be subject to errors and conflicts. This entails issues of varying representation, e.g., owner = “H. Müller” as opposed to owner = 5310. In our case, the required representation for each metadata attribute is specified in the templates and a population strategy that converts the representation based on the given value mappings can be selected. Issues of contradictory values, however, are not addressed through the marketplace, since the marketplace merely provides insights on the collected metadata, but is not the entity responsible for cleansing and improving this metadata. Depending on the chosen population strategy, only the value of one chosen source or all contradicting values can be displayed. Similarly, if systems provide metadata in different units, accuracy, or aggregation levels that do not contradict the template speci-

fications, then this information will simply be displayed in the marketplace for the different data assets as it contributes to a better understanding of the data regardless of its form. In this case, the population strategy for converting the input could adjust the values, e.g., by appending the units like euro and dollar to the values that are provided by the source to avoid misunderstandings. As illustrated by this example, the EDMP only focuses on simple actions in metadata cleansing. In addition to (meta)data errors, duplicates and quality differences also represent integration challenges. In this context, there are a few perspectives to consider: If two tools populate the same attribute, e.g. the owner, there are population strategies that either specify which tool takes prevalence or simply collect all values. Also, if the connection between two tools is not specified correctly, the marketplace might not recognize that metadata belongs to the same asset and consequently creates two entries for this asset and thus a duplicate. While this inconveniences the data consumers, there are no severe consequences.

Having illustrated how the marketplace handles metadata integration issues, it is worth pointing out that the marketplace provides an integrated view on a handful of metadata attributes from a variety of tools, which does not compare to integration scenarios for highly complex schemata of very large data volumes. Not only is the complexity significantly lower, but in most cases the consequences of conflicting, erroneous, or duplicate metadata are significantly less severe than, e.g., in an integration scenario for master data.

6.6 Summary

While Chapter 5 highlighted the architectural view on how an Enterprise Data Marketplace can be embedded in the system landscape of a company, this chapter covered an approach to how this type of marketplace can integrate with different tools and systems to leverage distributed metadata. In this regard, three concepts were presented: a) a plugin-based approach for integrating metadata-based tools with the marketplace, b) metadata templates to enable the support of diverse metadata for different data assets, and c) an approach to visualize this diverse metadata in the marketplace for the users. In conjunction, these three concepts address how the metadata related to data assets is

handled in an EDMP. As metadata management is a fundamental aspect in a data marketplace, the presented concepts lay the foundation for building such a platform in a company-internal context so that it integrates with the existent landscape and builds upon existent functionality and metadata. How the presented metadata management concepts can be implemented, while also considering the topics discussed throughout the previous chapters, is demonstrated in the following Chapter 7.

Chapter 7

Prototypical Implementation and Evaluation

In the scope of this thesis various topics including data consumer and data provider workflows, Enterprise Data Marketplace (EDMP) offerings and functionality, an EDMP platform and enterprise integration architecture, and an approach for leveraging distributed metadata have been discussed throughout the Chapters 3 to 6. In the following Section 7.1, we demonstrate how these topics can be realized in a prototypical EDMP implementation. This prototype is then utilized in an experiment described in Section 7.2, highlighting the impact of employing an EDMP within a company. Having proposed the EDMP as a platform for democratizing company data, we lastly assess to what extent the EDMP establishes data democratization in Section 7.3. Therein, it is examined how the EDMP addresses the data consumer and data provider challenges, and whether the EDMP addresses the five data democratization dimensions. Through this chapter, we thereby demonstrate the feasibility of the concepts presented in this thesis, highlight the necessity for an EDMP, and show that the EDMP is indeed suited for implementing data democratization within companies. This chapter is a revised and composite version of the author publications [EGH+22b; EGHS22; EGH+22a] and [EGH+23].

7.1 EDMP Prototype

The EDMP as presented throughout this thesis is not represented through commercial and open-source tools [JO23]. Therefore, we implemented an EDMP prototype to demonstrate how an EDMP supports key aspects of the data consumer and provider journeys illustrated in Chapter 3, and how the marketplace functionality, as well as platform and integration architecture presented in Chapters 4 and 5 can be realized, together with the approach for leveraging distributed metadata provided in Chapter 6. The prototype implements in particular the aspects of the concepts presented throughout this dissertation that are relevant for the evaluation. Figure 7.1 illustrates which aspects of the functionality framework discussed in Section 4.4.2 are presented through the prototype in the following. To begin with, an overview of the prototype is given in Section 7.1.1, followed by an explanation of how the metadata management concepts presented in Chapter 6 are implemented. This relates to the realization of functionality surrounding the *dataset-specific metadata* as highlighted in Figure 7.1. In terms of role-specific functionality, we present how the prototype offers *service publishing* and *data trading* functionality for the data provider and *discovery* and *data trading* functionality for the data consumer, in Sections 7.1.3 and 7.1.4 respectively.

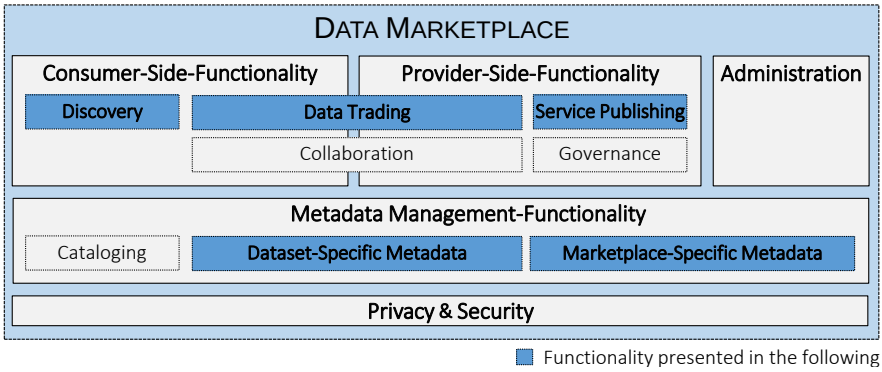


Figure 7.1: EDMP Prototype Functionality Discussed in the Following.

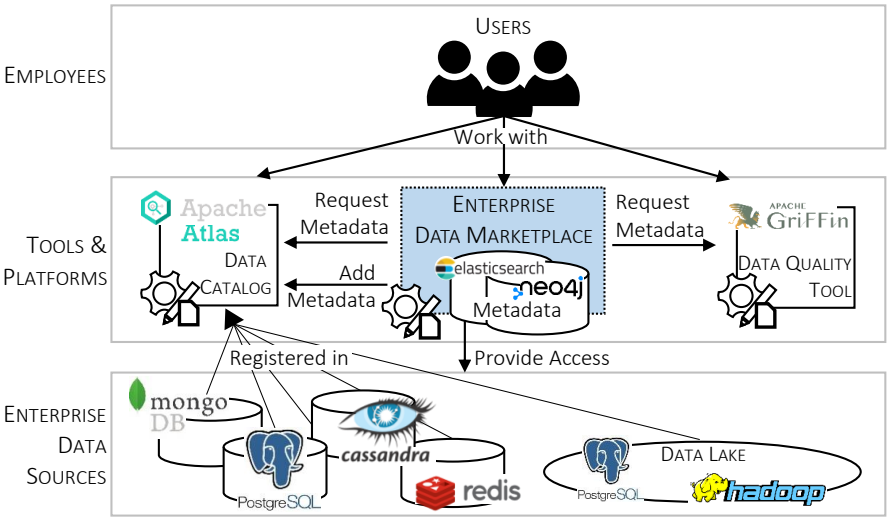


Figure 7.2: Emulated Enterprise System and Tool Landscape. The Blue/Dotted Box Represents the Marketplace. [EGH+23]

7.1.1 Prototype Overview

Before the following sections address the implementation of specific functionality and workflows in the EDMP prototype, we first provide a general overview of the prototype. The choice of tools that were incorporated in the prototype was based on non-commercial and open-source tools because we want to enable free usability and customization.

The EDMP prototype is structured according to the EDMP architecture presented in Section 5.3. It is implemented with the Spring framework[1] based on a microservices architecture including an authentication, discovery, order, security, access, and offerings service. What these services do is described in Section 5.3.2. The services communicate via the message broker RabbitMQ[2] and marketplace-specific metadata is stored in a Neo4J[3] graph database whereas the dataset-specific metadata is stored through Elasticsearch[4].

[1]Spring: https://spring.io
[2]RabbitMQ: https://rabbitmq.com
[3]Neo4J: https://neo4j.com
[4]Elasticsearch: https://elastic.co/de/elasticsearch

In this work, it has been repeatedly emphasized that the marketplace should be embedded in the corporate tool and system landscape, as is detailed in Chapter 5. We, therefore, emulated such a tool and system landscape. As depicted in Figure 7.2 the enterprise data sources are represented through a variety of database types and a data lake. The databases include the document store MongoDB[5], the object-relational database PostgreSQL[6], the columnar database Cassandra[7] and the key-value database Redis[8]. These databases contain a variety of structured, semi-, and unstructured sample datasets. In order to explore how a marketplace 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)[9] and PostgreSQL, and is based on the data lake zone model by Giebler et al. [GGH+20b]. Apache Airflow[10], a workflow management tool, is used to coordinate processes for moving the data into the appropriate zones.

In terms of tools, we have introduced the open source data catalog Apache Atlas[11] and the data quality tool Apache GriFFin[12]. The data sources are registered in Atlas. Amongst others, it provides governance and metadata management functionality for building an inventory 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. GriFFin is a data quality solution that 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. These two tools provide a minimal representation of a diverse tool landscape with distributed metadata which might be of relevance for data consumers when searching for a dataset in the marketplace. How the EDMP prototype deals with this distributed metadata is discussed in the following section.

[5]MongoDB: https://mongodb.com
[6]PostgreSQL: https://postgresql.org
[7]Cassandra: https://cassandra.apache.org
[8]Redis: https://redis.io
[9]HDFS: https://hadoop.apache.org
[10]Apache Airflow: https://airflow.apache.org
[11]Apache Atlas: https://atlas.apache.org
[12]Apache GriFFin: https://GriFFin.apache.org

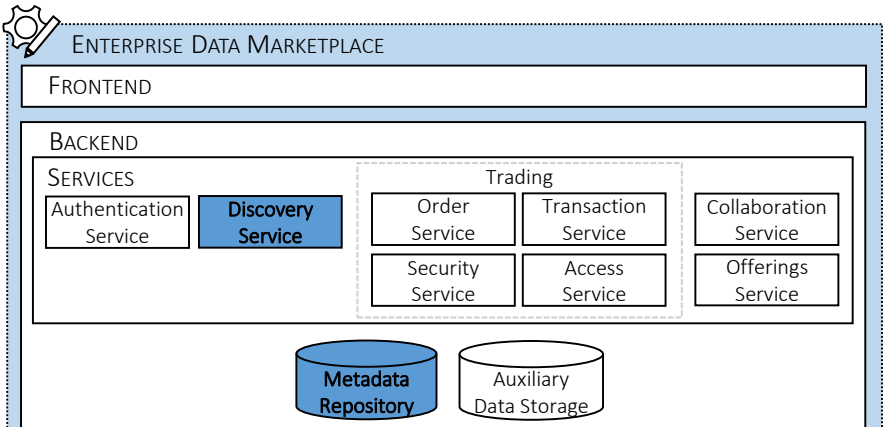


Figure 7.3: Extract of the Proposed EDMP Architecture in which the Relevant Components for Handling Dataset-Specific Metadata are Highlighted.

7.1.2 Metadata Management Functionality

In this section, we describe how our marketplace prototype realizes the concepts presented in Section 6.4 for handling distributed data asset metadata in a company. As illustrated in Figure 7.1 this mainly affects the metadata management functionality for dataset-specific metadata, i.e., the collection and handling of metadata on a specific data asset. Figure 7.3 additionally depicts which components in the proposed EDMP platform architecture (see Section 5.3 for more details) are involved in providing this metadata management functionality. This includes the *metadata repository* and the *discovery service* as it provides search functionality, based on elasticsearch, and an information page, i.e., a detailed view on the provided offerings in the marketplace. This detailed view effectively constitutes an integrated view on the distributed data asset metadata and additionally, the data product metadata if present. In the following, we discuss an implementation approach for realizing the plugin, metadata template, and display concepts presented in Sections 6.4.1 to 6.4.3.

For the sake of simplicity, we implemented our templates and associated files and configurations using json files. As mentioned in Section 6.4, it is

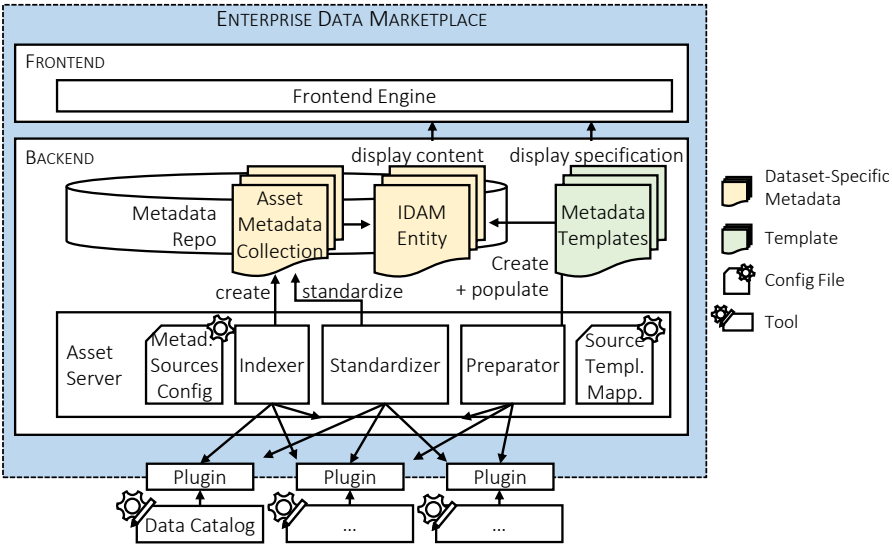


Figure 7.4: Component Overview for Handling Metadata in the Prototype.

also possible to implement the template concept using ontologies, which if a company already works with these may be a convenient approach.

Figure 7.4 illustrates the components within the EDMP prototype that are involved in handling the dataset-specific, i.e., data assets' metadata. There are *frontend* and *backend* components, wherein the prior are dedicated to dealing with the display of metadata and the latter to the processing and storage of the metadata. The tools from which the metadata shall be reused are depicted at the bottom and are connected via a *plugin* as described in Section 6.4.1. A component called *asset server* handles all the processing tasks for the data assets' metadata through its subcomponents *indexer, standardizer* and *preparator*.

To onboard a tool a plugin is created and the tool is registered in the asset server's *metadata-sources-config* file and its relation to other registered tools is specified to enable the allocation of metadata to the same data assets. In addition, how this tool's metadata belongs to which template attributes is specified in the asset server's *source-template-mapping* file with the population strategies for these template attributes, as described in Section 6.4.2.3. When

Extract – Metadata Sources Config.

```
sources:
[...]
- plugin: "griffin"
  mappers:
     - root:   "atlas"
       value: "atlas.attributes.id=griffin.attributes.id“
[...]
```

Extract – Source Template Mapping Config.

```
[...]
"atlas": [
     {
       "template":  "data-asset-base",
       "fields":  {
          {
            "field":     "description",
            "input":     "atlas.attributes.description"
            "strategy": ...
          },
[...]
```

Figure 7.5: Exemplary Extracts of the Asset Server's Configuration Files.

offboarding a tool the entries in the configuration files can either be disabled or removed. To exemplify this, in the prototype, GriFFin's asset entities can be related to the Atlas entities through the entity id, which is specified in the metadata-source-config as shown in Figure 7.5 on the top. The bottom depicts an extract of the source-template-mapping file, in which Atlas' *description* attribute for the data assets is mapped to the *data-asset-base* template's *description* attribute.

The metadata handling process, depicted in Figure 6.5, involves the indexer component extracting the list of tools that should be integrated with the EDMP from the metadata-sources-config file and the indexer component registering these in an indexing pipeline. It then connects to the tools and extracts the metadata specified in the according plugins from these tools. This metadata is associated with data assets, so the indexer creates *asset-metadata-collection* entities for each data asset, if not yet existent, and fills these entities with the according metadata. The metadata is inserted into the asset-metadata-collections sequentially per tool, as provided through the tools, meaning the

standardization and integration of the metadata are not yet considered. The indexer component therefore addresses steps one, two, and three, i.e., tool connection, metadata extraction, and assignment to assets, of the metadata handling process, illustrated in Figure 6.5. Next, the metadata is standardized. For this, the standardizer component uses the transformation strategies, provided through the plugins, to process the extracted metadata within the asset-metadata-collections. The metadata is transformed so it conforms to the supported value types in the marketplace, e.g., string or graph. This relates to step four, standardize metadata format, in Figure 6.5. The marketplace now has the extracted metadata grouped by data asset in a format it can process.

The next step in Figure 6.5's metadata handling process involves clarifying the semantics. The preparator component assembles the template bundles based on the source-template-mapping configuration which specifies for each tool which metadata belongs to what template attribute and therefore which templates are required for which data assets. The preparator creates an *Integrated Data Asset Metadata (IDAM) entity* based on the template bundle and populates it from the corresponding asset-metadata-collection by executing the population strategies. For instance, the asset-metadata-collection may contain the data owner's first and surname as one value, whereas the template may require these as separate values. The preparator finds the according fields in the asset-metadata-collection, reads and executes the population strategy to split these values, and adds them to the IDAM entity. At this point, the marketplace contains a set of semantically clarified IDAM entities for each data asset. An example extract of such an IDAM entity can be viewed in Figure 6.11 on the top right.

In order to display the collected metadata, the frontend engine reads the display specifications of all templates in a data asset's template bundle. It generates and populates the layout with the metadata from the IDAM entity and visualizes the metadata as specified, e.g., as text or a pie chart.

Figure 7.6 depicts a screenshot of the information page, i.e., a detailed view on a data asset, here "Welder - Error Report 09.01.2023", which is available through the marketplace. It shows part of the IDAM entity in the middle. The other parts of the IDAM entity are spread across the other groups, i.e., tabs. In the currently active tab, *Data Quality and Stats* metadata from both

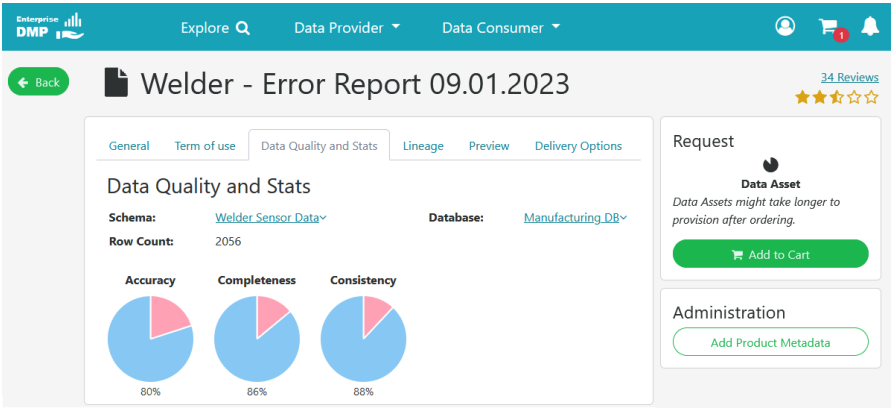


Figure 7.6: An Integrated View on a Data Asset's Metadata in the Enterprise Data Marketplace [EGH+23].

Apache Atlas and Apache GriFFin is displayed, wherein the key-value pairs like the schema information originate from the Atlas catalog and the data quality metrics from GriFFin. On the right and above the displayed metadata in the detailed view, other marketplace functionality is visible like the shopping cart for ordering data, which is introduced and discussed in more detail in the following with the role-specific functionality.

7.1.3 Data Provider Functionality

In the scope of this work, we have deduced that it is challenging and cumbersome for data providers to publish and provide their data within an enterprise. In Section 7.1.3.1 we, therefore, introduce the workflows and functionality for registering data in the marketplace prototype, which is part of the required *service publishing* functionality, highlighted in Figure 7.1 and described in Section 4.4.2. Furthermore, we provide brief insights into the implemented *data trading* functionality on the provider side in Section 7.1.3.2, as it involves fundamental features for sharing data within the enterprise.

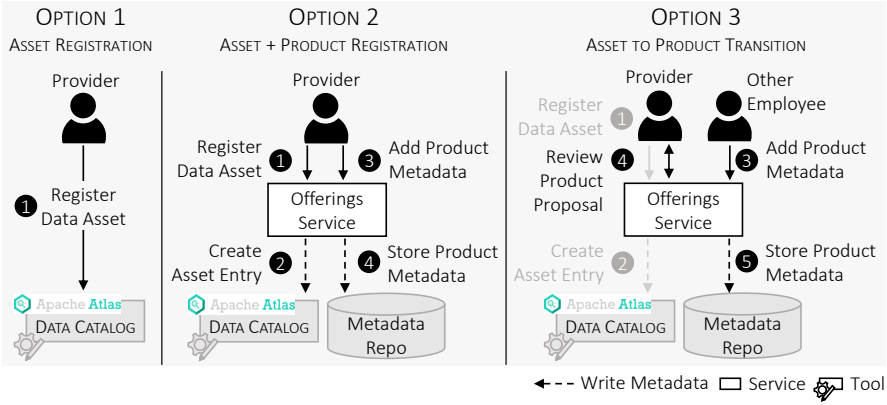


Figure 7.7: Data Registration Process Variants and Possible Implementation Variants with and without a Data Catalog (Based on [EGH+23].)

7.1.3.1 Service Publishing Functionality

The main offering of the EDMP prototype is data-as-a-service. Hence the functionality for service publishing in the following refers to the registration of data assets and data products in the marketplace. The registration process differs depending on whether the marketplace integrates with an existent data catalog or not. Without a data catalog, both the registration of data assets and products is done through the marketplace, and all the according metadata lives in the marketplace metadata repository. With a data catalog there are two options for registering data assets, through the catalog, as well as the marketplace, and the metadata can live in different places. As we have integrated our EDMP prototype with the data catalog Apache Atlas we will focus on the variant with the data catalog in the following. Our prototype supports three options for data providers to register data assets and products which are illustrated in Figure 7.7.

Option 1 - Asset Registration in the Catalog: The provider can register a data asset in the *data catalog Apache Atlas*. This is illustrated as provider option 1 in Figure 7.7, on the left. For this, the provider can dial into the Atlas GUI, as illustrated in Figure 7.8, and fills in the corresponding form fields for registering data assets. As the marketplace is integrated with the catalog this

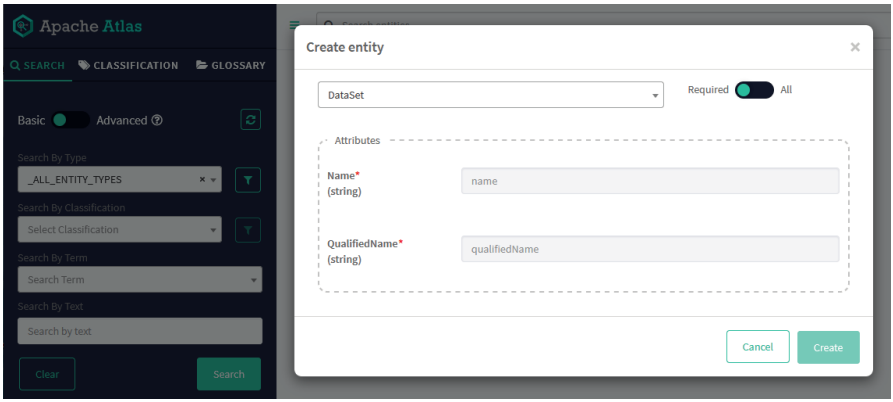


Figure 7.8: Data Asset Registration in Apache Atlas [EGH+22b].

entry can be found in the marketplace and is listed as a data asset. If this asset is then requested in the marketplace by a data consumer, the provider receives an access request and is prompted to add the product metadata, jumping into the workflow of option 2 at step 3.

Option 2 - Asset and Product Registration in the Marketplace: In step 1 the provider navigates to a data registration wizard in the marketplace to enter the asset metadata which is required in the corresponding catalog, i.e. Atlas. The *offerings service* then creates an according entry for the data asset in the catalog, in step 2. At this point the same state is achieved as in option 1 by directly registering the asset in the catalog, meaning the provider can stop here, as this dataset can now be found in the marketplace by potential consumers. The provider can decide to turn the data asset into a data product at a later point in time after, e.g., it has been requested by a data consumer, or alternatively, the provider can register the data product directly. The data registration wizard in the prototype guides the provider through three forms.

The first form prompts the provider to specify whether the data is already registered as an asset and if so, to enter the asset-id. This is relevant for the provider to be able to re-enter this process at a later point. Having specified the asset-id, the provider is led to the second form for registering or editing existing data assets, which is pre-populated with the metadata from the data

catalog if the asset already exists. According to step 3, the third form, which is illustrated in Figure 7.9, enables adding product metadata. In the prototype this constitutes the *terms of use* in which the provider indicates if it is personal data, the *permitted usage*, *conditions of use*, a *license*, and *data delivery options*. Laws like the General Data Protection Regulation (GDPR) allow people to influence how their personal data is processed [Eur16]. Therefore, if the data is personal and the processing has been restricted to a specific usage this must be specified in the field permitted usage. Other conditions such as restriction of usage by a specific team can be specified in the field conditions of use. If none of the licenses fit the requirements, the provider can create a customized license in the prototype. Additionally, the *data delivery*

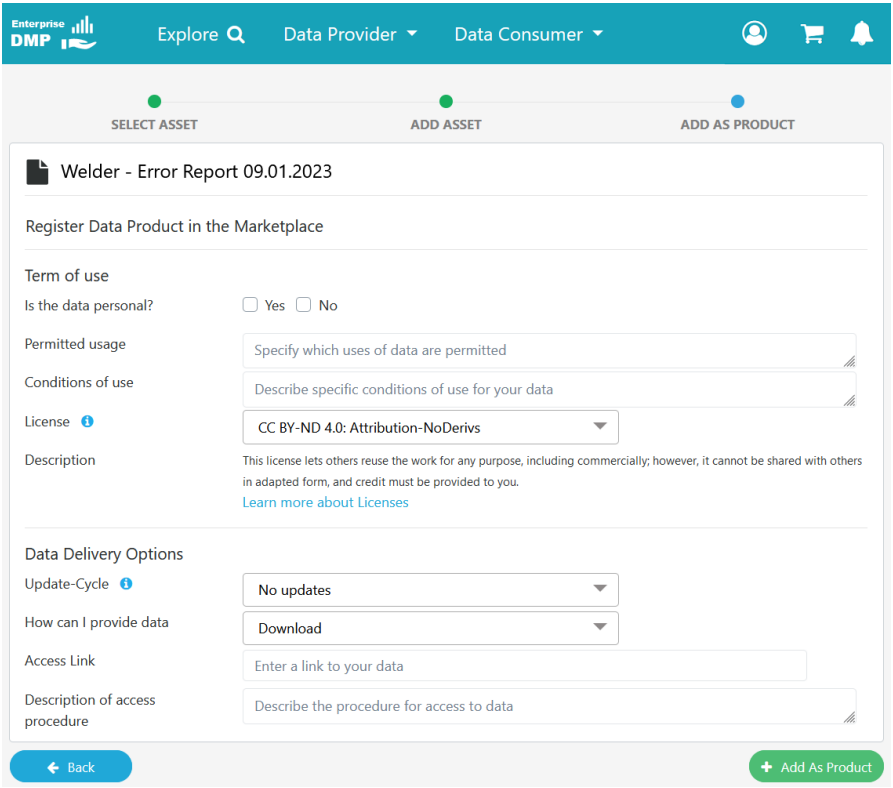


Figure 7.9: Data Product Registration Wizard [EGH+23].

options are specified. This includes information on the data's *update cycle*, the *provisioning options*, and *description of the access procedure*. Having specified this information, the provider clicks on the button "Add as Product" after which the offerings service stores the data product metadata in the Neo4J metadata repository in step 4. The data can now be ordered through the data marketplace as a data product.

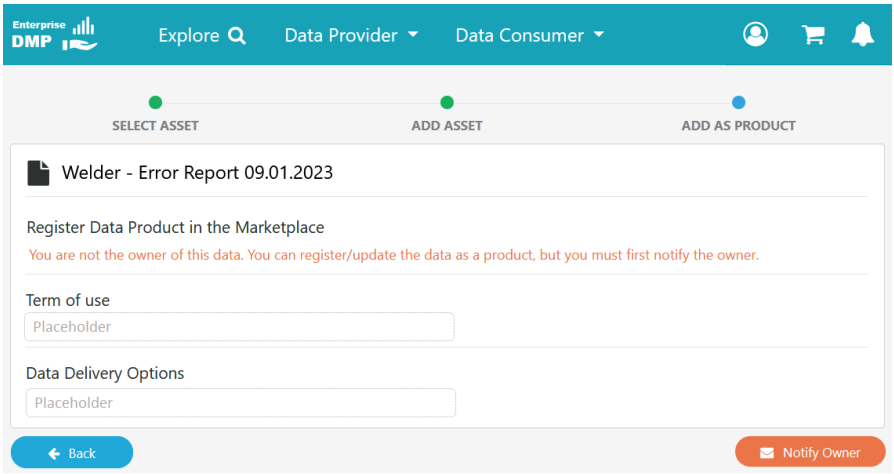


Figure 7.10: Product Registration Wizard viewed as not the Data Owner.

Option 3 - Asset to Product Transition Initiated Through Other Employees: The third option entails that another employee can fill in the required product metadata for the provider, an alternative step 3 in Figure 7.7. If a user selects a search result they are taken to a detailed view with metadata providing details on the dataset as depicted in Figure 7.6. This detailed view also specifies whether the result is an asset or a product. In the case that it is an asset, a button is displayed "add product metadata". Clicking on this button will take the user to the registration wizard where the user can navigate to the "add product" form, illustrated in Figure 7.9. If the user is not the data owner, this is displayed with a message as shown in Figure 7.10. The form is submitted using the "Notify Owner" button. In step 4 the owner then receives the request for asset-to-product transition, as depicted in Figure 7.11. They can edit the

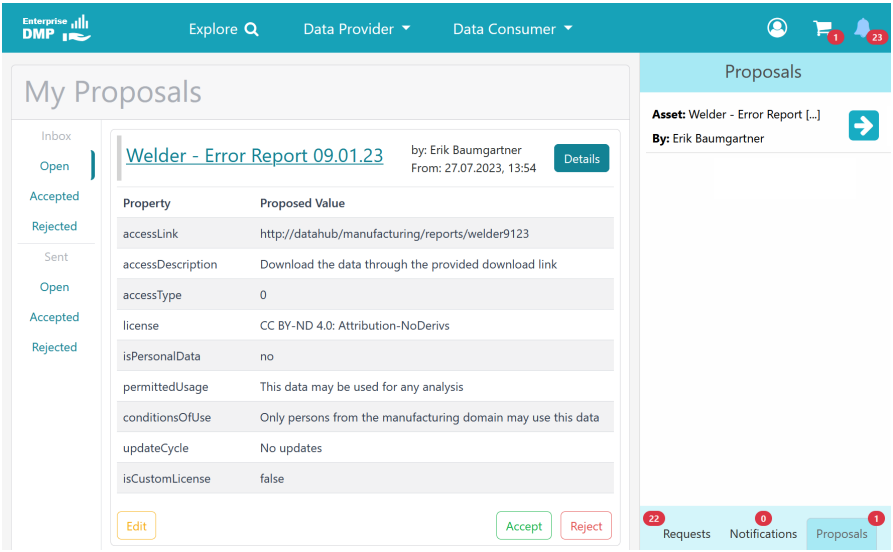


Figure 7.11: Transition Request as Viewed by the Data Owner.

metadata, and accept, or reject the request. In our prototype, this request is sent to the owner, but ultimately it should be possible that this request is sent to both the owner and original provider who registered the data asset so that both can review the entered product information. If the request is accepted, the offerings service creates the data product by adding the information to the data product index in the metadata repository.

Having registered a data asset or product a provider and owner might wish to have an overview of all registered items in the marketplace. This functionality is available in our prototype through the drop-down menu in the navigation bar on the top under *data provider* which can be seen in the various screenshots.

7.1.3.2 Provider Side Data Trading Functionality

Besides the ability to register data assets and products in the marketplace, we also implemented functionality that enables the providers to trade data assets and products. This includes functionality for access management of offered

assets and products, as well as subscription and order management for those which have already been shared.

Figure 7.12 illustrates how a data owner is notified of open access requests on both their data assets and data products. These are listed within the *bell* on the top right through which the providers can navigate the different requests, or they can open all requests through the *data provider* drop-down in the navigation bar on the top. They see who is requesting access to what asset or products and what they intend to do with the data. If the owner accepts the request and it is an asset, this request will move from the *open* requests selection to the *accepted but require action* selection of requests and they will be prompted to fill in the data product metadata. As long as the asset is not transitioned into a product the request will be pending for the data consumer.

The data provider can also view all ongoing as well as closed subscriptions on their data products, meaning who still has access to the data and who used to have access to it. This for instance enables the providers to contact all consumers if changes are made in the data, or also enables the provider to close a running subscription if necessary.

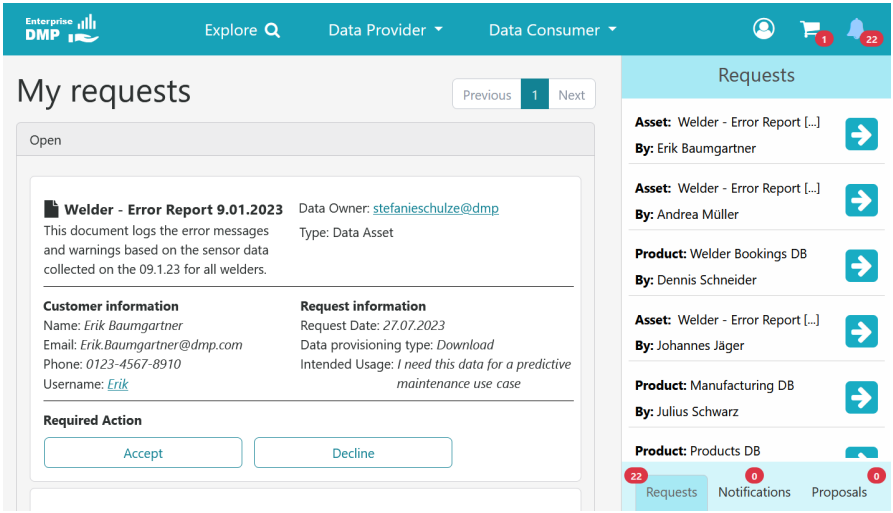


Figure 7.12: Access Request as Viewed by the Data Owner.

7.1.4 Data Consumer Functionality

The main steps of the consumer journey presented in Section 3.3 involve finding and requesting access to data. Therefore we briefly show how the EDMP prototype enables finding data through its discovery functionality in Section 7.1.4.1 and how a consumer can order, i.e., request access to this data as part of the marketplace's data trading functionality in Section 7.1.4.2.

7.1.4.1 Discovery Functionality

Figure 7.13 illustrated the process for finding data. The consumer enters a request into the frontend search bar in step 1. Based on the search string the *discovery service*, which is based on elastic search, collects entries from the *data catalog Atlas*, in step 2. Then it collects additional metadata such as product metadata from the *metadata repository* and according metadata from other tools such as quality metadata from GriFFin in step 3. How this metadata is extracted and integrated is explained in Section 7.1.2. A list of search results as shown in Figure 7.14 is returned to the consumer in step 4. Each entry is flagged as either a data asset or data product as can be seen in Figure 7.14 on the right-hand side. The single results can be selected to provide a detailed view, i.e., information page on all the collected metadata as previously shown in Figure 7.6. This detailed view is one of the features that sets the EDMP

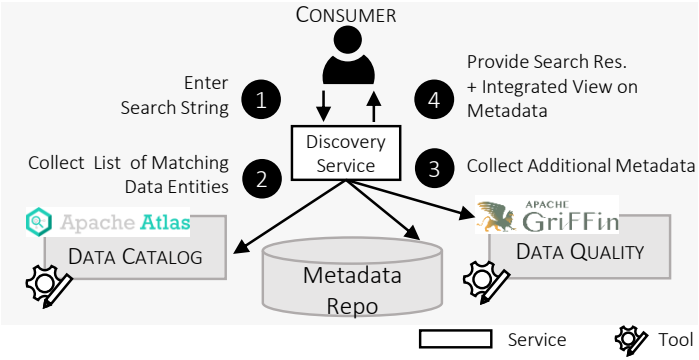


Figure 7.13: Search Process for Data with Involved Tools and Components [EGH+23].

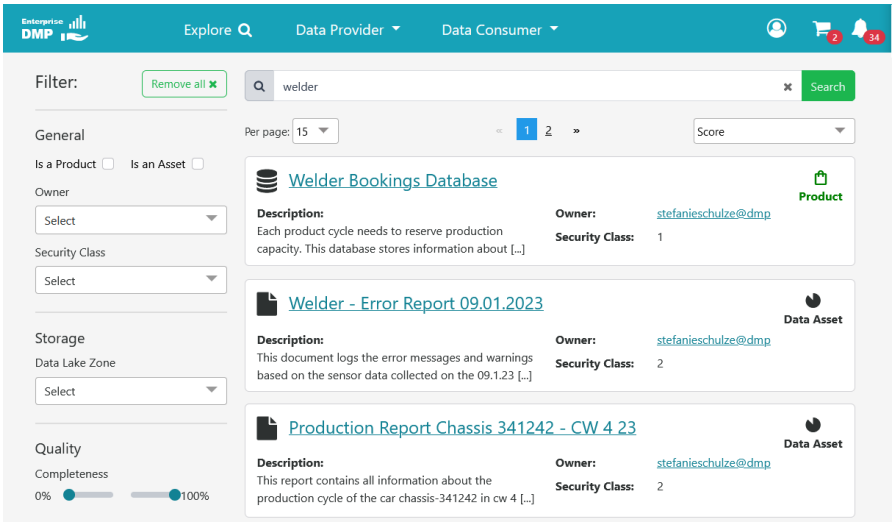


Figure 7.14: Prototype - Search Results View.

apart from external data marketplaces, as it demonstrates how the EDMP can tightly integrate with the existent tools and provide a comprehensive view on data assets and products by leveraging and integrating the existent metadata.

7.1.4.2 Consumer Side Data Trading Functionality

For the data consumers functionality for trading data also involves the ability to request access to data, transaction management, and subscription/order management as described in Section 4.4.2. The prototype supports an access request process for requesting access to assets and products as well as aspects from the subscription/order management which are outlined in the following.

In the marketplace's detailed view, as shown in Figure 7.6, the consumer can add data assets or products to a shopping cart. The consumer is guided to a form in which they must specify the intended usage and which provisioning option they would prefer, e.g., download or direct access to the source. Having specified this information, they can issue the order which is illustrated as step 1 in Figure 7.15. 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

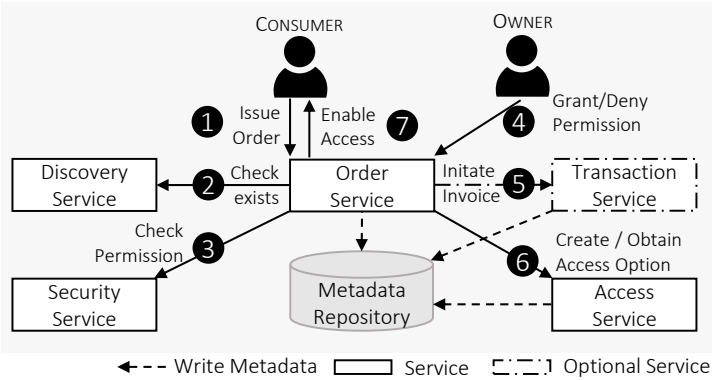


Figure 7.15: Access Request Process with Involved Components and Their Interaction Patterns [EGH+23].

been verified the order service transfers the request to the *security service* in step 3 through 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 service 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 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.

In terms of order management, the data consumer can track the progress of the request as depicted in Figure 7.16. They can also view their active, expired, or rejected subscriptions and terminate active subscriptions if required.

Based on this prototype, we have demonstrated how the marketplace supports central aspects of the workflows for data consumers and data providers in their journeys presented in Chapter 3, how some of the required marketplace functionality discussed in Chapter 4 can be realized, how the platform

and integration architecture introduced in Chapter 5 can be implemented, and lastly, how distributed metadata can be leveraged in the marketplace as discussed in Chapter 6.

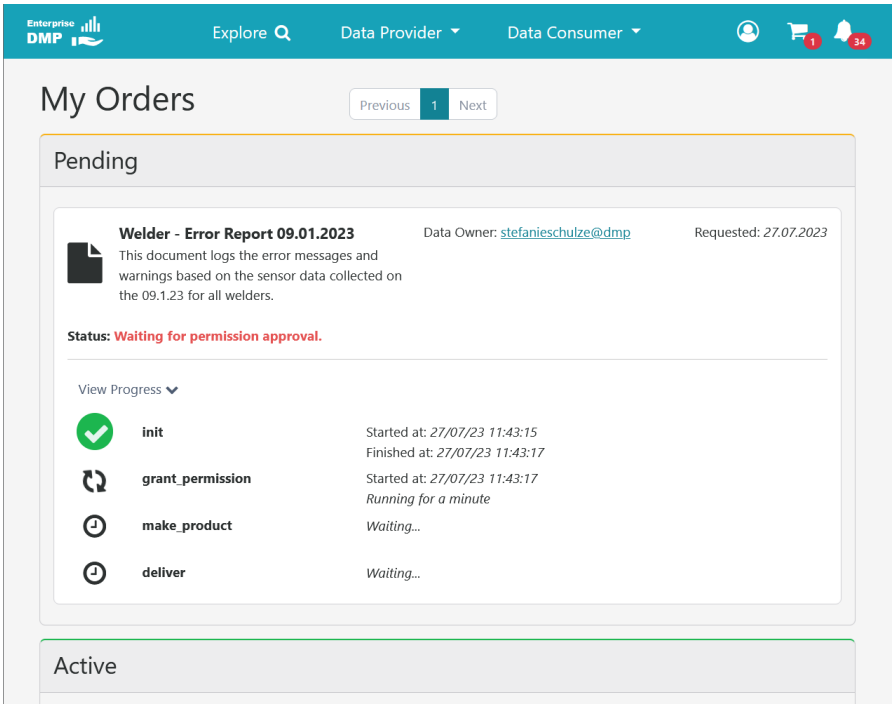


Figure 7.16: Tracking the Progress of Open Order Requests.

7.2 Experiment: Evaluating the Impact of an EDMP

Throughout this thesis, we stipulate that the EDMP improves the data consumer and data provider journeys, discussed in Chapter 3. We claim that by using an EDMP and integrating it with the company tool and system landscape several of the consumer and provider challenges are addressed. In order to verify these assumptions, we leveraged the EDMP prototype described in the previous section to conduct an experiment 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.

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 Sections 7.2.1 to 7.2.3 we outline the experiment design, the results of the experiment, and a discussion and conclusion of the results.

7.2.1 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 Section 7.1.1 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. The participants were tasked to find equivalent datasets within this lineage graph within the two scenarios. 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: The participants were tasked to act as data consumers in both scenarios and, in accordance with the first two segments of the data consumer journey presented in Figure 3.3, 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" [CGK12], 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, the author 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. 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.

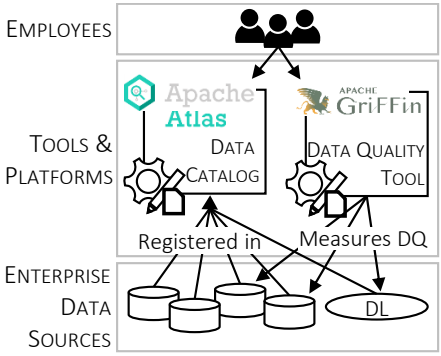


Figure 7.17: System and Tool Landscape in Scenario S1 without an EDMP.

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.

We based the experiment on our prototypical enterprise tool and system landscape as presented in Section 7.1.1 and depicted in Figure 7.2. Within this scenario the participants do not get access to the data marketplace but to the data catalog Apache Atlas and the data quality tool Apache GriFFin. This results in the tool and system landscape as depicted in Figure 7.17. As a starting point, the participants were given a link to the data catalog tool together with user account details.

Figure 7.18 depicts the workflow they had to figure out to find and request the according data asset. The consumer journey steps are illustrated on the top and the according steps which had to be conducted in the scenario and the tools or methods for these steps, on the bottom.

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 data quality (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

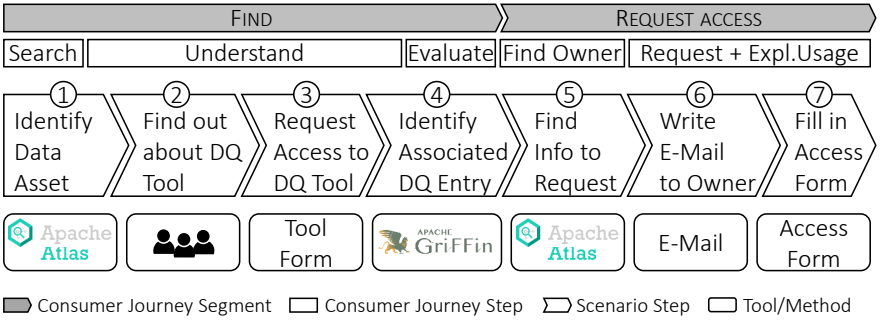


Figure 7.18: The Workflow and Tools Without the Use of an EDMP (Based on [EGH+23].)

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 Figure 7.18 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 second scenario the participants worked with the EDMP prototype as described in Section 7.1. 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. Their current objective was the creation of a dashboard to enable predictive maintenance throughout the production line for a car chassis, for the production steps “welding” and “painting”. The

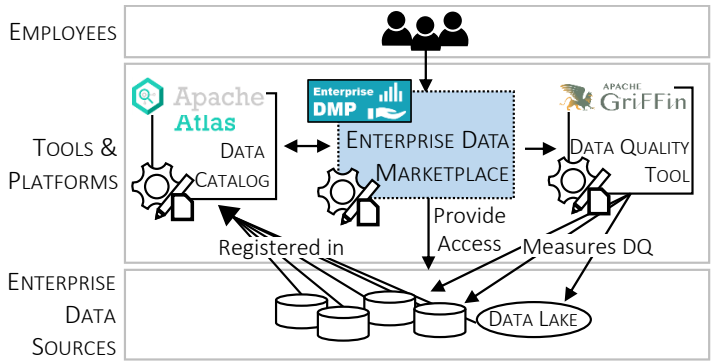


Figure 7.19: System and Tool Landscape in in Scenario S2 with an EDMP.

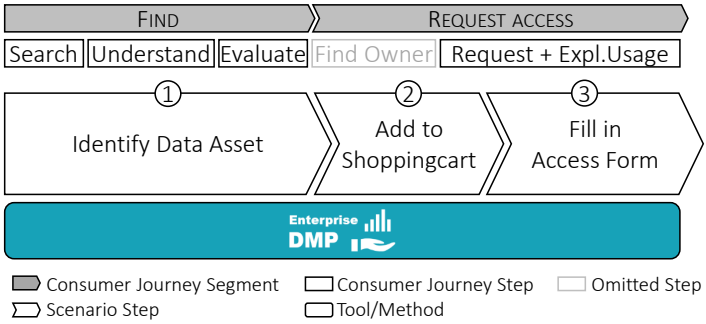


Figure 7.20: The Workflow and Tools with the Use of an EDMP (Based on [EGH+23].)

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. The tool and system landscape used within this scenario is depicted in Figure 7.19.

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 Figure 7.20, 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. As they do not need to manually find the data owner, the according data consumer journey step "Find Owner" is faded out in Figure 7.20. The last step involves filling out and submitting the form to request access in the shopping cart.

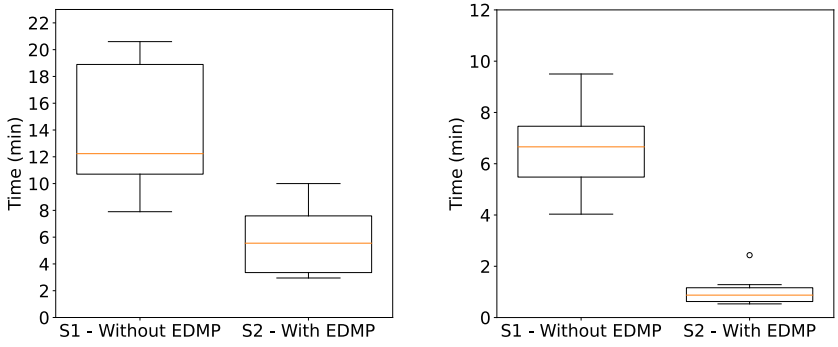
Measures: To determine whether the hypothesis holds true, the three metrics, efficiency, effectiveness, and complexity had to be measured. The efficiency relates to the time required to perform the assigned task. We, there-

fore, 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. The three questionnaires with the given answers are provided in the appendix of this thesis. 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 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. The questionnaires also inquired about other aspects, yet as these did not yield any further findings in terms of efficiency, effectiveness, and complexity, we focus on the relevant questions and results for examining the hypothesis.

7.2.2 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 Section 7.2.3.

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 Figure 7.21. As can be seen on the left in Figure 7.21a, 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 is shown in Figure 7.21b. 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



(a) Find Process

(b) Access Request Process

Figure 7.21: Time Required for the Find and Access Request Process in both Scenarios [EGH+23].

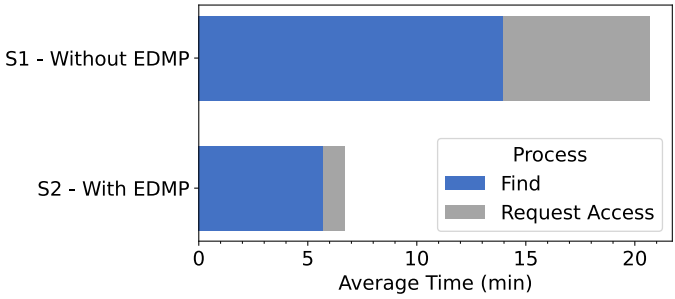


Figure 7.22: Scenario Runtime Comparison [EGH+23].

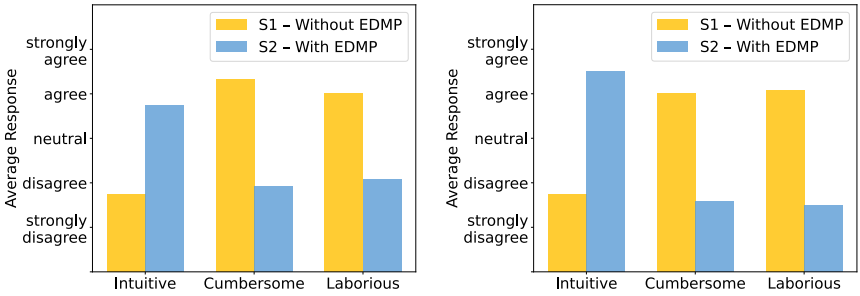
the processes in scenario S2, with an EDMP. This is especially pronounced in the process of requesting access.

Figure 7.22 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:

Figure 7.23 depicts the results of the two questionnaires the participants filled out after completing each scenario. The bars reflect the average answer given 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 under-*



(a) Find Process

(b) Access Request Process

Figure 7.23: A Comparison of the Participants Perception, if the Consumer Process in the Scenarios was Intuitive, Cumbersome, or Laborious [EGH+23].

standing 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 Figure 7.23b, 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 7.24 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.

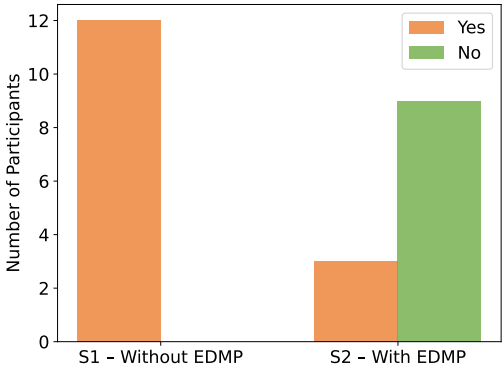


Figure 7.24: Required Guidance [EGH+23].

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 7.1 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.

7.2.3 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,

Question	Result (Mean)
Simplified finding & understanding	agree
simplified requesting data	strongly agree

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

Table 7.1: Process Simplification through the EDMP.

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 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. We attribute the prolonged times to find data in scenario S1 without the EDMP to the fact that the participants first had to figure out where the required information is available, then had to request access to the according tool and figure out how to operate this tool. Additionally, the metadata in the various tools in a company are not necessarily aligned, meaning the participants had to figure out which entries in the data catalog and quality tool belong together. By integrating this available metadata, the EDMP addresses the consumer challenge of distributed metadata (Cons-C2) which has both the effect that the required information is available in one place and that employees do not need to find and operate a variety of tools. As shown through this experiment, this reduces the time required for finding data by more than half.

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. 7.21b, 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. Also, aspects like contacting the data owner were automated in scenario S2 in the EDMP, addressing the consumer challenge (Cons-C1) of having to identify and contact various parties. 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, addressing part of the consumer challenge (Cons-C3), concerning the challenge that tools are not integrated across the consumer process.

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, 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 the participants on average agree that the process to find data is intuitive with an EDMP and strongly agree that the access request process is intuitive therein, whereas they disagree on both accounts without the EDMP, we deduce that the EDMP increases the intuitiveness of the consumer process. We believe this is due to the fact that it integrates with the available tools and thereby offers an integrated view on metadata, as well as that it offers one specific workflow for the data consumer process. In this workflow, the users are guided through a set of steps so they do not have 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. In scenario S1 without the EDMP the participants disagree that it was clear which steps had to be followed in the find and access request process whereas in scenario S2 they agree for the finding process and strongly agree for the access request process. 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. 7.21, 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 7.2. 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 company's goal to democratize their data, the consumer process relates to the data democratization dimension one, which involves the

Efficiency	Effectiveness	Complexity
✔	-	✔

✔ improved, - unchanged

Table 7.2: Hypothesis Evaluation: The EDMP Improves the Consumer Process in Terms of Efficiency, Effectiveness, and Complexity.

accessibility of data [LLF21]. 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 impact 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 had 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.

7.3 Assessing the EDMP for Establishing Data Democratization

In the scope of this thesis, the EDMP is proposed as a tool for democratizing company data. Therefore, this section evaluates to what extent the EDMP as presented within this thesis is suited to further the democratization initiative. To begin with, we explore in Sections 7.3.1 and 7.3.2 to what extent the EDMP addresses the data consumer and data provider challenges presented in Chapter 3. This is relevant as the goal of data democratization involves the

enablement of employees to find, understand, access, use, and share data. If the EDMP addresses the challenges of the data consumers and data providers, it directly contributes to the democratization of data. Thereafter, we discuss in Section 7.3.3 to what extent the EDMP contributes to the five data democratization dimensions, how the challenges which accompany an EDMP impact the data democratization initiative, and to what extent the concepts presented throughout this work already address these challenges. In conjunction, we thereby gain perspective on the impact the EDMP has on a data democratization initiative.

7.3.1 Addressing the Data Consumer Challenges

The data consumer journey presented in Section 3.3 and the challenges therein correlate to the data democratization dimension one, which aims for broader accessibility of data. The data consumers are the party that wants to acquire access to data and thus, the challenges they face throughout this process can be translated into data democratization challenges within dimension one. Therefore we examine how and to what extent the EDMP addresses these challenges. Table 7.3 lists an overview of the challenges and which are addressed.

The consumer challenge one (Cons-C1) refers to the number of *people that need to be found in the consumer process, contacted* individually, and that there will be delays until each processes the tasks. Companies are currently building data catalogs part of which involves assigning a data owner to data assets. As our EDMP prototype integrates with existing metadata management tools, in this case, the catalog, and extracts metadata such as the data owner or other related employees, the consumer does not have to find the relevant

Challenge ID	Description	Addressed
Cons-C1	process involves several parties	✔
Cons-C2	distributed metadata	✔
Cons-C3	tool integration access process	✔

✔ fully addressed

Table 7.3: Addressing of Consumer Challenges.

people manually. The prerequisite for this, however, is that this metadata must be maintained within at least one of the existent metadata management tools. Additionally, some communication steps can be automated, for instance, the access request is sent to the respective people automatically. Also, if the according people from management can be extracted from any source, the consumer journeys step, request for resources, to move the data into another system, or create provisioning options can be realized in the consumer workflow in the marketplace and sent to the according people automatically. Communication between the parties is handled through the marketplace interface which provides consumers and producers with a workflow and overview of new, active, and expired requests and subscriptions. This simplifies the finding of and exchange of data with colleagues, and as shown through the experiment described in Section 7.2, this also accelerates the consumer process.

The second challenge (Cons-C2) concerns the *distribution of metadata in different tools*. The EDMP addresses Cons-C2 by requesting and displaying the metadata spread across the tools and integrating it according to the concepts presented in Chapter 6. This also significantly reduces the time required to discover data assets and products as demonstrated through the experiment described in Section 7.2.

Lacking *integration in the toolchain across the access process* is consumer challenge three (Cons-C3). As shown in Section 7.2, an EDMP provides support through most of the consumer journey's workflow. Finding data and requesting access are supported in the EDMP and data that can be found in the marketplace can also be requested as the marketplace is integrated with the metadata management tools including the data catalog. Enabling access to data by moving it into another system is, however, exempt from the EDMP's capabilities and currently not supported through an integrated toolchain.

Summing up, it has been shown that an EDMP addresses the three data consumer challenges and thus advances the data democratization initiative on the data consumer side.

7.3.2 Addressing the Data Provider Challenges

Like the consumer journey, the provider journey presented in Section 3.2 is related to data democratization dimension one. The data providers are the

party that has to make their data accessible so that it becomes available to data consumers. Therefore, also the data provider challenges constitute data democratization challenges in the scope of dimension one. In this section, we, therefore, discuss to which extent the EDMP addresses the data provider challenges, which is also consolidated in Table 7.4.

The first challenge (Prov-C1) signifies the *assembly of metadata*, i.e., meaning the collection of metadata currently not maintained throughout the tool landscape. In effect, this task is supported to a certain extent through tools that can automatically capture metadata. For instance, the data catalog Alation uses AI to suggest business glossary terms and suggests links to relevant data [Sha]. This concerns the second step of the provider journey. While the marketplace can extract existent metadata from the tools, it does not support the maintenance of this metadata within these tools and hence does not fully address this challenge.

Challenge two (Prov-C2) refers to the *effort of supplying provisioning options*, even if these may not be required. This issue is addressed by the EDMP through the differentiation of data assets and data products. It is dealt with by allowing the provider to supply product metadata and thereby make provisioning options available only when a request is made for a data asset. Therefore, the effort relating to provisioning options is only undertaken if this data is actually relevant for other employees.

Challenge three (Prov-C3) deals with the necessity of *registering data in several publishing tools* like data catalogs and a data marketplace. Whether this challenge is addressed by the EDMP depends on the implementation

Challenge ID	Description	Addressed
Prov-C1	assembly of metadata	~
Prov-C2	supplying provisioning options	✔
Prov-C3	registering data in publishing tools	✔
Prov-C4	process involves several parties	✔

✔ fully addressed ~ partly addressed

Table 7.4: Addressing of Provider Challenges.

approach that is chosen. The EDMP can be built as a standalone platform with its own inventory. In this case, challenge three is not addressed, data must be registered in all tools, i.e. catalog and marketplace, and some metadata must be maintained twice. As explained in Chapter 5, the implementation alternative involves integrating the marketplace with a company's existing data catalog. If data catalogs are used as an inventory for the EDMP, so it can find the data assets that are registered in them, the provider only has to register data in the data catalog or data marketplace. This avoids the need to register data in more than one tool. In addition, as the marketplace reads metadata from the data catalog, the duplication of the same metadata and the duplicate maintenance of it is avoided. Hence, this implementation option addresses the challenge of redundant data registration and metadata maintenance. It can also be added that users who are not data owners or dedicated data providers can update data assets to products in the data marketplace, which eliminates the need for the data provider to do this. In this case, the data provider will merely receive an asset-to-product transition request, reducing the provider's efforts even further.

That the data provider's journey *involves several parties* which have to be found, contacted, and coordinated constitutes challenge four (Prov-C4). There are two steps in the journey, which involve a request to third parties that can be partially automated through the marketplace. This includes the *request to publish data*. For this, however, the owner and the legal experts must be known and specified. If this is the case, the marketplace represents a platform via which a workflow for the request and approval of such processes can be implemented. The same is true for the *request for resources*. If the according people from management are known and can be identified by the EDMP, then it can also ensure a regulated workflow for the resolution of this subject matter.

Consequently, all of the challenges are addressed through the EDMP, yet it specifically resolves the challenges two through four. The EDMP therefore also furthers the data democratization initiative on the data provider side.

7.3.3 Addressing Data Democratization Dimensions

As described in Section 2.2, data democratization entails the five dimensions, broader access to data for users with varying skill-sets (Dim 1), broader access

to self-serve analytics tools (Dim 2), the development of data and analytics skills (Dim 3), collaborative knowledge-sharing (Dim 4), as well as the promotion of data value (Dim 5) [LLF21]. Hereonforth, we examine the extent to which an EDMP supports these five dimensions, how the EDMP challenges introduced in Section 4.6 relate to these dimensions, and how the concepts provided in this thesis already address these. Lastly, we examine how the EDMP functionality contributes to data democratization.

Several of the democratization dimensions are addressed through the EDMP offerings and functionality as described in Section 4.4. Broader access to data is supported through the marketplace offering Data-as-a-Service, as it makes all kinds of data available both as assets and products. In addition, both data consumers and data providers are supported in their processes surrounding data access and provisioning. Since the EDMP guides the users through the workflows, users with different skill-sets are supported. This is also evidenced through the experiment, described in Section 7.2, wherein the use of an EDMP leads to a smaller standard deviation for completing the consumer process. Furthermore, the offering of Infrastructure-as-a-service supports various non-technical user groups as these do not need the skill for setting up a data analytics environment. Thus, Dim 1 is addressed. The offering Software-as-a-Service supports easy and broader access to tools, as required per Dim 2. The last offering, Professional services, enables the development of data and data analytic skills which supports Dim 3. The collaboration functionality offered through the consumer and provider-side functionality like commenting,

Dimension	Description	Addressed
Dim 1	broader access to data	✔
Dim 2	broader access to self-service analytics tools	✔
Dim 3	development of data and analytics kills	✔
Dim 4	collaborative knowledge-sharing	~
Dim 5	promotion of data value	✘

✔ fully addressed ~ partly addressed ✘ not addressed

Table 7.5: Addressing of Data Democratization Dimensions.

rating, and documentation features address Dim 4, concerning collaborative knowledge-sharing to a certain degree. The last dimension, the promotion of data value, is partly covered through the collaboration functionality as a user may stress the value of data through these. Yet, we see the comprehensive realization of this dimension as part of the development of the corporate culture, not in a tool. Hence, the EDMP supports four out of the five data democratization dimensions through its offerings and functionality.

As explained in Section 4.6, introducing an EDMP in a company also poses a few challenges. In this context, we briefly highlight which of the challenges affect which democratization dimension, as also summarized in Table 7.6. Furthermore, we discuss to what extent the concepts provided throughout this thesis already address these challenges.

The first challenge entails the *lack of incentives* for data providers to share their data. As the monetization of data hinders the goal of data democratization, this has to be reduced in the EDMP, even though this poses the main incentive for data providers. Without incentives, the providers may choose to not invest the initial effort in making their data available and accessible. This would directly impact the data democratization dimension one of enabling broader access to data. While the concepts in this thesis do not explicitly address incentives, the concept for distinguishing assets and products in the EDMP, presented in Section 5.2, does reduce the effort of providers and thus contributes to this topic.

Challenge Description	Affected Democratization Dimension	Addressed
Lack of incentives	Dim 1	~
Retaining or passing data ownership	Dim 1 + 3	✘
Flooding EDMP with unusable data	Dim 1	~
Integrating EDMP in company tool and system landscape	Dim 1 + 2	✔

✔ fully addressed ~ partly addressed ✘ not addressed

Table 7.6: Relation of the EDMP Specific Challenges to Data Democratization.

The question *how data ownership can be retained or passed* constitutes the second challenge. This relates to dimension one since this issue arises with broader access to data and data sharing. As we learned through interviews with the industrial manufacturer presented in Section 3.1, employees currently seem to be hesitant to take the role of a data owner because they do not fully understand what this role entails. This challenge therefore also relates to dimension three of developing data skills. In this case, this refers to acquiring the knowledge of what has to be done as a data owner, and how to do this. This challenge is currently not addressed through the concepts presented in this thesis and thus constitutes future work.

With the goal of making as much data as possible available through the EDMP, there is the risk of *flooding the marketplace with unusable data*. In this case, the marketplace could turn into a data dump, making it increasingly difficult for data consumers to find relevant and high-quality data. Therefore this challenge impedes data democratization dimension one as data may become less accessible. However, by differentiating data assets and data products as described in Section 5.2 this issue is addressed, as data products are specifically prepared for access and consumption by data consumers and thus, constitute reliable data. Having said this, data assets are not necessarily unusable data, if not documented properly these may be difficult to find and understand, yet this specifically depends on how much effort data providers invest when registering these in the catalog or marketplace. Thus, by providing guidance for data providers this challenge can be solved.

The last challenge involves the *integration of the Enterprise Data Marketplace into the existing system landscape*. By integrating the marketplace with the available data sources more of the data is made available, and by integrating it with e.g., analytics tools, the data may be made available within these tools, thus supporting dimensions one and two. How the marketplace can fit into the tool and system landscape and how it integrates with metadata management tools is addressed in Chapter 5 and Chapter 6. Hence, three of the four challenges are addressed through the concepts presented in this thesis.

In addition to supporting the democratization dimensions, the EDMP functionality as described in Section 4.4.2 also addresses almost all parts of the

definition of data democratization. This definition states that data democratization involves empowering and motivating the majority of company employees to find, understand, access, use, and share data across the enterprise, in a secure and compliant way [LLF21]. The discovery features enable finding, the data trading features enable accessing, service publishing features enable sharing, cataloging, and dataset-specific-metadata support understanding, and lastly, the governance, privacy, and security aspects enable this in a secure and compliant way.

We have therefore shown that the EDMP explicitly supports four out of five data democratization dimensions through its offerings and functionality, that the concepts presented throughout the scope of this thesis address three of the four EDMP challenges, and that the EDMP functionality generally assists in democratizing data. Therefore, we conclude that the EDMP is specifically suited to support the realization of data democratization initiatives within an enterprise.

7.4 Summary

Within this section, the applicability and feasibility of the concepts presented throughout this thesis were demonstrated through a prototypical Enterprise Data Marketplace. Therein concepts of the Chapters 3 to 6 were taken into account, displaying key aspects of the consumer and provider workflows, the EDMP offerings and functionality, architectures, and enterprise integration. Furthermore, based on an experiment we have established that an EDMP makes the consumer journey significantly more efficient and less complex. The EDMP therefore considerably benefits the goal of greater accessibility of data within a company. Contrary to the expectations, it does not explicitly improve the effectiveness of data consumers, yet whether the effectiveness might be improved under circumstances in which data consumers have a time constraint has yet to be determined. Lastly, a discussion yielded that the concepts presented throughout this thesis address three out of the four EDMP challenges, that the EDMP addresses all the data consumer challenges and three out of four data provider challenges, and benefits four out of the five data democratization dimensions.

Through this chapter, we have validated the feasibility and advantages of the concepts put forth in this thesis, have exhibited the benefits of employing an EDMP, and have established that the EDMP is well-suited for the effective realization of data democratization within organizations.

CHAPTER 8

CONCLUSION AND FUTURE WORK

Companies today have increasing amounts of data at their disposal, most of which is not used, leaving the data value unexploited. In order to leverage the data value, the data must be democratized, i.e., made available to the company employees. In this context, the use of Enterprise Data Marketplaces (EDMPs), platforms for trading data within a company, are proposed. However, specifics of EDMPs have not been investigated in detail so far. In this dissertation, the EDMP is examined as a platform for democratizing company data. By introducing the EDMP as a distinct marketplace type and providing insights into various dimensions ranging from its characteristics, over requirements, to, a platform architecture, it is the objective of this dissertation to lay the foundation for establishing the EDMP within a company. Pertaining to this objective, the contributions of this work are summarized in Section 8.1. In closing, an outlook on future work is given in Section 8.2.

8.1 Summary of the Contributions

The central research question of this thesis considers whether the Enterprise Data Marketplace is suited to address data democratization based on the extent to which it enables company employees with varying skill-sets to find,

understand, access, use, and share data within a company. Four research goals RG1-RG4 were defined, which jointly address this research question. The research goals encompass the identification of the processes and challenges company employees face in finding, understanding, accessing, and sharing data in their enterprise (RG1) without an EDMP, the identification of the distinctive aspects of an EDMP (RG2), establishing an architectural foundation for building an EDMP (RG3), and lastly, the goal of leveraging existent metadata in the EDMP (RG4). Each of these goals is backed by a set of research contributions which are summarized in the following:

C1.1 Identification of the data provider journey for sharing data, and the data consumer journey for finding, understanding, and accessing data [EGH+22a; EGHS22]

Based on expert interviews with a globally active industrial manufacturer and a literature study, a generalized data provider and data consumer journey are introduced. The data provider journey illustrates the steps, involved participants, and tools required to make data available. Complementary, the data consumer journey yields a number of steps, the participants, and potential tools required for finding, understanding, and accessing data. By this means, insight is gained into the workflows of data democratization dimension one, for broader access to data, independent of the use of data marketplaces. This serves as the basis for considering how a marketplace influences these workflows. To this end, an EDMP prototype is implemented, and an experiment is conducted to gain insight into the effect the introduction of an EDMP has. The prototype and experiment reveal that the EDMP significantly improves the data consumer journey in terms of efficiency and reduces the complexity of the workflow. The data provider journey becomes more efficient in that the provider has less effort in registering data, benefiting data democratization, as this may lead to an increased amount of data that is made available in the enterprise.

C1.2 Identification of challenges throughout the data provider and data consumer journey [EGH+22a; EGHS22]

The data consumer and data provider journeys give rise to a number of challenges that arise throughout the respective workflows and hinder data democratization. On the data provider side, the challenges encompass the assembly of metadata, the required effort for supplying provisioning options, and the necessity to register data in various publishing tools, as well as that this process involves several parties. Similarly, on the data consumer side, the number of involved parties poses a challenge, in addition to metadata required for understanding and selecting datasets being distributed across the tool landscape and lacking integration across the toolchain throughout the overall process. On the basis of these challenges, the EDMP is assessed as to whether it assists the data consumer and data provider by addressing the challenges. The discussion yields that three out of three consumer challenges and three out of four data provider challenges are improved through the use of an EDMP.

C2.1 Establishing a type distinction by providing a tailored EDMP definition and identifying the EDMP specific characteristics [EGH+23]

Initially, it is clarified what an EDMP is, by complementing the EDMP definition of Wells [Wel18] with facets on the intended scope of offered datasets, the variety of targeted users, and also the range of different datasets and services which are to be supported in this marketplace. The ability to distinguish the EDMP from external data marketplaces is provided through a classification framework, based on which the distinctive profile of an EDMP's characteristics is demonstrated. Thereby the EDMP is highlighted as a distinct type of data marketplace.

C2.2 Identification of EDMP requirements [EGH+23; EGHS22]

The required EDMP offerings, the required functionality as well as requirements concerning the integration into an existing enterprise system and tool landscape are listed as part of this contribution. Based on the type definition and the defining characteristics, it is highlighted which of the offerings and functionality are specific to the EDMP, as opposed

to external data marketplaces. Amongst others, this entails offerings such as infrastructure and software-as-a-service to also assist less data-literate employees. In addition, functionality like metadata management and privacy and security aspects may be relevant in a larger scope or more challenging to implement in the company-internal context. Requirements concerning the integration into the existing company tool and system landscape are distinctive for the EDMP, as an external data marketplace will not be able to tightly integrate with a variety of company landscapes. The distinctive requirements provide the basis for designing and consequently implementing an EDMP.

C2.3 Identification of EDMP specific challenges [EGH+22b]

With distinctive characteristics and a specialized setting, i.e., the company-internal context, there are a number of challenges that must be resolved in order to effectively use an EDMP. This includes topics such as a lack of incentives for data providers to share their data, as monetization is less prevalent in the company-internal context as it hinders the goal of data democratization. As discussed in the scope of the evaluation, the concepts presented throughout this dissertation address the challenge of integrating an EDMP in the company system and tool landscape, partly address the issue of lacking incentives as well as partly address the issue of flooding the EDMP with unusable data. The issue of retaining and passing data ownership is not addressed and thus subject to future work.

C3.1 Derivation of an enterprise integration architecture for the EDMP [EGH+23]

A distinctive aspect of an EDMP is that it can tightly integrate into a company's system and tool landscape, as external stand-alone marketplaces, for instance, are usually not tightly connected with various data management systems and tools within each participating company. In the scope of this contribution, an integration architecture demonstrates how the EDMP fits into the company's IT landscape, how it should interact with a variety of systems and tools and it is elaborated how this integration is advantageous. Thence, existent functionality is reused,

there is a comprehensive view on metadata, the metadata management effort and errors are reduced, and there is less redundant data.

C3.2 Designing of an Enterprise Data Marketplace platform architecture [EGH+23]

In order to establish a basis for building an EDMP, a platform architecture is presented that reflects the idiosyncrasies of integrating the EDMP with an existent company's infrastructure. The platform architecture outlines the components and services required in a data marketplace and which of these can be implemented differently within an EDMP.

C4.1 An approach for integrating with different tools

While contribution C3.1 indicates through an architectural view how the EDMP can be integrated in the company tool and system landscape, this contribution features a plugin-based approach for integrating the EDMP with different metadata-based tools. It enables connecting a variety of tools to the EDMP so selected metadata can be extracted from these, enables mapping this metadata to the according data assets and standardizing it. This approach lays the foundation for leveraging distributed metadata across a company's system and tool landscape as it makes the metadata available to the EDMP.

C4.2 An approach for supporting diverse metadata per dataset and visualizing this metadata in an integrated view

In addition to making existent metadata available to the EDMP, as tackled through contribution C4.1, the EDMP must also be able to support all the metadata provided through the tools and systems and must also make this metadata available to the data consumer. In this regard, the plugin-based integration approach is extended through another approach based on metadata templates for enabling a diverse set of metadata per data asset, and an approach for displaying an integrated view on this diverse set of metadata. In conjunction, these concepts enable realizing the metadata management, accentuated as distinctive functionality in the EDMP so that the advantages of integrating the EDMP in the company's tool and system landscape as highlighted in contribution C3.1 are harnessed.

We, therefore, achieved research goal RG1 involving the exploration of the current state and challenges without an EDMP and revealed the necessity to explore new solutions, such as a data marketplace. In accomplishing research goal RG2 by identifying the distinct aspects of the EDMP, we set forth to not only assess how these alleviate the previously highlighted issues but also draft a set of requirements that benefit the data democratization objective. The contents of RG2 also served as a basis for establishing the foundation for building an EDMP (RG3) through an enterprise integration and platform architecture. Again, in the realization of RG3, the architectures are tailored to address the current challenges of RG1. The fulfillment of research goal RG4 yields an in-depth insight into how to exploit the potential advantage of existent metadata and thus one of the distinguishing aspects given through RG2. We have thereby revealed the need for an EDMP, demonstrated how it addresses current issues, what its benefits are, and how to implement it. Concluding, we have answered the research questions and demonstrated that introducing the EDMP significantly benefits a company as it effectively supports data democratization, by enabling employees with varying skill-sets to find, understand, access, use, and share data within a company.

8.2 Future Work

There are a number of remaining research topics in the context of the Enterprise Data Marketplace which have not been addressed in the scope of this dissertation. Future work includes, but is not limited to the following topics:

Open challenges in the EDMP context: While the concepts presented in this dissertation address aspects of three of the four presented EDMP challenges, a few sub-aspects remain unresolved. Firstly, this includes the topic of effectively incentivizing data providers to share their data in the company-internal context by another means as monetization. Secondly, the topic of handling data ownership when data is moved through a chain of use cases remains a challenge. Lastly, how to retain the quality of the EDMP offerings even as the EDMP spans the scope of the entire company data, has to be addressed.

Privacy and security in the EDMP: We have emphasized that both metadata management, as well as privacy and security, are especially relevant in the EDMP context. As this dissertation focuses on metadata management aspects, privacy and security aspects have not been discussed in depth. For instance, in offering the whole scope of a company's data, the EDMP includes the highly confidential or also personal data in the company. The marketplace must ensure that it is used and shared for only approved purposes. For example, some parties may access the entire dataset, other parties may access an anonymized version of the data, and some may not be allowed to know that this data exists. Therefore, one open topic in this regard is how the EDMP enables a demand-oriented provision of data while remaining compliant with legal regulations like the General Data Protection Regulation (GDPR). In this regard, existent work in which, for instance, the demand-oriented generation of data products is considered with regard to privacy and security [Sta23], or the application of privacy filters is evaluated for different use cases [SGB+22], could be considered for data marketplaces in the company-internal context.

The EDMP in other environments like the data mesh: As we have explained in Chapter 2, Background, there are research areas in the context of data exchanges that overlap with the EDMP, such as the data mesh. In this regard, an exploration of how the EDMP behaves in the context of the organizational paradigm data mesh also constitutes a research direction. Since the EDMP trades data products, it would be of interest to consider what aspects it must take into account in order to support the data products as defined in the data mesh. Furthermore, it may be of interest to explore how the EDMP ties in with the self-serve platform in the data mesh concept.

Author Publications

[1] R. Eichler, C. Gröger, E. Hoos, C. Stach, H. Schwarz, and B. Mitschang. "Introducing the Enterprise Data Marketplace: A Platform for Democratizing Company Data." In: *Journal of Big Data* 10 (2023). DOI: 10.1186/s40537-023-00843-z.

[2] C. Stach, R. Eichler, and S. Schmidt. "A Recommender Approach to Enable Effective and Efficient Self-Service Analytics in Data Lakes." In: *Datenbank-Spektrum* 23.2 (2023), pages 123–132. DOI: 10.1007/s13222-023-00443-4.

[3] R. Eichler, C. Göger, E. Hoos, H. Schwarz, and B. Mitschang. "Data Shopping – How an Enterprise Data Marketplace supports Data Democratization in Companies." In: *CAiSE Forum on Intelligent Information Systems*. CAiSE Forum '22. 2022, pages 19–26. DOI: 10.1007/978-3-031-07481-3_3.

[4] R. Eichler, C. Gröger, E. Hoos, and H. Schwarz. "From Data Asset to Data Product – The Role of the Data Provider in the Enterprise Data Marketplace." In: *Symposium and Summer School on Service-Oriented Computing*. SummerSoc '22. 2022, pages 119–138. DOI: 10.1007/978-3-031-18304-1_7.

[5] R. Eichler, C. Gröger, E. Hoos, C. Stach, and H. Schwarz. "Establishing the Enterprise Data Marketplace: Characteristics, Architecture, and Challenges." In: *Workshop on Data Science for Data Marketplaces in Conjunction with the International Conference on Very Large Data Bases*. DSDM '22. 2022.

[6] C. Stach, J. Bräcker, R. Eichler, C. Giebler, and B. Mitschang. "Simplified Specification of Data Requirements for Demand-Actuated Big Data

Refinement." In: *Journal of Data Intelligence* 3.3 (2022), pages 366–400. DOI: 10.26421/JDI3.3-5.

[7] R. Eichler, C. Giebler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang. "Enterprise-Wide Metadata Management: An Industry Case on the Current State and Challenges." In: *International Conference on Business Information Systems*. BIS '21. 2021, pages 269–279. DOI: 10.52825/bis.v1i.47.

[8] R. Eichler, C. Giebler, C. Gröger, H. Schwarz, and B. Mitschang. "Modeling Metadata in Data Lakes – A Generic Model." In: *Data & Knowledge Engineering* 136 (2021). DOI: 10.1016/j.datak.2021.101931.

[9] C. Giebler, C. Gröger, E. Hoos, R. Eichler, H. Schwarz, and B. Mitschang. "The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture." In: *Fachtagung Datenbanksysteme für Business, Technologie und Web*. BTW '21. 2021, pages 351–370. DOI: 10.18420/btw2021-19.

[10] C. Stach, J. Bräcker, R. Eichler, C. Giebler, and B. Mitschang. "Demand-Driven Data Provisioning in Data Lakes: BARENTS — A Tailorable Data Preparation Zone." In: *International Conference on Information Integration and Web Intelligence*. iiWAS '21. **iiWAS 2021 Best Paper Award**. 2021, pages 191–202. DOI: 10.1145/3487664.3487784.

[11] R. Eichler, C. Giebler, C. Gröger, H. Schwarz, and B. Mitschang. "Handle - a Generic Metadata Model for Data Lakes." In: *International Conference on Big Data Analytics and Knowledge Discovery*. DaWaK '20. 2020, pages 73–88. DOI: 10.1007/978-3-030-59065-9_7.

[12] C. Giebler, C. Gröger, E. Hoos, R. Eichler, H. Schwarz, and B. Mitschang. "Data Lakes auf den Grund gegangen." In: *Datenbank-Spektrum* 20.1 (2020), pages 57–69. DOI: 10.1007/s13222-020-00332-0.

[13] C. Stach, J. Bräcker, R. Eichler, C. Giebler, and C. Gritti. "How to Provide High-Utility Time Series Data in a Privacy-Aware Manner: A VAULT to Manage Time Series Data." In: *International Journal on Advances in Security* 13.3 & 4 (2020), pages 88–108.

Supervised Student Projects

[1] S. Chowdhury and M. Debnath. “Implementation of a Multi-layered Framework for the Holistic Management of Monitoring Data.” Practical Course. University of Stuttgart, Nov. 2020.

[2] L. Schuiki. “Needs-based data provision in a trustworthy data science platform.” Bachelor Thesis. University of Stuttgart, Nov. 2020.

[3] H. Düsseldorf. “Development of a data provisioning platform for data-intensive IoT applications.” Bachelor Thesis. University of Stuttgart, Dec. 2020.

[4] G. Berezhna, C. Eberlein, M. Fischer, L. Glandier, K. Mendes Guido, F. S. Müller, and A. Schmid. “Enterprise Data Marketplace: Design and Implementation of a Platform for the Exchange of Data and Data-related Services.” Study Project. University of Stuttgart, Sept. 2021.

[5] G. Berezhna. “Developement of Metadata Templates for Enterprise Data Marketplaces.” Bachelor Thesis. University of Stuttgart, July 2022.

[6] G. Berezhna, M. Fischer, and F. S. Müller. “The Modular Connection of Metadata Management Tools to the Data Marketplace.” Bachelor Research Project. University of Stuttgart, July 2022.

[7] M. Weilinger, J.-P. Thewes, and M. Hack. “Evaluation of Data Marketplace Platforms.” Practical Course. University of Stuttgart, Apr. 2023.

[8] P. Fuchs. “Design of a System Landscape for Data Management in the Data Mesh Approach.” Master Thesis. University of Stuttgart, June 2023.

[9] A. Wurster. “Development of a Framework for Creating Data Products in the Data Mesh.” Bachelor Thesis. University of Stuttgart, June 2023.

[10] L. Schuiki. “Erweiterung des Data-Mesh-Konzepts um Datenschutzaspekte.” Master Thesis. University of Stuttgart, Nov. 2023.

Bibliography

[ABC+20] R. Albertoni, D. Browning, S. Cox, A. G. Beltran, A. Perego, and P. Winstanley. *Data Catalog Vocabulary (DCAT) - Version 2*. 2020. URL: https://www.w3.org/TR/vocab-dcat-2/ (cit.on p. 33).

[ABF+18] S. Alpers, S. Betz, A. Fritsch, A. Oberweis, G. Schiefer, and M. Wagner. "Citizen Empowerment by a Technical Approach for Privacy Enforcement." In: *International Conference on Cloud Computing and Services Science*. CLOSER' 18. 2018, pages 589–595. DOI: 10.5220/0006789805890595 (cit.on p. 69).

[ACD+07] C. A. Ardagna, M. Cremonini, E. Damiani, S. De Capitani di Vimercati, and P. Samarati. "Location Privacy Protection Through Obfuscation-Based Techniques." In: *Data and Applications Security XXI*. DBSec' 07. 2007, pages 47–60. DOI: 10.1007/978-3-540-73538-0_4 (cit.on p. 69).

[ADS19] A. Agarwal, M. Dahleh, and T. Sarkar. "A Marketplace for Data: An Algorithmic Solution." In: *ACM Conference on Economics and Computation*. EC '19. 2019, pages 701–726. DOI: 10.1145/3328526.3329589 (cit.on p. 54).

[AFSC23] M. S. Azari, F. Flammini, S. Santini, and M. Caporuscio. "A Systematic Literature Review on Transfer Learning for Predictive Maintenance in Industry 4.0." In: *IEEE Access* 11 (2023), pages 12887–12910. DOI: 10.1109/ACCESS.2023.3239784 (cit.on p. 39).

[AG20] P. Awasthi and J. George. "A case for Data Democratization." In: *Americas Conference on Information Systems*. AMCIS '20. 2020, page 23 (cit.on pp. 14, 27, 43, 61).

[AL22] S. A. Azcoitia and N. Laoutaris. “A Survey of Data Marketplaces and Their Business Models.” In: *SIGMOD Rec.* 51.3 (2022), pages 18–29. DOI: 10.1145/3572751.3572755 (cit.on pp. 24, 53, 54, 74).

[ALL20] A. S. Alrawahi, K. Lee, and A. Lotfi. “AMACoT: A Marketplace Architecture for Trading Cloud of Things Resources.” In: *IEEE Internet of Things Journal* 7.3 (2020), pages 2483–2495. DOI: 10.1109/JIOT.2019.2957441 (cit.on pp. 54, 88, 101).

[AN20] A. Alsharif and M. Nabil. “A Blockchain-based Medical Data Marketplace with Trustless Fair Exchange and Access Control.” In: *IEEE Global Communications Conference*. GLOBECOM’ 20. 2020, pages 1–6. DOI: 10.1109/GLOBECOM42002.2020.9348192 (cit.on p. 24).

[AOA17] J. Attard, F. Orlandi, and S. Auer. “Exploiting the Value of Data through Data Value Networks.” In: *International Conference on Theory and Practice of Electronic Governance*. ICEGOV’ 17. 2017, pages 475–484. DOI: 10.1145/3047273.3047299 (cit.on pp. 23, 24).

[AS19] E. Ahmed and M. Shabani. “DNA Data Marketplace: An Analysis of the Ethical Concerns Regarding the Participation of the Individuals.” In: *Frontiers in Genetics* 10 (2019). DOI: 10.3389/fgene.2019.01107 (cit.on p. 24).

[ASvB19] R. Abraham, J. Schneider, and J. vom Brocke. “Data Governance: A Conceptual Framework, Structured Review, and Research Agenda.” In: *International Journal of Information Management* 49 (2019), pages 424–438. DOI: 10.1016/j.ijinfomgt.2019.07.008 (cit.on p. 22).

[ASZ+19] C. Anhalt-Depies, J. L. Stenglein, B. Zuckerberg, P. A. Townsend, and A. R. Rissman. “Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science.” In: *Biological Conservation* 238 (2019). DOI: 10.1016/j.biocon.2019.108195 (cit.on p. 69).

[Bar18] S. Barns. “Smart cities and urban data platforms: Designing interfaces for smart governance.” In: *City, Culture and Society* 12 (2018), pages 5–12. DOI: 10.1016/j.ccs.2017.09.006 (cit.on p. 24).

[BCMC22] A. Bernasconi, A. Canakoglu, M. Masseroli, and S. Ceri. “META-BASE: A Novel Architecture for Large-Scale Genomic Metadata Integration.” In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.1 (2022), pages 543–557. DOI: 10.1109/TCBB.2020.2998954 (cit.on p. 104).

[BG00] B. E. Bargmeyer and D. W. Gillman. “Metadata Standards and Metadata Registries: An Overview.” In: *International Conference on Establishment Surveys II*. ICES II’ 00. 2000 (cit.on p. 33).

[Cao17] L. Cao. “Data Science: A Comprehensive Overview.” In: *ACM Comput. Surv.* 50.3 (2017). DOI: 10.1145/3076253 (cit.on pp. 13, 14).

[CCB19] A. Colman, M. J. M. Chowdhury, and M. Baruwal Chhetri. “Towards a Trusted Marketplace for Wearable Data.” In: *IEEE International Conference on Collaboration and Internet Computing*. CIC ’19. 2019, pages 314–321. DOI: 10.1109/CIC48465.2019.00044 (cit.on p. 23).

[CGK12] G. Charness, U. Gneezy, and M. A. Kuhn. “Experimental methods: Between-subject and within-subject design.” In: *Journal of Economic Behavior & Organization* 81.1 (2012), pages 1–8. DOI: 10.1016/j.jebo.2011.08.009 (cit.on p. 140).

[CKD+04] C. Clifton, M. Kantarcioundefinedlu, A. Doan, G. Schadow, J. Vaidya, A. Elmagarmid, and D. Suciu. “Privacy-Preserving Data Integration and Sharing.” In: *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. DMKD ’04. 2004, pages 19–26. DOI: 10.1145/1008694.1008698 (cit.on p. 69).

[CSN+14] M. Chessell, F. Scheepers, N. Nguyen, R. van Kessel, and R. van der Starre. *Governing and Managing Big Data for Analytics and Decision Makers*. Technical report. IBM Redguides for Business Leaders, 2014 (cit.on p. 98).

[Deh22] Z. Dehghani. *Data Mesh: Delivering Data-driven Value at Scale*. O’Reilly Media, 2022. ISBN: 9781492092391 (cit.on p. 25).

[DMV22] S. W. Driessen, G. Monsieur, and W.-J. Van Den Heuvel. “Data Market Design: A Systematic Literature Review.” In: *IEEE Access* 10 (2022), pages 33123–33153. DOI: 10.1109/ACCESS.2022.3161478 (cit.on pp. 22–24, 52, 54, 74, 75, 83).

[DMvdH23] S. Driessen, G. Monsieur, and W.-J. van den Heuvel. “Data Product Metadata Management: an Industrial Perspective.” In: *Service-Oriented Computing*. ICSOC Workshops ’22. 2023, pages 237–248. DOI: 10.1007/978-3-031-26507-5_19 (cit.on pp. 53, 101, 103, 104).

[DMvdHT21] S. W. Driessen, G. Monsieur, W.-J. van den Heuvel, and D. A. Tamburri. "Validated Data Quality Assessment with "Skin in the Game": A Smart Contract Approach." In: *Service-Oriented Computing*. SummerSOC '21. 2021, pages 119–130. DOI: 10.1007/978-3-030-87568-8_8 (cit.on p. 23).

[EGG+20] R. Eichler, C. Giebler, C. Gröger, H. Schwarz, and B. Mitschang. "Handle - a Generic Metadata Model for Data Lakes." In: *International Conference on Big Data Analytics and Knowledge Discovery*. DaWaK '20. 2020, pages 73–88. DOI: 10.1007/978-3-030-59065-9_7 (cit.on pp. 96, 98–100).

[EGG+21a] R. Eichler, C. Giebler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang. "Enterprise-Wide Metadata Management: An Industry Case on the Current State and Challenges." In: *International Conference on Business Information Systems*. BIS '21. 2021, pages 269–279. DOI: 10.52825/bis.v1i.47 (cit.on pp. 17, 25, 31, 41, 46, 48, 53, 66, 68, 78, 80, 92).

[EGG+21b] R. Eichler, C. Giebler, C. Gröger, H. Schwarz, and B. Mitschang. "Modeling Metadata in Data Lakes – A Generic Model." In: *Data & Knowledge Engineering* 136 (2021). DOI: 10.1016/j.datak.2021.101931 (cit.on pp. 29, 30, 32, 96, 99, 100, 111).

[EGH+22a] R. Eichler, C. Göger, E. Hoos, H. Schwarz, and B. Mitschang. "Data Shopping – How an Enterprise Data Marketplace supports Data Democratization in Companies." In: *CAiSE Forum on Intelligent Information Systems*. CAiSE Forum '22. 2022, pages 19–26. DOI: 10.1007/978-3-031-07481-3_3 (cit.on pp. 36, 45, 53, 54, 78, 104, 119, 164, 165).

[EGH+22b] R. Eichler, C. Gröger, E. Hoos, C. Stach, and H. Schwarz. "Establishing the Enterprise Data Marketplace: Characteristics, Architecture, and Challenges." In: *Workshop on Data Science for Data Marketplaces in Conjunction with the International Conference on Very Large Data Bases*. DSDM '22. 2022 (cit.on pp. 52, 77, 104, 119, 129, 166).

[EGH+23] R. Eichler, C. Gröger, E. Hoos, C. Stach, H. Schwarz, and B. Mitschang. "Introducing the Enterprise Data Marketplace: A Platform for Democratizing Company Data." In: *Journal of Big Data* 10 (2023). DOI: 10.1186/s40537-023-00843-z (cit.on pp. 52, 57, 63, 72, 77, 79, 89, 119, 121, 127, 128, 130, 134, 136, 141, 143, 145–148, 165–167).

[EGHS22] R. Eichler, C. Gröger, E. Hoos, and H. Schwarz. “From Data Asset to Data Product – The Role of the Data Provider in the Enterprise Data Marketplace.” In: *Symposium and Summer School on Service-Oriented Computing*. SummerSoc ’22. 2022, pages 119–138. DOI: 10.1007/978-3-031-18304-1_7 (cit.on pp. 36, 39, 52–54, 59, 63, 65, 67, 71, 74, 77, 78, 82, 84, 104, 119, 164, 165).

[Eur16] 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, Apr. 27, 2016 (cit.on pp. 40, 46, 69, 130).

[Fan19] L. Fan. “Practical Image Obfuscation with Provable Privacy.” In: *IEEE International Conference on Multimedia and Expo*. ICME ’19. 2019, pages 784–789. DOI: 10.1109/ICME.2019.00140 (cit.on p. 69).

[FPM+20] G. Fierro, A. K. Prakash, C. Mosiman, M. Pritoni, P. Raftery, M. Wetter, and D. E. Culler. “Shepherding Metadata Through the Building Lifecycle.” In: *ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. BuildSys ’20. 2020, pages 70–79. DOI: 10.1145/3408308.3427627 (cit.on p. 104).

[FRP20] M. Fruhwirth, M. Rachinger, and E. Prlja. “Discovering Business Models of Data Marketplaces.” In: *Hawaii International Conference on System Sciences*. HICSS ’20. 2020, pages 5738–5747. DOI: 10.24251/HICSS.2020.704 (cit.on p. 53).

[FSF20] R. C. Fernandez, P. Subramaniam, and M. J. Franklin. “Data Market Platforms: Trading Data Assets to Solve Data Problems.” In: *Proceedings of the VLDB Endowment* 13.12 (2020), pages 1933–1947. DOI: 10.14778/3407790.3407800 (cit.on pp. 14, 15, 23, 25, 38, 43, 52–54, 62, 74, 75, 88, 104).

[FSS20] K. Figueredo, D. Seed, and V. Subotic. “Preparing for Highly Scalable and Replicable IoT Systems.” In: *IEEE Internet of Things Magazine* 3.3 (2020), pages 94–98. DOI: 10.1109/IOTM.0001.1900022 (cit.on p. 23).

[Fur20] J. Furner. “Definitions of ’Metadata’: A Brief Survey of International Standards.” In: *Journal of the Association for Information Science and Technology* 71.6 (2020), E33–E42. DOI: 10.1002/asi.24295 (cit.on p. 29).

[Gan23] A. Ganti. *Clearinghouse: An Essential Intermediary in the Financial Markets*. 2023. URL: https://www.investopedia.com/terms/c/clearinghouse.asp (cit.on p. 23).

[Gar22] Gartner. *Understand the Role of Data Fabric*. Guide. 2022 (cit.on p. 26).

[GGH+20a] C. Giebler, C. Gröger, E. Hoos, R. Eichler, H. Schwarz, and B. Mitschang. “Data Lakes auf den Grund gegangen.” In: *Datenbank-Spektrum* 20.1 (2020), pages 57–69. DOI: 10.1007/s13222-020-00332-0 (cit.on pp. 32, 98).

[GGH+20b] C. Giebler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang. “A Zone Reference Model for Enterprise-Grade Data Lake Management.” In: *International Enterprise Distributed Object Computing Conference*. EDOC ’20. 2020, pages 57–66. DOI: 10.1109/EDOC49727.2020.00017 (cit.on pp. 32, 70, 98, 100, 122).

[GGH+21] C. Giebler, C. Gröger, E. Hoos, R. Eichler, H. Schwarz, and B. Mitschang. “The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture.” In: *Fachtagung Datenbanksysteme für Business, Technologie und Web*. BTW’21. 2021, pages 351–370. DOI: 10.18420/btw2021-19 (cit.on p. 32).

[GH19] C. Gröger and E. Hoos. “Ganzheitliches Metadatenmanagement im Data Lake: Anforderungen, IT-Werkzeuge und Herausforderungen in der Praxis.” In: *Fachtagung für Datenbanksysteme für Business, Technologie und Web*. BTW’19. 2019, pages 435–452. DOI: 10.18420/btw2019-26 (cit.on pp. 29, 38, 43).

[GMV+21] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. “Datasheets for datasets.” In: *Commun. ACM* 64.12 (2021), pages 86–92. DOI: 10.1145/3458723 (cit.on p. 103).

[GÖM19] C. Gritti, M. Önen, and R. Molva. “Privacy-Preserving Delegable Authentication in the Internet of Things.” In: *ACM/SIGAPP Symposium on Applied Computing*. SAC ’19. 2019, pages 861–869. DOI: 10.1145/3297280.3297365 (cit.on p. 68).

[Gri20] C. Gritti. “Publicly Verifiable Proofs of Data Replication and Retrievability for Cloud Storage.” In: *International Computer Symposium*. ICS ’20. 2020, pages 431–436. DOI: 10.1109/ICS51289.2020.00091 (cit.on p. 68).

[Grö18] C. Gröger. “Building an Industry 4.0 Analytics Platform.” In: *Datenbank-Spektrum* 18 (2018), pages 5–14. DOI: 10.1007/s13222-018-0273-1 (cit.on p. 37).

[Grö21] C. Gröger. “There is no AI without Data.” In: *Communications of the ACM* 64.11 (2021), pages 98–108. DOI: 10.1145/3448247 (cit.on pp. 14, 15, 17, 21, 22, 25, 28, 33, 38, 43, 48, 51, 52, 54, 61, 71, 78, 80, 92, 104).

[GSB+22] C. Ge, W. Susilo, J. Baek, Z. Liu, J. Xia, and L. Fang. “Revocable Attribute-Based Encryption With Data Integrity in Clouds.” In: *IEEE Transactions on Dependable and Secure Computing* 19.5 (2022), pages 2864–2872. DOI: 10.1109/TDSC.2021.3065999 (cit.on p. 68).

[GWF17] D. Grewe, M. Wagner, and H. Frey. “ICN-based open, distributed data market place for connected vehicles: Challenges and research directions.” In: *IEEE International Conference on Communications Workshops*. ICC Workshops’ 17. 2017, pages 265–270. DOI: 10.1109/ICCW.2017.7962668 (cit.on p. 24).

[HES17] D. Henderson, S. Earley, and L. Sebastian-Coleman, editors. *DAMA-DMBOK: Data Management Body of Knowledge*. Basking Ridge, NJ, USA: Technics Publications, 2017. ISBN: 9781634622349 (cit.on pp. 17, 22, 29–32, 97).

[HL19] H. Halpin and I. Lykourentzou. “Crowdsourcing High-Quality Structured Data.” In: *Information Management and Big Data*. SIMBig ’18. 2019, pages 304–319. DOI: 10.1007/978-3-030-11680-4_29 (cit.on p. 23).

[HL23] M. R. Hasan and C. Legner. “Data Product Canvas: A Visual Inquiry Tool Supporting Data Product Design.” In: *Design Science Research for a New Society: Society 5.0*. DESRIST’ 23. 2023, pages 191–205. ISBN: 978-3-031-32807-7. DOI: 10.1007/978-3-031-32808-4_12 (cit.on p. 14).

[HWW23] E. Hechler, M. Weihrauch, and Y. Wu. *Data Fabric and Data Mesh Approaches with AI*. Apress Berkeley, CA, 2023. DOI: 10.1007/978-1-4842-9253-2 (cit.on p. 26).

[Ish20] G. Ishmaev. “The Ethical Limits of Blockchain-Enabled Markets for Private IoT Data.” In: *Philosophy and Technology* 33 (2020), pages 411–432. DOI: 10.1007/s13347-019-00361-y (cit.on p. 24).

[ISO18] ISO/IEC 27000:2018. *Information Technology–Security Techniques–Information Security Management Systems–Overview and Vocabulary*. ISO Standard. Geneva, Switzerland: International Organization for Standardization, 2018 (cit.on p. 68).

[JCZ12] M. Janssen, Y. Charalabidis, and A. Zuiderwijk. “Benefits, Adoption Barriers and Myths of Open Data and Open Government.” In: *Information Systems Management* 29.4 (2012), pages 258–268. DOI: 10.1080/10580530.2012.716740 (cit.on pp. 22, 101).

[JF20] S. Judah and T. Friedman. *Magic Quadrant for Data Quality Tools*. Report. Gartner, 2020 (cit.on pp. 17, 32).

[JO23] N. Jahnke and B. Otto. “Data Catalogs in the Enterprise: Applications and Integration.” In: *Datenbank-Spektrum* 23 (2023), pages 89–96. DOI: 10.1007/s13222-023-00445-2 (cit.on pp. 52, 53, 72, 73, 92, 120).

[JSR22] N. Jahnke, M. Spiekermann, and B. Ramouzeh. *Data Catalogs: Implementing Capabilities for Data Curation, Data Enablement and Regulatory Compliance*. Technical report. 2022 (cit.on p. 31).

[JYJS20] B.-G. Jeong, T.-Y. Youn, N.-S. Jho, and S. U. Shin. “Blockchain-Based Data Sharing and Trading Model for the Connected Car.” In: *Sensors* 20.11 (2020). DOI: 10.3390/s20113141 (cit.on p. 23).

[KGC+16] Y. M. Kassa, J. Gonzalez, Á. Cuevas, R. Cuevas, M. Marciel, and R. González. “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: *International Conference on Availability, Reliability and Security*. ARES ’16. 2016, pages 701–705. DOI: 10.1109/ARES.2016.55 (cit.on p. 88).

[KJAJ23] S. H. Kim, J. H. Jeon, A. Aridi, and B. Jun. “Factors That Affect the Technological Transition of Firms Toward the Industry 4.0 Technologies.” In: *IEEE Access* 11 (2023), pages 1694–1707. DOI: 10.1109/ACCESS.2022.3233390 (cit.on p. 37).

[KLT17] P. Koutroumpis, A. Leiponen, and L. D. W. Thomas. *The (Unfulfilled) Potential of Data Marketplaces*. ETLA Working Papers 53. The Research Institute of the Finnish Economy (ETLA), 2017 (cit.on pp. 24, 53, 56, 62, 88).

[KLT20] P. Koutroumpis, A. Leiponen, and L. D. W. Thomas. "Markets for data." In: *Industrial and Corporate Change* 29.3 (2020), pages 645–660. DOI: 10.1093/ICC/DTAA002 (cit.on pp. 54, 96, 103).

[KPKS18] B. Krishnamachari, J. Power, S. H. Kim, and C. Shahabi. "I3: An IoT Marketplace for Smart Communities." In: *ACM International Conference on Mobile Systems, Applications, and Services*. MobiSys '18. 2018, pages 498–499. DOI: 10.1145/3210240.3223573 (cit.on pp. 54, 88).

[LLEF20] C. Labadie, C. Legner, M. Eurich, and M. Fadler. "FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs." In: *IEEE Conference on Business Informatics*. CBI '20. 2020, pages 201–210. DOI: 10.1109/CBI49978.2020.00029 (cit.on pp. 14, 28, 31, 38, 43, 71, 72, 83).

[LLF21] H. Lefebvre, C. Legner, and M. Fadler. "Data democratization : toward a deeper understanding." In: *International Conference on Information Systems*. ICIS '21. 2021 (cit.on pp. 14, 27, 28, 37, 43, 61, 64, 152, 157, 160).

[LLL+21] J. Liu, J. Lou, J. Liu, L. Xiong, J. Pei, and J. Sun. "Dealer: an end-to-end model marketplace with differential privacy." In: *Proceedings of the VLDB Endowment* 14.6 (2021), pages 957–969. DOI: 10.14778/3447689.3447700 (cit.on p. 54).

[LN06] U. Leser and F. Naumann. *Informationsintegration - Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen*. dpunkt.verlag, 2006. ISBN: 978-3-89864-400-6 (cit.on pp. 82, 116).

[LSR19] S. Lawrenz, P. Sharma, and A. Rausch. "Blockchain Technology as an Approach for Data Marketplaces." In: *International Conference on Blockchain Technology*. ICBCT '19. 2019, pages 55–59. DOI: 10.1145/3320154.3320165 (cit.on p. 54).

[LSV18] J. Lange, F. Stahl, and G. Vossen. “Datenmarktplätze in verschiedenen Forschungsdisziplinen: Eine Übersicht.” In: *Informatik-Spektrum* 41 (2018), pages 170–180. DOI: 10.1007/s00287-017-1044-3 (cit.on pp. 14, 21, 24, 53, 56).

[Mat17] C. Mathis. “Data Lakes.” In: *Datenbank-Spektrum* 17 (2017), pages 289–293. DOI: 10.1007/s13222-017-0272-7 (cit.on p. 32).

[Mez23] S. Mezzetta. *Principles of Data Fabric: Become a data-driven organization by implementing Data Fabric solutions efficiently*. Packt Publishing, 2023. ISBN: 9781804613092 (cit.on p. 26).

[MS19] L. Meisel and M. Spiekermann. *Datenmarktplätze – Plattformen für Datenaustausch und Datenmonetarisierung in der Data Economy*. ISST-Bericht. Fraunhofer ISST, Feb. 18, 2019 (cit.on pp. 14, 21, 53, 56, 57, 62, 96, 103).

[MSLV13] A. Muschalle, F. Stahl, A. Löser, and G. Vossen. “Pricing Approaches for Data Markets.” In: *Enabling Real-Time Business Intelligence*. BIRTE’ 12. 2013, pages 129–144. DOI: 10.1007/978-3-642-39872-8_10 (cit.on p. 23).

[MWZ+19] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. “Model Cards for Model Reporting.” In: *Conference on Fairness, Accountability, and Transparency*. FAT* ’19. 2019, pages 220–229. DOI: 10.1145/3287560.3287596 (cit.on p. 103).

[Obj19] Object Management Group. *OMG Meta Object Facility (MOF) Core Specification - Version 2.5.1*. 2019. URL: https://www.omg.org/spec/MOF/2.5.1 (cit.on p. 29).

[OJS+16] B. Otto, J. Jürjens, J. Schon, S. Auer, N. Menz, S. Wenzel, and J. Cirullies. *Industrial Data Space - Digitale Souveränität Über Daten*. White Paper. 2016 (cit.on p. 27).

[OSTL+19] B. Otto, S. Steinbuß, A. Teuscher, S. Lohmann, et al. *IDS Reference Architecture Model Version 3.0*. Steinbuss, S. (ed.) International Data Spaces Association, 2019 (cit.on pp. 23, 26, 27).

[Ott11] B. Otto. “A Morphology of the Organisation of Data Governance.” In: *European Conference on Information Systems*. ECIS’ 11. 2011 (cit.on p. 22).

[Pei22] J. Pei. “A Survey on Data Pricing: from Economics to Data Science.” In: *IEEE Transactions on Knowledge and Data Engineering* 34.10 (2022), pages 4586–4608. DOI: 10.1109/TKDE.2020.3045927 (cit.on p. 54).

[Pou89] M. Pourahmadi. “Estimation and Interpolation of Missing Values of a Stationary Time Series.” In: *Journal of Time Series Analysis* 10.2 (1989), pages 149–169. DOI: 10.1111/j.1467-9892.1989.tb00021.x (cit.on p. 69).

[PWZ+17] J. Pillmann, C. Wietfeld, A. Zarcula, T. Raugust, and D. C. Alonso. “Novel common vehicle information model (CVIM) for future automotive vehicle big data marketplaces.” In: *IEEE Intelligent Vehicles Symposium*. IV’ 17. 2017, pages 1910–1915. DOI: 10.1109/IVS.2017.7995984 (cit.on p. 24).

[QHV16] C. Quix, R. Hai, and I. Vatov. “Metadata Extraction and Management in Data Lakes With GEMMS.” In: *Complex Systems Informatics and Modeling Quarterly* 9 (2016), pages 67–83. DOI: 10.7250/csimq.2016-9.04 (cit.on p. 98).

[RD18] A. Radhakrishnan and S. Das. “Data Markets for Smart Grids: An Introduction.” In: *IEEE Innovative Smart Grid Technologies - Asia*. ISGT Asia’ 18. 2018, pages 1010–1015. DOI: 10.1109/ISGT-Asia.2018.8467818 (cit.on pp. 23, 24).

[RP19] B. Ramosaj and M. Pauly. “Predicting Missing Values: A Comparative Study on Non-parametric Approaches for Imputation.” In: *Computational Statistics* 34 (2019), pages 1741–1764. DOI: 10.1007/s00180-019-00900-3 (cit.on p. 69).

[RPT+17] D. Roman, J. Paniagua, T. Tarasova, G. Georgiev, D. Sukhobok, N. Nikolov, and T. C. Lech. “proDataMarket: A Data Marketplace for Monetizing Linked Data.” In: *ISWC Posters & Demonstrations and Industry Tracks co-located with the International Semantic Web Conference*. Volume 1963. ISWC Workshop Proceedings ’17. 2017 (cit.on p. 62).

[RRK18] G. S. Ramachandran, R. Radhakrishnan, and B. Krishnamachari. “Towards a Decentralized Data Marketplace for Smart Cities.” In: *IEEE International Smart Cities Conference*. ISC2 ’18. 2018, pages 1–8. DOI: 10.1109/ISC2.2018.8656952 (cit.on pp. 23, 24, 88, 96, 101, 103).

[RS16] D. Roman and G. Stefano. “Towards a Reference Architecture for Trusted Data Marketplaces: The Credit Scoring Perspective.” In: *International Conference on Open and Big Data*. OBD ’16. 2016, pages 95–101. DOI: 10.1109/OBD.2016.21 (cit.on p. 88).

[Sax18] S. Saxena. *Enterprise Data Marketplace: Democratizing Data within Organizations*. White Paper. Tata Consultancy Services, 2018 (cit.on p. 62).

[SBK+17] S. Schmid, A. Bröring, D. Kramer, S. Käbisch, A. Zappa, M. Lorenz, Y. Wang, A. Rausch, and L. Gioppo. “An Architecture for Interoperable IoT Ecosystems.” In: *International Workshop on Interoperability and Open-Source Solutions for the Internet of Things*. InterOSS-IoT ’16. 2017, pages 39–55 (cit.on pp. 54, 88).

[SD20] P. Sawadogo and J. Darmont. “On data lake architectures and metadata management.” In: *Journal of Intelligent Information Systems* 56 (2020). DOI: 10.1007/s10844-020-00608-7 (cit.on p. 98).

[Sea20] Seagate Technology. *Rethink Data – Bessere Nutzung von mehr Unternehmensdaten – vom Netzwerkrand bis hin zur Cloud*. Report. Seagate, 2020 (cit.on pp. 14, 51).

[Seb13] L. Sebastian-Coleman. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. The Morgan Kaufmann Series on Business Intelligence. Elsevier, 2013. DOI: 10.1016/C2011-0-07321-0 (cit.on p. 32).

[SGB+22] C. Stach, C. Gritti, J. Bräcker, M. Behringer, and B. Mitschang. “Protecting Sensitive Data in the Information Age: State of the Art and Future Prospects.” In: *Future Internet* 14.11 (2022). DOI: 10.3390/fi14110302 (cit.on p. 169).

[Sha] M. Shah. *The People’s Data Catalog: Alation Featured as Top Choice in Eckerson’s Latest Report*. URL: https://www.alation.com/blog/the-peoples-data-catalog/ (cit.on p. 155).

[SLR20] P. Sharma, S. Lawrenz, and A. Rausch. “Towards Trustworthy and Independent Data Marketplaces.” In: *International Conference on Blockchain Technology*. ICBCT ’20. 2020, pages 39–45. DOI: 10.1145/3390566.3391687 (cit.on pp. 88, 96, 103, 104).

[SLS+18] L. Sánchez, J. Lanza, J. R. Santana, R. Agarwal, P. G. Raverdy, T. Elsaleh, Y. Fathy, S. Jeong, A. Dadoukis, T. Korakis, S. Keranidis, P. O'Brien, J. Horgan, A. Sacchetti, G. Mastandrea, A. Fragkiadakis, P. Charalampidis, N. Seydoux, C. Ecrepont, and M. Zhao. "Federation of Internet of Things Testbeds for the Realization of a Semantically-Enabled Multi-Domain Data Marketplace." In: *Sensors* 18.10 (2018). DOI: 10.3390/s18103375 (cit.on p. 23).

[SLT+20] M. Spiekermann, S. Lehmann-Brauns, R. Tontsch, B. Otto, and M. Hoffmann. *Datenmarktplätze in Produktionsnetzwerken*. Technical report. Plattform Industrie 4.0, Bundesministerium für Wirtschaft und Energie (BMWi), 2020 (cit.on p. 62).

[SML+23] P. Subramaniam, Y. Ma, C. Li, I. Mohanty, and R. C. Fernandez. *Comprehensive and Comprehensible Data Catalogs: The What, Who, Where, When, Why, and How of Metadata Management*. 2023 (cit.on p. 14).

[Spi19] M. Spiekermann. "Data Marketplaces: Trends and Monetisation of Data Goods." In: *Intereconomics* 54 (2019), pages 208–216. DOI: 10.1007/s10272-019-0826-z (cit.on pp. 22, 53, 56–59, 83).

[spl19] splunk. *The State of Dark Data*. Report. 2019 (cit.on pp. 14, 51).

[SSF+19] P. N. Sawadogo, É. Scholly, C. Favre, É. Ferey, S. Loudcher, and J. Darmont. "Metadata Systems for Data Lakes: Models and Features." In: *European Conference on Advances in Databases and Information Systems*. ADBIS' 19. 2019, pages 440–451. DOI: 10.1007/978-3-030-30278-8_43 (cit.on pp. 32, 98).

[SSV13] F. Schomm, F. Stahl, and G. Vossen. "Marketplaces for data: an initial survey." In: *ACM SIGMOD Record* 42.1 (2013), pages 15–26. DOI: 10.1145/2481528.2481532 (cit.on pp. 53, 54, 56).

[SSVV16] F. Stahl, F. Schomm, G. Vossen, and L. Vomfell. "A classification framework for data marketplaces." In: *Vietnam Journal of Computer Science* 3 (2016), pages 137–143. DOI: 10.1007/s40595-016-0064-2 (cit.on pp. 23, 53, 56).

[SSVV17] F. Stahl, F. Schomm, L. Vomfell, and G. Vossen. "Marketplaces for Digital Data: Quo Vadis?" In: *Computer and Information Science* 10.4 (2017), pages 22–37. DOI: 10.5539/cis.v10n4p22 (cit.on pp. 24, 53, 56).

[Sta23] C. Stach. “Data Is the New Oil—Sort of: A View on Why This Comparison Is Misleading and Its Implications for Modern Data Administration.” In: *Future Internet* 15.2 (2023). DOI: 10.3390/fi15020071 (cit.on pp. 69, 169).

[STWO18] M. Spiekermann, D. Tebernum, S. Wenzel, and B. Otto. “A Metadata Model for Data Goods.” In: *Multikonferenz Wirtschaftsinformatik*. MKWI ’18. 2018, pages 326–337 (cit.on p. 103).

[TA23] T. M. Takang and A. O. Amaechi. “Considerations for a Planned Democratizing Data Framework for Valid and Trusted Data.” In: *Journal of Data Analysis and Information Processing* 11.3 (2023), pages 240–261. DOI: 10.4236/jdaip.2023.113013 (cit.on p. 51).

[TL18] K. Täuscher and S. M. Laudien. “Understanding platform business models: A mixed methods study of marketplaces.” In: *European Management Journal* 36.3 (2018), pages 319–329. DOI: 10.1016/j.emj.2017.06.005 (cit.on p. 53).

[WDA+16] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. “The FAIR Guiding Principles for scientific data management and stewardship.” In: *Scientific Data* 3 (2016). DOI: 10.1038/sdata.2016.18 (cit.on p. 28).

[Wel17] D. Wells. *The Rise of the Data Marketplace: Data as a Service*. Report. Eckerson Group, 2017 (cit.on pp. 14, 15, 52, 54, 61, 62, 88).

[Wel18] D. Wells. *Dynamic Data Marketplace: Fast Data for Fast Business*. Report. Eckerson Group, 2018 (cit.on pp. 22, 48, 51, 54, 58, 78, 104, 165).

[WLYH19] Z.-J. Wang, C.-H. V. Lin, Y.-H. Yuan, and C.-C. J. Huang. "Decentralized Data Marketplace to Enable Trusted Machine Economy." In: *IEEE Eurasia Conference on IOT, Communication and Engineering*. ECICE' 19. 2019, pages 246–250. DOI: 10.1109/ECICE47484.2019.8942729 (cit.on p. 23).

[WZJT21] Z. Wang, Z. Zheng, W. Jiang, and S. Tang. "Blockchain-Enabled Data Sharing in Supply Chains: Model, Operationalization, and Tutorial." In: *Production and Operations Management* 30.7 (2021), pages 1965–1985. DOI: 10.1111/poms.13356 (cit.on p. 24).

[YW10] X. Yu and Q. Wen. "A View about Cloud Data Security from Data Life Cycle." In: *International Conference on Computational Intelligence and Software Engineering*. CISE' 10. 2010, pages 1–4. DOI: 10.1109/CISE.2010.5676895 (cit.on p. 30).

[YZK+17] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan. "iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning." In: *IEEE Transactions on Information Forensics and Security* 12.5 (2017), pages 1005–1016. DOI: 10.1109/TIFS.2016.2636090 (cit.on p. 69).

[ZDED17] E. Zaidi, G. De Simoni, R. Edjlali, and A. D. Duncan. *Data Catalogs Are the New Black in Data Management and Analytics*. Gartner Research. Gartner, 2017 (cit.on pp. 14, 31, 55, 71, 83).

[ZG18] J. Zeng and K. W. Glaister. "Value creation from big data: Looking inside the black box." In: *Strategic Organization* 16.2 (2018), pages 105–140. DOI: 10.1177/1476127017697510 (cit.on pp. 38, 43).

[ZMWC19] Z. Zheng, W. Mao, F. Wu, and G. Chen. "Challenges and Opportunities in IoT Data Markets." In: *International Workshop on Social Sensing*. SocialSense '19. 2019, pages 1–2. DOI: 10.1145/3313294.3313378 (cit.on p. 54).

[ZTTR21] M. Zasadzinski, M. Theodoulou, M. Thurner, and K. Ranganath. "The Trip to The Enterprise Gourmet Data Product Marketplace through a Self-service Data Platform." In: *arXiv preprint arXiv:2107.13212* (2021) (cit.on pp. 53, 88).

[ZYC+19] H. Zhu, Y. Yuan, Y. Chen, Y. Zha, W. Xi, B. Jia, and Y. Xin. “A Secure and Efficient Data Integrity Verification Scheme for Cloud-IoT Based on Short Signature.” In: *IEEE Access* 7 (2019), pages 90036–90044. DOI: 10.1109/ACCESS.2019.2924486 (cit.on p. 68).

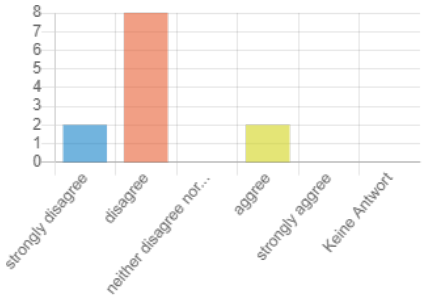
All URLs were last accessed on February 04, 2024.

List of Figures

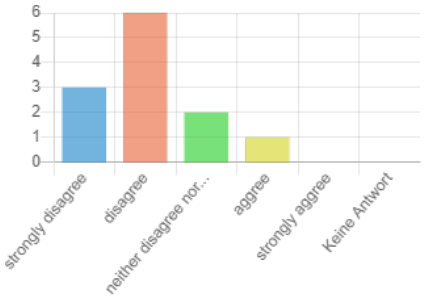
List of Tables

EXPERIMENT QUESTIONNAIRES: QUESTIONS AND RESPONSES

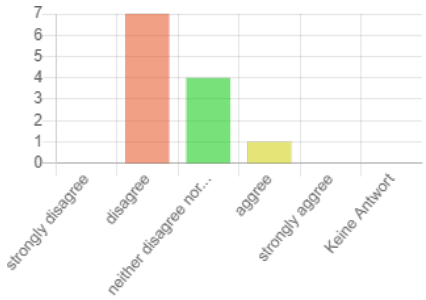
A.1 Scenario S1: Without the Use of an EDMP



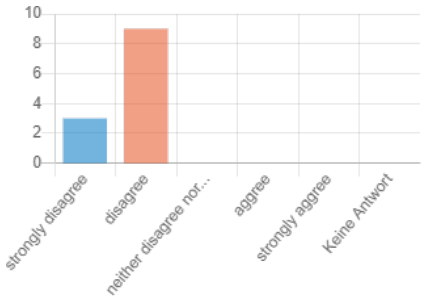
It was easy to find a dataset that matched the use case requirements.



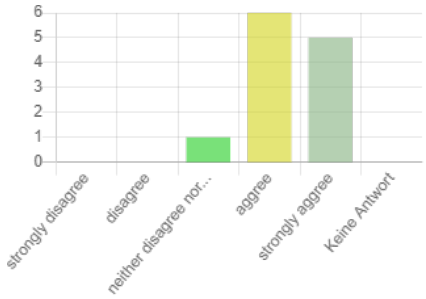
It was easy to find the metadata you were looking for in the data catalog.



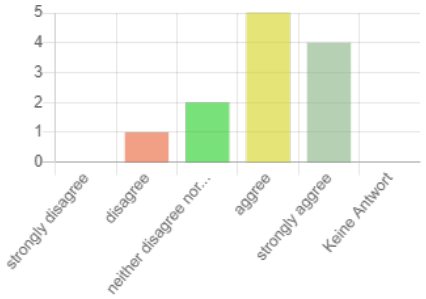
It was easy to find the metadata you were looking for in the Data Quality Tool.



The process for finding and identifying a relevant dataset was intuitive.



The process for finding and identifying a relevant dataset was cumbersome (umständlich).

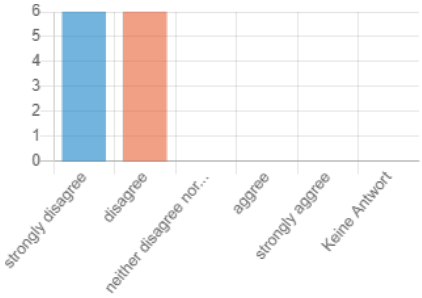


The process for finding and identifying a relevant dataset was laborious (aufwändig).

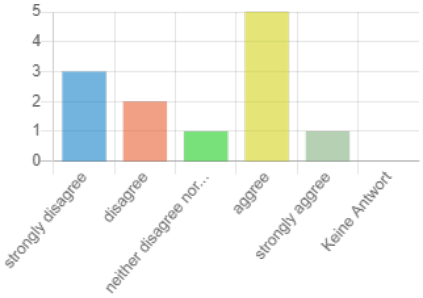
Missing Metadata	**Affected Participants**
Information on the tools and interconnection	2
Mapping between datasets in these tools	4
Quality metadata in data catalog	6
More detailed information on content of the dataset	1

Summarization of the textual responses.

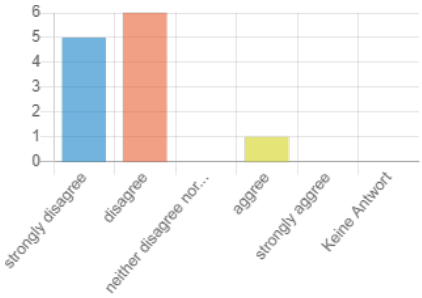
Table A.1: Question: Did you require additional metadata for selecting a dataset? If yes, what metadata was missing?



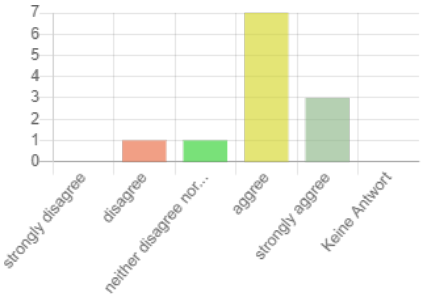
It was clear which steps you had to follow in the process of finding and identifying a relevant dataset.



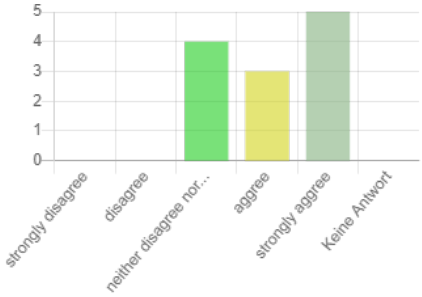
All metadata you were looking for was provided through the data catalog or data quality tool.



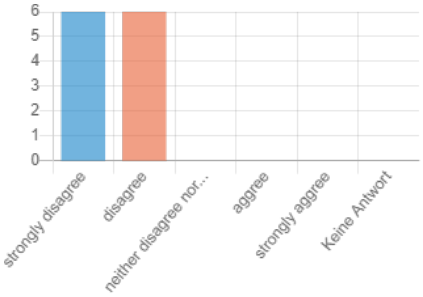
The process for requesting data was intuitive.



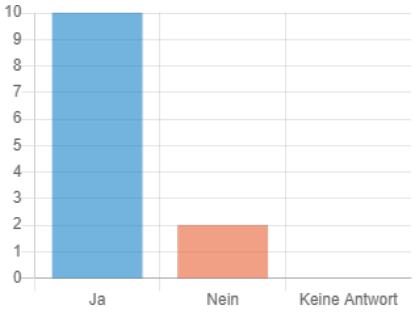
The process for requesting data was cumbersome (umständlich).



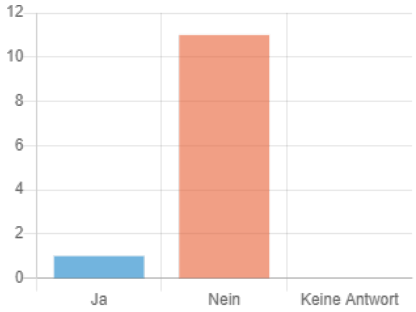
The process for requesting data was laborious (aufwändig).



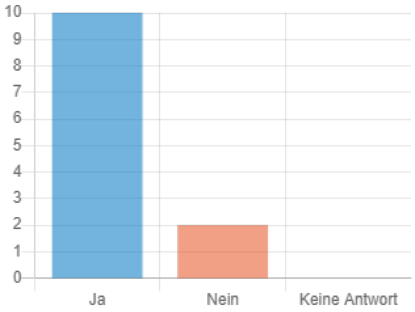
It was clear which steps had to be followed to request data.



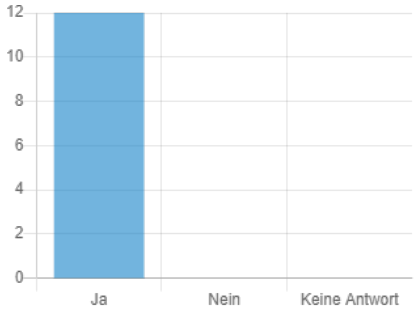
Did you consider looking for more metadata in other tools besides the data catalog?



Did you consider looking for more metadata in other tools besides the data quality tool (and data catalog)?



Would you have liked guidance in using the different tools?



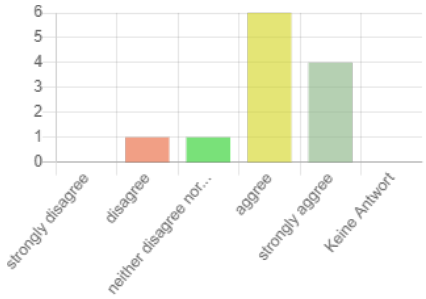
Did you ask for guidance in using the tools?

Difficulty	**Affected Participants**
Overall process was cumbersome due to many manual steps	2
Process for gaining access was unclear/unintuitive	4
Finding & gaining access to required tools was cumbersome	4
Lacking integration between tools (Relation between datasets in the tools was unclear)	5
Finding the required data quality information was difficult	4
Required guidance from colleague	1

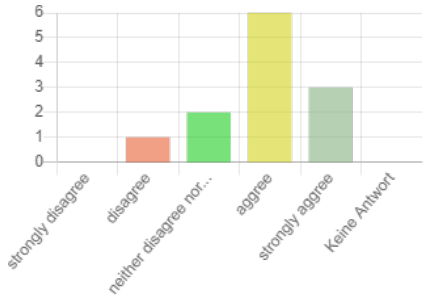
Summarization of the textual responses.

Table A.2: Question: Did you encounter difficulties? If yes, please describe these

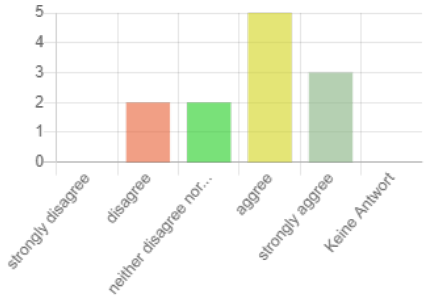
A.2 Scenario S2: With the Use of an EDMP



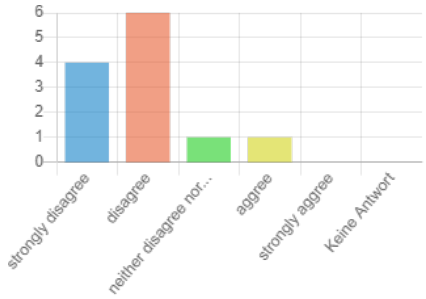
It was easy to find a dataset that matched the use case requirements.



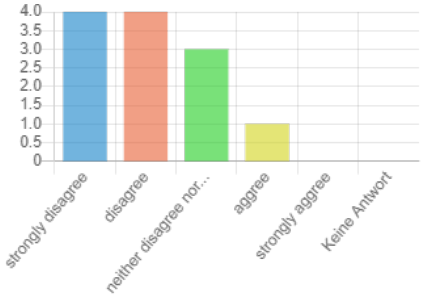
It was easy to find the metadata you were looking for in the Enterprise Data Marketplace.



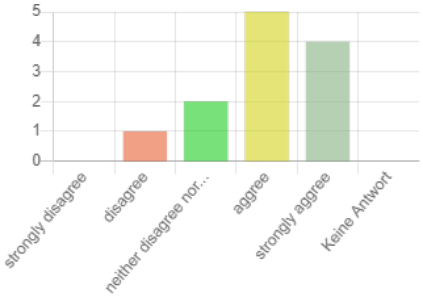
The process for finding and identifying a relevant dataset was intuitive.



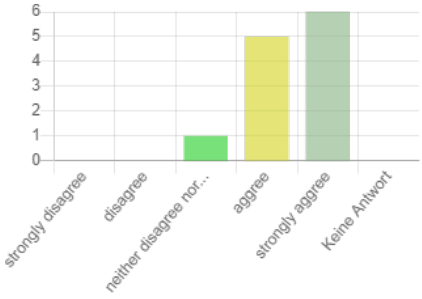
The process for finding and identifying a relevant dataset was cumbersome (umständlich).



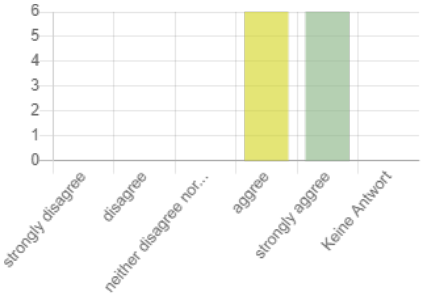
The process for finding and identifying a relevant dataset was laborious (aufwändig).



It was clear which steps you had to follow in the process to finding and identifying a relevant dataset.



All metadata you were looking for was provided through the Enterprise Data Marketplace.

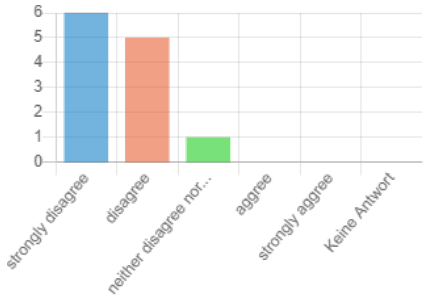


The process for requesting data was intuitive.

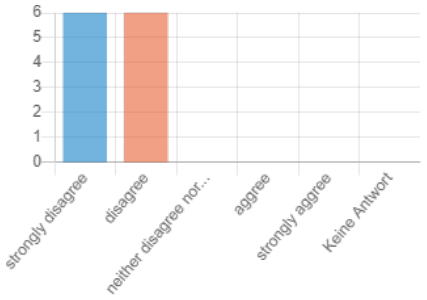
Missing Metadata	**Affected Participants**
The collection date	1

Summarization of the textual responses.

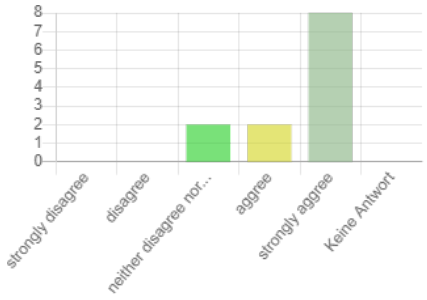
Table A.3: Question: Did you require additional metadata for selecting a dataset? If yes, what metadata was missing?



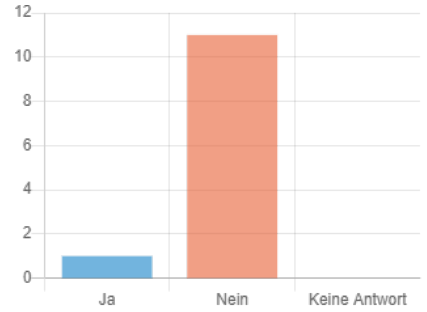
The process for requesting data was cumbersome (umständlich).



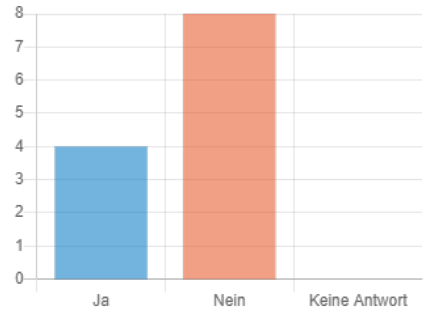
The process for requesting data was laborious (aufwändig).



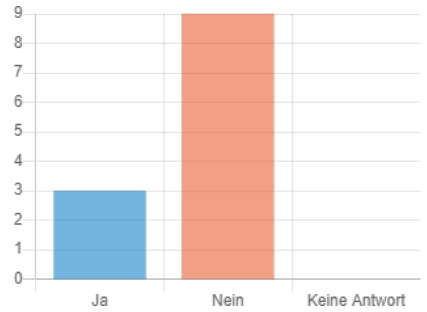
It was clear which steps had to be followed to request data.



Did you consider looking for more metadata in other tools besides the marketplace?



Would you have liked guidance in using the different tools?



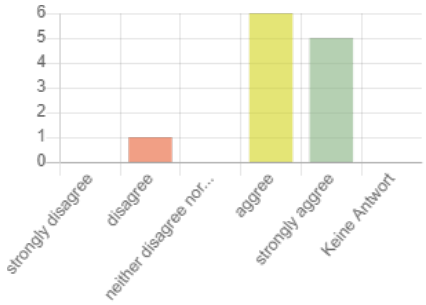
Did you ask for guidance in using the tools?

Difficulty	**Affected Participants**
Comparing two datasets was inconvenient	1
Desired more insight into dataset to check if all data required for task is contained	1
Difficulty defining search query and settings to filter for correct datasets	2

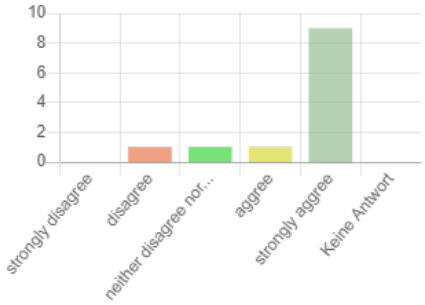
Summarization of the textual responses.

Table A.4: Question: Did you encounter difficulties? If yes, please describe these

A.3 Questionnaire on Scenario Comparison



The enterprise data marketplace simplified the process for finding and understanding data.



The enterprise data marketplace simplified the process for requesting data.