Enabling Trusted Data Sharing in Data Spaces: PROTON—A Privacy-by-Design Approach to Data Products

Laura Schuiki[1][a], Christoph Stach[1][b], Corinna Giebler[2][c], Eva Hoos[2][d] and Bernhard Mitschang[1][e]

[1] *Institute for Parallel and Distributed Systems, University of Stuttgart, Stuttgart, Germany*

[2] *Robert Bosch GmbH, Stuttgart, Germany*

Keywords: Distributed Data Management, Data Product, Privacy.

Abstract: In the current era of data-driven innovation, the value of data can be significantly enhanced by facilitating its dissemination. In this context, the data mesh concept has gained popularity in recent years. Data Mesh includes domain experts who design so-called data products. It is imperative that all parties involved have trust in these data products. This applies in particular to data subjects who share their data, data owners who create the data products, and data consumers who use them. To establish such trust, privacy approaches are key. Due to the decentralized and distributed nature of data mesh, however, traditional privacy strategies cannot be applied. To address this issue, we present **PROTON**, a concept that facilitates the handling of **PR**ivacy-c**O**mpliant da**T**a pr**O**ducts by desig**N**. PROTON is based on three pillars: a comprehensive **description model** for privacy requirements, an extended **creation process** that adheres to these requirements when compiling data products, and a refined **access process** for verifying compliance prior to data sharing. The practical applicability of PROTON is illustrated by means of a real-world application scenario that has been devised in collaboration with domain experts from our industry partner.

1 INTRODUCTION

In the age of digitalization, data has become a highly valuable asset, frequently compared to the new oil that fuels innovation and decision-making processes (Stach, 2023). The true potential of data, however, lies not in its collection but in its strategic dissemination across organizational boundaries, thereby creating new avenues for value creation (Reiberg et al., 2022). In this context, the data mesh concept has gained popularity in recent years. The data mesh is a new organizational approach to exchanging analytical data in the form of so-called *data products* (Dehghani, 2019). A data product is not a mere aggregation of raw data; rather, it is data that has been curated, refined, and enriched with properties such as discoverability, interoperability, and value (Dehghani, 2022). As such, it reflects a product-oriented mindset.

As data sharing becomes increasingly integral to collaborative ecosystems, it is crucial to preserve the integrity and value of data while ensuring trust in its exchange. In this context, trust is distributed across multiple stakeholders associated with data products: *Data subjects* whose data is included in a data product must be reassured that their privacy is protected. *Data owners* are responsible for ensuring that the data products they create comply with privacy regulations, including legal frameworks such as the General Data Protection Regulation (GDPR), as well as privacy requirements specific to a particular domain or data subject. Finally, *data consumers* must trust that the data products they access meet all relevant privacy standards.

To meet these requirements, a holistic privacy-by-design approach to the creation, management, and usage of data products is key. To this end, we have collaborated with our industry partner to make three contributions: **1.** We introduce a **description model** for data products that captures privacy requirements of data subjects, domain-specific privacy requirements, and legal privacy regulations. **2.** We extend the data product **creation process**, enabling data owners to enforce privacy requirements as required. **3.** We refine the data product **access process**, incorporating a verifi-

[a] https://orcid.org/0009-0008-0219-5485
[b] https://orcid.org/0000-0003-3795-7909
[c] https://orcid.org/0000-0002-5726-0685
[d] https://orcid.org/0000-0003-0040-9562
[e] https://orcid.org/0000-0003-0809-9159

Schuiki, L., Stach, C., Giebler, C., Hoos, E. and Mitschang, B.
Enabling Trusted Data Sharing in Data Spaces: PROTON - A Privacy-by-Design Approach to Data Products.
DOI: 10.5220/0013372900003899

In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025) - Volume 1*, pages 95-106
ISBN: 978-989-758-735-1; ISSN: 2184-4356

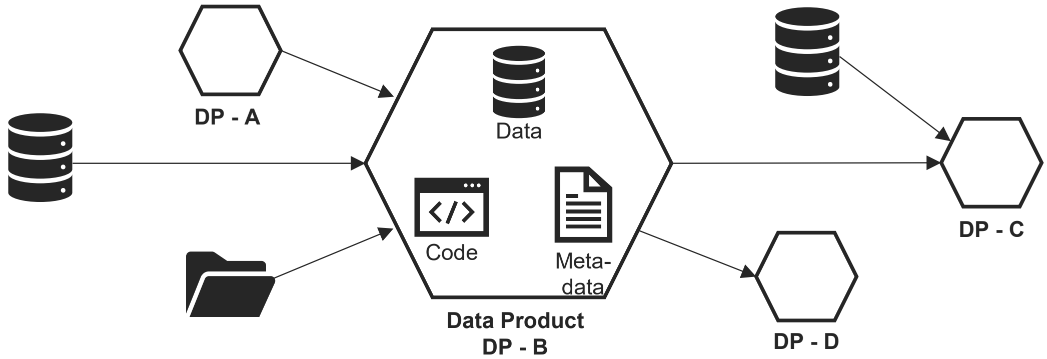

Figure 1: Structure of a Data Product and Its Role within the Ecosystem of a Data Mesh.

cation mechanism to ensure compliance with privacy standards before access is granted to data consumers.

These components are the foundation for **PROTON**, our **PR**ivacy-c**O**mpliant da**T**a pr**O**ducts by desig**N** concept. We assess PROTON by means of a real-world application scenario in the manufacturing domain, leveraging insights from our industry partner.

The remainder of this paper is structured as follows: Section 2 provides a definition of data products and an overview of the current state of research. Section 3 introduces an application scenario for data products from our industry partner, emphasizing key privacy considerations. Related work is discussed in Section 4. Section 5 presents PROTON. Together with our industry partner we assess PROTON in Section 6 before Section 7 concludes this paper.

2 DATA PRODUCTS

Given the current lack of well-defined standards for data products (Hasan and Legner, 2023), we initially present the key characteristics of data products as identified in literature (see Section 2.1). Subsequently, we examine the current state of research on data products, focusing in particular on the issue of data privacy (see Section 2.2).

2.1 A Harmonized View on Data Products

In the traditional approach to data management, data is stored and only processed when it is required for a specific purpose. However, nowadays data is regarded as a raw material that can be refined to create added value (Blohm et al., 2024). This view on data means that data is treated like a product and consumers are regarded as customers (Dehghani, 2022). This shift in thinking necessitates the transformation of raw data into self-contained and ready-to-use products.

To achieve this, further components are required in addition to raw data. Figure 1 depicts how Dehghani (2022) envisions the design of such a data product and its embedding in a data mesh. This design exceeds the paradigm of data as a product (Huang et al., 2015).

Data. The first component is the raw data itself, which is gathered about data subjects. This data may undergo changes over time, e.g., as more data is provided by a data subject. Data products are designed to dynamically reflect these changes by retrieving data directly from its sources (González-Velázquez et al., 2024).

Code. The second component is responsible for ensuring that the raw data is usable. This encompasses all stages from data retrieval to data transformation and data access. To this end, the data owner has to provides the necessary code and/or software (Jeffar and Plebani, 2024).

Metadata. The third component is the metadata, e.g., a description of the raw data and quality guarantees. That is, it encompasses all the information that data consumers need for identifying an appropriate data product and handling it properly (Driessen et al., 2023a).

The data owner is responsible for assembling these components, maintaining their quality, and providing access to them (Falconi and Plebani, 2023).

As depicted in Figure 1, data products are not limited to the usage of raw data; they can also leverage other data products. To illustrate, data product DP-B merges data product DP-A with raw data from external data sources as well as internal domain-specific data. DP-B can then be utilized as input for further data products (DP-C and DP-D), establishing a chain of interconnected data products that collectively constitute a data mesh as a result.

It is important to mention that any alteration to the source data inevitably results in a ripple effect, affecting all data products that are either directly or indirectly derived from it.

2.2 Privacy in the Realm of Data Products

Having established a harmonized understanding of data products, it is important to determine how to handle them in a trustworthy manner. In doing so, the three main stakeholders must be considered: the data subjects whose data is included in data products, the data owners who process that data to create data products, and the data consumers who use the data products. It is evident that privacy is key for trustworthy data products (Houser and Bagby, 2023). Privacy in this context goes far beyond mere data protection laws, as it involves individual privacy requirements that must be met (Quach et al., 2022). To build trust in data meshes, a data processing strategy must be implemented that automatically enforces compliance with these requirements (Podlesny et al., 2022). This can only be guaranteed by a holistic privacy-by-design approach covering all data processing steps involved (Borovits et al., 2024).

To gain insight into the current state of research on privacy in the context of data products, we conducted a literature review, differentiating between case studies and theoretical research.

Case Studies. Studies such as Chee and Sawade (2021), Joshi et al. (2021), and Lei et al. (2022), provide insight into the practical implementation of data products in a data mesh environment. Yet, none of these studies mention privacy considerations. A review of grey literature (Goedegebuure et al., 2024) as well as interviews with industry experts (Bode et al., 2024) attest to the importance of privacy in this context. However, neither source provides any specific approaches for its practical implementation.

Theoretical Research. In a 2022 publication, Dehghani (2022) introduced the concept of the data mesh and provided an overview of the design principles for data products, emphasizing the significance of data privacy while offering limited insights into the practical aspects of its implementation. Machado et al. (2021) propose a data mesh architecture that includes a security component. However, they only briefly mention the integration of privacy, without specifying at what level or in what manner this should be done. Driessen et al. (2023b) present a metamodel for data products where data contracts ensure certain technical and legal standards are met. Meanwhile, Podlesny et al. (2022) survey the privacy challenges within a data mesh, arguing against the use of centralized components for privacy management, concluding that existing privacy research therefore is not easily transferable to data mesh environments.

In conclusion, while existing research underscores the significance of privacy in data products, it lacks practical implementation solutions. However, a privacy solution for data products is a prerequisite for the establishment of trusted data sharing in data meshes. In the following section, we therefore introduce a real-world application scenario from our industry partner to identify the requirements towards such a privacy solution.

3 APPLICATION SCENARIO INVOLVING DATA PRODUCTS

This section presents an application scenario inspired by a globally active manufacturer of hybrid car components. This scenario serves to illustrate both the practical utility of data products and the imperative for addressing critical privacy considerations.

In our scenario, drivers interact with an application called CarApp during their trips. CarApp captures a variety of data including location, velocity, battery charge, fuel levels, and driving patterns. Users can also provide feedback by commenting on their routes, rating aspects such as parking convenience or perceived driving enjoyment, and receiving recommendations for future routes. Additionally, CarApp assesses the sustainability of the user's driving behavior. All collected data is managed by the car domain.

This data encompasses a range of sensitive information directly associated with a specific driver, including, but not limited to, location, velocity, driving patterns, and personal commentary. The potential for this data to be used to infer behaviors such as speeding or unsafe driving raises significant privacy concerns.

In this scenario, three key roles are involved:

Data Subject. The individual or entity whose data is being collected. In this scenario, this refers to the CarApp users, i.e., the drivers.

Data Owner. The individual or entity responsible for processing the collected data to create a data product. This entails managing the actual data and resources, and ensuring compliance with privacy regulations. In this scenario, there is a distinct data owner for each defined data product.

Data Consumer. The individual or entity that leverages the data product for analytical purposes or as an input for another data product. In this scenario, the data product DP-Car is utilized by two distinct data consumers.

Our application scenario encompasses three use cases, each of which is illustrated in Figure 2:

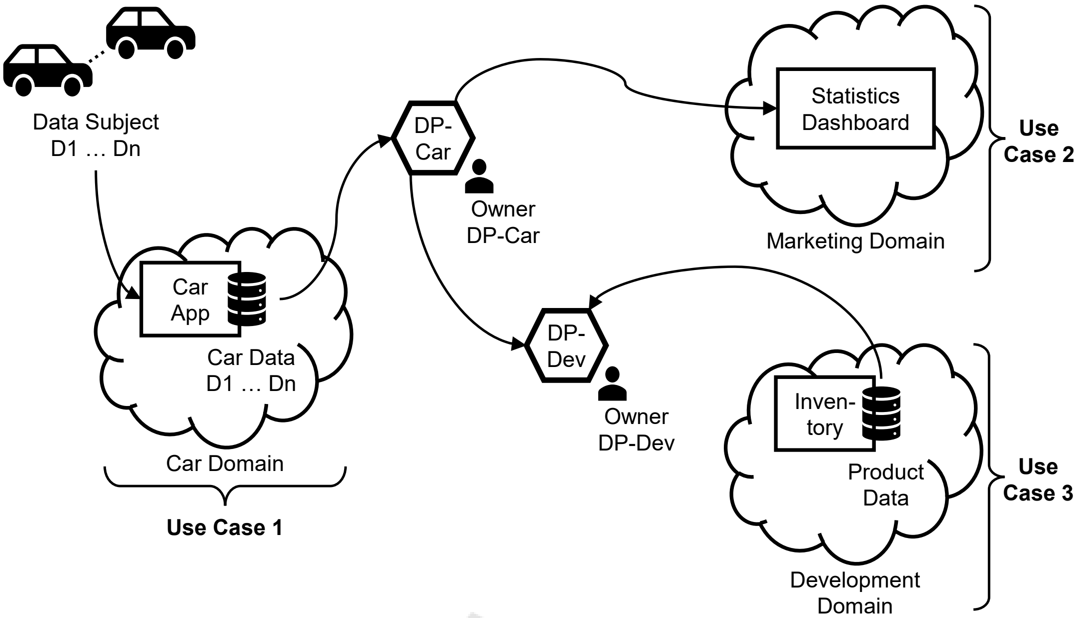

Figure 2: The Application Scenario Is Comprised of Three Use Cases, Each Associated with a Different Domain.

Use Case 1. This **intra-domain use case** pertains to the operation of CarApp within the context of the hybrid car domain. In light of the fact that the data remains within the same domain, it is imperative that both legal privacy regulations, such as the purpose limitation set forth in GDPR Article 5(b), and data subject privacy requirements are adhered to. For instance, a data subject might opt out of data collection entirely or consent to data collection on the condition that it is only used for CarApp operations or is anonymized before being used outside the car domain.

Use Case 2. In this **cross-domain use case**, the marketing domain leverages CarApp data indirectly via the data product DP-Car. The marketing team computes statistics on the distance traveled with recuperated energy, utilizing these findings for advertising purposes. Given that this data transcends domain boundaries, supplementary domain-specific privacy regulations come into effect. To illustrate, the car domain may necessitate the anonymisation of data prior to its proliferation to other domains.

Use Case 3. In this **enhanced data product use case**, the development domain enriches the CarApp data with supplementary information regarding automotive components, e.g., to gain insights that may facilitate improvements in battery performance. As with the cross-domain case, legal and domain-specific privacy regulations as well as privacy requirements of data subjects have to be observed. If the development domain decides to create a new data product (DP-Dev) using this enhanced data, it is their responsibility to ensure that the privacy regulations initially imposed on DP-Car are still met. Prior to the deployment of DP-Dev, the data owner of DP-Car therefore has to verify these regulations.

In these use cases, the car data is made available as a data product (DP-Car) for utilization by other domains. The process of creating a data product is comprised of three steps:

1. The **conceptualization step** requires the responsible data owner to determine the appropriate data, transformation code, and metadata for the data product. From a privacy perspective, it is important to identify the regulations that apply to the data and to ascertain how these regulations can be complied with during the processing of the data.
2. In the **construction step**, the data, code, and metadata are assembled based on the design, and the requisite resources are allocated. From a privacy perspective, it is therefore necessary in this step to adjust and supplement the transformation code so that data processing complies with privacy requirements.
3. In the **deployment step**, the data product is deployed on the provided infrastructure and registered in the data product catalog. From this point onward, the data owner is responsible for maintaining the data product. From a privacy perspective, mechanisms must be implemented to verify that access does not violate any privacy requirements. Furthermore, data products must be revised if privacy requirements change.

Table 1 presents a synthesis of our key insights regarding the roles and responsibilities associated with the trustworthy, i.e., privacy-aware handling of data products. It is evident that support is required in three areas: the collection and description of privacy requirements, their implementation and application, and the verification of their fulfillment.

4 RELATED WORK

As our literature review in Section 2.2 indicates, there is currently no dedicated approach to privacy for data products. We thus assess the feasibility of applying traditional approaches to data products in the three areas where assistance is required (see Table 1).

Elicitation. The establishment of privacy policies is of paramount importance whenever personal data is involved. They enact regulations that stipulate the manner in which data may be utilized. Miyazaki et al. (2009) introduce a computer-aided technique for eliciting privacy policies, emphasizing the significance of user involvement, particularly that of data subjects, in the process. Effective privacy protection is contingent upon data subjects who are adequately informed as to when and how their data is being utilized, as well as their right to formulate individual privacy requirements. Murmann et al. (2019) investigated the efficacy of notifications as a method of informing data subjects about privacy policy settings. Elicitation occurs prior to the conceptualization step during the process of data collection. Consequently, the distributed nature of a data mesh has no impact on this process. These approaches can thus also be applied to data products.

Description. In order for data owners to ensure that privacy policies are upheld when data is shared, it is essential that these policies are linked to the data in question (Stach et al., 2020). Pearson and Casassa-Mont (2011) propose that this can be achieved through the use of extended metadata. Eichler et al. (2021) address the challenges associated with the storage and utilization of such metadata, while Alshugran and Dichter (2014) investigate the development of descriptive metadata with input from domain experts. He and Antón (2003) focus on defining roles and permissions to regulate data access, which must be included in such a metadata model. These models are, however, designed for centralized data environments. As data products are deployed in distributed environments, such as data meshes, it is necessary to apply distinct policies depending on the data origin. In our application scenario (see Section 3), e.g., there are legal regulations that apply to all data, domain policies that apply to data products administered by a particular domain, and data subject privacy requirements that only apply to data products that contain data about this data subject. Therefore, dedicated metadata models are required for data products that are capable of reflecting such complex privacy requirements.

Verification. The establishment of trust is of the utmost importance when data consumers access data products, as they require assurance that the data in question adheres to the relevant privacy policies. Verification and enforcement of these policies are indispensable for the maintenance of trust. Ahmadian et al. (2018) introduce a model-based privacy analysis approach for data spaces. This approach is geared towards identifying potential privacy violations proactively, so that they can be prevented. This can be achieved by means of privacy enforcing technologies (Ahmadian et al., 2019). As the focus is on generic data sets, this approach requires adaptations when applied to data products. McSherry (2009) presents an approach for the automatic enforcement of general privacy policies. This approach, however, is not designed to cope with individual and dynamically

Table 1: Need for Assistance for the Three Roles in Handling Data Products to Improve Privacy.

Role	Required Type of Assistance
Data Subject	It must be facilitated to communicate individual privacy requirements in the **elicitation** process.
Data Owner	Applicable privacy requirements must be available in a machine-processable **description** to enable guidance on privacy-compliant data processing.
Data Consumer	A **verification** of privacy-compliance is required before access to exclude the risk of unauthorized data usage by design.

Table 2: Applicability of Traditional Privacy Approaches.

Area	Applicability to Data Products
Elicitation	Existing approaches can be applied to data products.
Description	**Description models** must be adapted to the more complex privacy requirements of data products.
Verification	Adapted processes are needed for the **creation** of and **access** to data products to ensure compliance with privacy requirements.

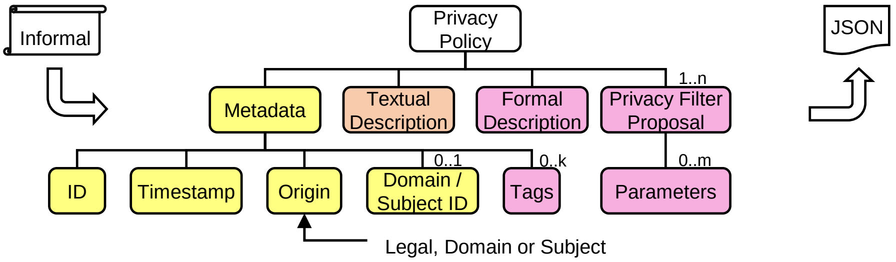

Figure 3: Blueprint for a PROTON Privacy Policy.

changing privacy requirements, which are common in the context of data products. Stach et al. (2022) present a framework that facilitates the definition and application of data processing steps that could be used for the enforcement of privacy policies. Nevertheless, it is not feasible to ascertain whether these steps are indeed sufficient to fulfill the specified privacy requirements. To be applicable to data products, existing solutions require significant adaptation to address the distributed nature and high complexity of privacy requirements that are ubiquitous in data mesh.

Table 2 summarizes our findings by identifying research gaps, particularly with regard to the description and verification of privacy policies. PROTON is designed to address these gaps.

5 PROTON: ENABLING PRIVACY-COMPLIANT DATA PRODUCTS BY DESIGN

To achieve a privacy-by-design approach to data products, we introduce PROTON, which enables the creation of privacy-compliant data products from the outset. PROTON integrates privacy considerations into the entire data product lifecycle. In the following, the fundamental elements of PROTON are delineated, including a description model for privacy policies (see Section 5.1) as well as revised processes for data product creation (see Section 5.2) and access requests (see Section 5.3).

5.1 Description Model for Privacy Policies

It is of the utmost importance to have comprehensive privacy policies in place to guarantee that data products not only comply with applicable privacy regulations but also meet the privacy requirements of the data subjects. PROTON introduces a description model that extends the metadata of data products to include detailed privacy policies. This model enables data subjects to delineate their privacy requirements while facilitating the identification and verification of relevant policies for data owners.

Privacy policies are classified into three categories in PROTON: Legal, domain, and data subject policies. Legal policies are applicable across the entire organization, whereas domain policies are specific to a particular domain. Data subject policies, meanwhile, are tied to the individual whose data is being processed. These policies can apply to an entire data product or specific parts of it. To illustrate, if a data product includes location and velocity data from multiple users, and one user has specified that their data should not be shared with other domains, only the processing of that user's data is restricted, while all other data remain unaffected.

The description model incorporates these policies as metadata linked to the relevant data. The model offers a structured approach to identifying, enforcing, and verifying applicable privacy policies, thereby supporting data owners and data consumers.

Figure 3 depicts a UML representation of the blueprint for our privacy policy. The following color coding is applied to describe the manner in which the respective components are determined: Elements shown in yellow are set automatically when a privacy policy is generated. Elements shown in orange are obtained directly from the respective issuer. For instance, privacy requirements of a data subject are collected by means of a questionnaire or an interview. Elements shown in pink are elicited by a privacy expert.

Each policy is assigned a unique 'ID', a 'Timestamp', and its 'Origin', i.e., whether it is a legal, domain, or data subject policy. In order to identify the creator of the privacy policy, the ID of the data subject or domain that created the privacy policy is specified in the 'Domain / Subject ID' element. In the case of legal privacy policies, this element is not required, as such policies are managed by a central governance team. Optional 'Tags' can be added to provide further detail regarding the privacy policy.

The model incorporates both a 'Textual Description' of the privacy requirements (e.g., the applicable legal text or declarations by a data subject) and a for-

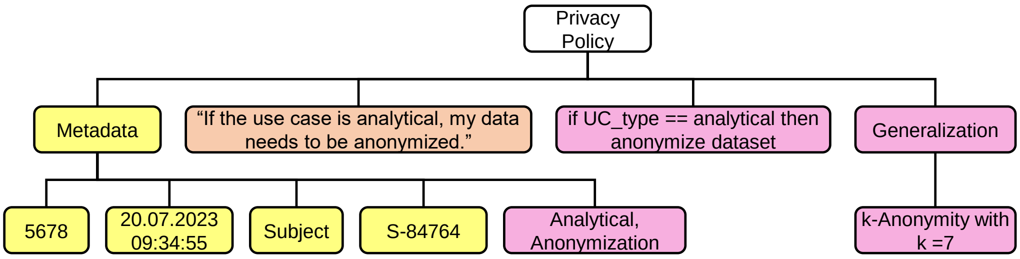
Privacy Policy
Metadata
"If the use case is analytical, my data needs to be anonymized."
if UC_type == analytical then anonymize dataset
Generalization
5678
20.07.2023 09:34:55
Subject
S-84764
Analytical, Anonymization
k-Anonymity with k =7

Figure 4: Sample Instance of the Model for a Data Subject Privacy Policy.

malized version, facilitating automated processing of the policy at subsequent stages. This 'Formal Description' is the result of a multi-stage formalization process adapted from Stach and Steimle (2019). This process facilitates the conversion of an informal description of the privacy requirements of a data subject into a machine-processable format, in our case, JSON.

Furthermore, suitable privacy measures can be recommended as part of the model as a means of ensuring that the relevant requirements are met. This 'Privacy Filter Proposal' must be applied to all associated data products at the point of creation and prior to access. Additionally, relevant 'Parameters' can be specified for the filter in question, which have to be used when applying the filter.

Each privacy requirement is defined as a distinct instance of the model. The sum of all instances thus represents the set of rules to be observed when handling data products. To ensure that these privacy policies are available to all data owners and data consumers so that they are able to take them into account when creating and accessing data products, it is essential that these policies are managed centrally, e.g., by the federated governance of a data mesh. The identifiers ('ID' and 'Domain / Subject ID') can be used to trace the data to which the policies apply. This even allows to identify the applicable policies for data products made from other data products by tracing the underlying source data via the data lineage. Additionally, the scope of a policy can be identified via the 'Origin' specifications, which can be global (for 'Legal'), domain-specific (for 'Domain'), or data-specific (for 'Subject').

Figure 4 shows a privacy requirement that was created on July 20, 2023, at 09:34:55, pertaining to the data subject with the ID 'S-84764'. This policy has the internal ID '5678'. The policy stipulates that the data may be utilized for analytical purposes, provided that it has been anonymized beforehand. The two tags, 'Analytical' and 'Anonymization', were derived from the textual description. The textual description was then converted into a formal description: 'if UC_type == analytical then anonymize dataset'. A privacy expert has recommended that the privacy filter 'Generalization' with the parameterization 'k-Anonymity with k=7' be applied as a suitable measure for fulfilling this privacy requirement.

5.2 Adaptations to the Creation Process of a Data Product

In order for the PROTON privacy policies to be applied in a systematic manner, it is necessary to extend the process of creating data products. Traditionally, privacy is often regarded as an afterthought. In contrast, the PROTON approach is based on a privacy-by-design philosophy, meaning that privacy measures are seamlessly integrated into the process of creating data products.

Figure 5 illustrates the process of creating data products. The process steps that originate from the traditional creation process are depicted in blue, while the PROTON-specific process steps are depicted in purple.

The creation of a new data product (DP_{new}) necessitates the definition of its three essential components: data, code, and metadata **(1)**. Internal source data as well as DP_{exist} are used as sources for the data component. DP_{exist} is an already existing data product from another domain. It is not necessary to make any adaptations to this process step, as the privacy requirements are already captured in the metadata at the point of data collection. The description model in PROTON ensures that no privacy requirements are lost.

Once the required components have been identified and acquired, DP_{new} is built **(2)**. Subsequently, novel PROTON-specific checks are mandatory. Initially, it is imperative to ascertain whether any privacy policies are applicable to the newly created data product **(3)**. If this is not the case, DP_{new} is deployed, and responsibility is transferred to its data owner **(6)**.

In the event that privacy requirements must be met, it is necessary to verify whether the data compilation in DP_{new} violates an applicable privacy policy **(4)**. If not, DP_{new} can also be deployed. Whereas, should any privacy requirement not be met, the appropriate privacy filters must be applied **(5)**. This may entail the removal of specific data features (Majeed and Lee,

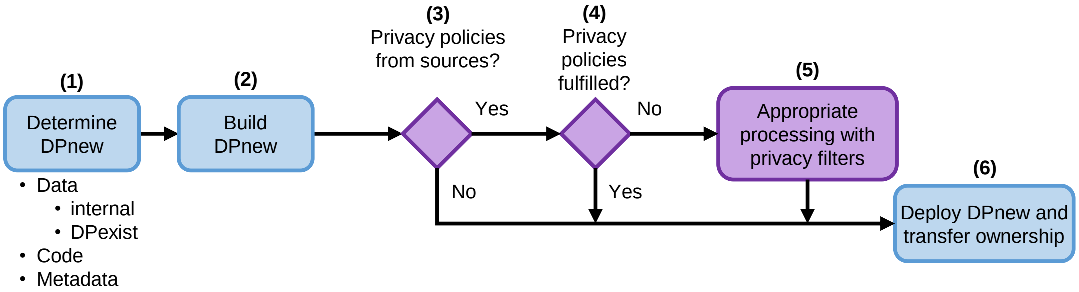
(1)
Determine DPnew
Data
internal
DPexist
Code
Metadata
(2)
Build DPnew
(3)
Privacy policies from sources?
Yes
No
(4)
Privacy policies fulfilled?
No
Yes
(5)
Appropriate processing with privacy filters
(6)
Deploy DPnew and transfer ownership

Figure 5: Extended Process of Creating Data Products in Compliance with PROTON Privacy Policies.

2021), or the addition of noise to the data (Deshkar et al., 2023). If there are multiple conflicting privacy policies, the most restrictive one has to be applied.

5.3 Adaptations to the Access Process to a Data Product

The extended process of creating data products in PROTON ensures that all created data products are privacy-compliant at the time of creation. However, it can happen that other, more restrictive requirements may apply to certain data consumers or that privacy requirements may subsequently become more restrictive. In order to be able to reflect these dynamics, it is necessary to expand the process of accessing data products in PROTON as well.

Figure 6 illustrates the process of accessing data products. The process steps that originate from the traditional access process are depicted in green, while the PROTON-specific process steps are once again depicted in purple.

When a data consumer requests access to a data product **(a)**, an initial check is made as to whether the requesting party is permitted to access such a data product **(b)**. If this is not the case, data access is denied immediately **(c)**.

If access is generally permitted, the PROTON-specific checks are initiated. To this end, in analogy to the creation process, it is first checked whether privacy requirements apply to the data product **(d)** and then whether these are fulfilled for the data consumer in question **(e)**. If there are no privacy requirements to be observed or if these requirements are already satisfied, access is granted **(g)**. In all other cases, the privacy filters specified in the privacy policy are applied prior to granting access **(f)**.

In this way, PROTON ensures that data consumers only receive data products that comply with all relevant privacy requirements, regardless of whether these regulations have been modified since the data product in question was initially created.

Synopsis. The PROTON approach effectively integrates privacy into the creation and management of data products, adhering to the principles of privacy by design. Our description model captures legal, domain, and data subject privacy requirements, thereby assisting data owners in identifying and enforcing applicable privacy policies throughout the entire lifecycle of a data product. This approach ensures that, on the one hand, data subjects have trust that their sensitive data is processed in accordance with their requirements. On the other hand, data consumers can have trust in the privacy compliance of any data product they access.

6 ASSESSMENT OF PROTON

In this section, we assess the practical applicability of PROTON based on the application scenario presented in Section 3. To this end, we have implemented and simulated the creation and management of data products with synthetic sample data using a proof-of-concept prototype of PROTON. The results of this feasibility and effectiveness assessment were reviewed

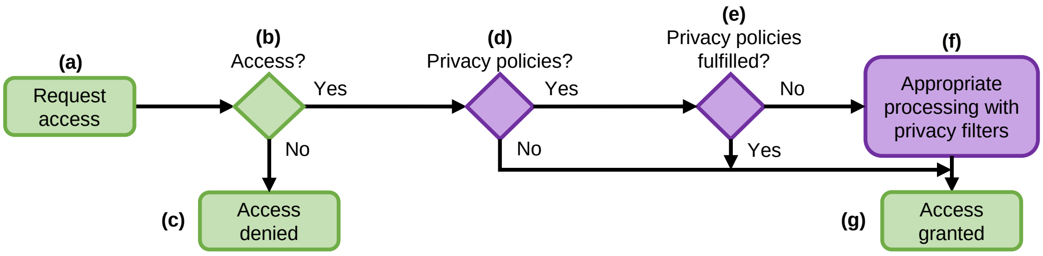

Figure 6: Extended Process of Accessing Data Products in Compliance with PROTON Privacy Policies.

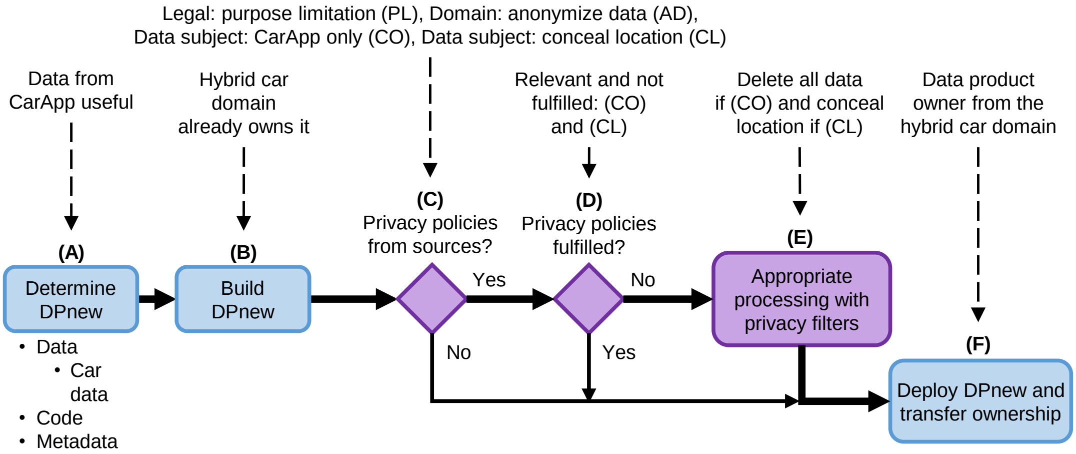
Legal: purpose limitation (PL), Domain: anonymize data (AD),
Data subject: CarApp only (CO), Data subject: conceal location (CL)
Data from CarApp useful
Hybrid car domain already owns it
Relevant and not fulfilled: (CO) and (CL)
Delete all data if (CO) and conceal location if (CL)
Data product owner from the hybrid car domain
(A)
Determine DPnew
• Data
• Car data
• Code
• Metadata
(B)
Build DPnew
(C)
Privacy policies from sources?
Yes
No
(D)
Privacy policies fulfilled?
No
Yes
(E)
Appropriate processing with privacy filters
(F)
Deploy DPnew and transfer ownership

Figure 7: DP-Car Creation Process.

with domain experts from our industry partner. Initially, Section 6.1 presents a detailed, step-by-step walkthrough of the processes of creating and accessing data products with PROTON. Subsequently, in Section 6.2, the assessment concludes with a summary of the lessons learned.

6.1 Review of the Practical Applicability

To assess the practical applicability of PROTON, we simulated Use Case 3 from Section 3 with the help of our industry partner. Domain experts assumed the roles of data owner and data consumer for the two data products involved, namely DP-Car and DP-Dev. Our assessment consists of three phases: creation of DP-Car, access request to DP-Car, and creation of DP-Dev. Prior to these phases, we generated synthetic data and defined privacy requirements. Based on the domain experts' feedback, we examined whether PROTON enables a privacy-compliant handling of data products.

Creation of DP-Car. Initially, a domain expert designed **(A)** and built **(B)** the new data product DP-Car. The use of PROTON did not entail any requisite alterations in this respect. The PROTON description model, which is available to all domains for all data in the data mesh, was used to ascertain the privacy requirements relevant to DP-Car **(C)**. This revealed that four distinct privacy policies must be applied to DP-Car. First, the purpose limitation stipulated in the GDPR is applicable (PL). Second, some data subjects have consented solely to the use of their data for the operation of CarApp (CO). Third, the additional domain-specific directive applies that all data must be anonymized if it leaves the domain (AD). Forth, one data subject has additionally indicated that their location data must be concealed (CL). Once all relevant policies have been identified, it is then necessary to verify whether DP-Car is in compliance with said policies **(D)**. As all data subjects have consented to the collection of data by the CarApp, the legal policy (PL) has been satisfied. In light of the assumption by the domain expert that DP-Car is utilized exclusively within the car domain, the domain-specific directive (AD) is also satisfied. Since DP-Car exceeds the operation of CarApp, all data of the data subjects who have not consented to this use must be removed (CO) and the location data of the respective data subjects must be concealed (CL). The formal description and proposal for privacy filters facilitate the identification of the necessary measures for domain experts in this phase **(E)**. Following this, DP-Car is deployed **(F)**. The sequence of operations is illustrated in Figure 7 by means of bold arrows.

Access Request to DP-Car. A domain expert from the development domain recognizes the benefits of DP-Car and intends to create a novel data product based on it. To this end, the domain expert submits an access request (α). In general, such an access request would be approved (β). However, PROTON requires the data owner to ascertain whether any existing privacy policies are in conflict with the intended use of the DP-Car (γ). As the requirements CO and CL have already been addressed during the development of DP-Car, only the purpose limitation (PL) and the domain policy requiring data to be anonymized before leaving the domain (AD) require verification. The domain expert's review (δ) indicates that the permitted purposes also cover the use by the development domain. Yet, the domain-specific directive that data has to be anonymized must be applied, as the data is now shared with another domain. The proposed privacy filters are therefore applied (ϵ), and access is then granted (ζ). PROTON supports the domain expert in two ways: first, by including reverification steps of privacy requirements in the extended access process, and second,

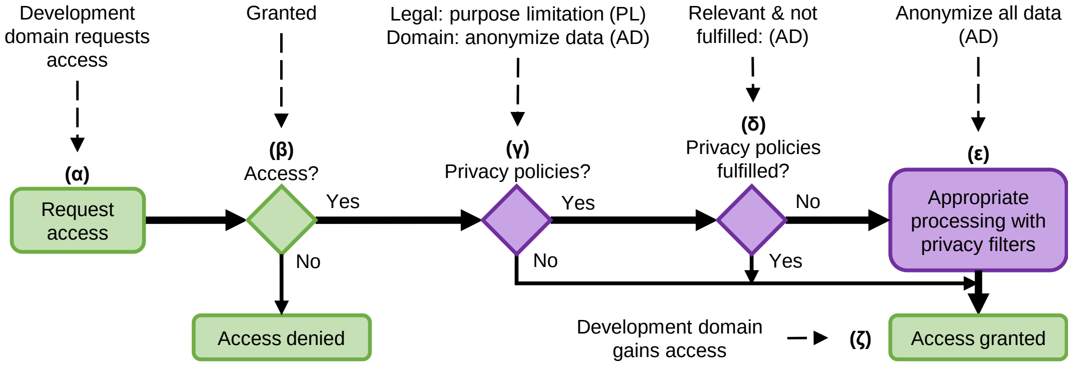
Development domain requests access
Granted
Legal: purpose limitation (PL)
Domain: anonymize data (AD)
Relevant & not fulfilled: (AD)
Anonymize all data (AD)
(α)
Request access
(β)
Access?
Yes
No
Access denied
(γ)
Privacy policies?
Yes
No
(δ)
Privacy policies fulfilled?
No
Yes
(ε)
Appropriate processing with privacy filters
Development domain gains access
(ζ)
Access granted

Figure 8: Process of Access Request to DP-Car.

by facilitating the identification of necessary measures due to the information in the PROTON description model. The sequence of operations is illustrated in Figure 8 by means of bold arrows.

Creation of DP-Dev. Once the domain expert has gained access to DP-Car, the creation of DP-Dev can begin. To this end, DP-Car is enhanced with internal data from the development domain. The required internal data is determined **(A)** and the new data product DP-Dev is built **(B)**. Subsequently, the PROTON-specific adaptations of the creation process take effect. As there are no policies for the internal data, it is only necessary to check which privacy policies are carried over from DP-Car **(Γ)**. As the domain policy AD was verified during the access request, it is only necessary to check whether the purpose limitation (PL) is violated by DP-Dev. As this is not the case, no further measures need to be taken **(Δ)** and DP-Dev can be deployed **(E)**. The sequence of operations is illustrated in Figure 9 by means of bold arrows.

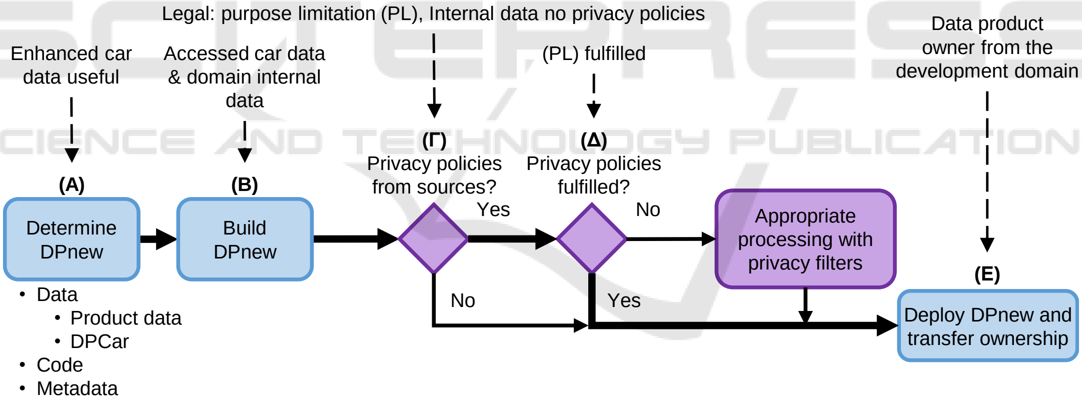
Legal: purpose limitation (PL), Internal data no privacy policies
Enhanced car data useful
Accessed car data & domain internal data
(PL) fulfilled
Data product owner from the development domain
(A)
Determine DPnew
(B)
Build DPnew
(Γ)
Privacy policies from sources?
Yes
No
(Δ)
Privacy policies fulfilled?
No
Yes
Appropriate processing with privacy filters
(E)
Deploy DPnew and transfer ownership
Data
Product data
DPCar
Code
Metadata

Figure 9: DP-Dev Creation Process.

Table 3: Summary of the Key Findings regarding the Benefits for Each Role Gained from PROTON.

Role	**Benefits of PROTON**
Data Subject	The **description model** provides a means of eliciting privacy policies. Data subjects are thus able to specify their privacy policies in a descriptive manner.
Data Owner	The description model assists data owners in identifying relevant privacy policies, while the adapted data product **creation process** ensures the enforcement of these policies.
Data Consumer	The revised **access process** for data products entails the verification of privacy policies. Consequently, data consumers are able to ascertain that the data product they access is in compliance with all relevant privacy policies.

6.2 Lessons Learned

The assessment conducted in collaboration with domain experts from our industry partner highlighted three aspects of PROTON:

1. **Data Subject Empowerment.** PROTON enables data subjects to delineate their privacy requirements at the time of data collection, which are then adhered to throughout the data product lifecycle. The description model allows data subjects to provide privacy policies in natural language, which are subsequently transformed for automated processing.
2. **Data Owner Responsibility.** It is the responsibility of data owners to ensure that their data products comply with all relevant privacy policies. To this end, the description model provided by PROTON offers a comprehensive overview of legal, domain, and data subject privacy policies, as well as proposed measures to satisfy them. This facilitates the identification and enforcement of these policies decisively.
3. **Data Consumer Trust.** It is also important to note that data consumers accessing data products require assurance that these products adhere to all applicable privacy policies. PROTON guarantees this by extending the access request process to include privacy policy verification and enforcement.

In summary, the results of our assessment, based on our industrial use cases presented in Section 3, demonstrate that PROTON effectively integrates privacy by design into existing data product processes. Our approach addresses the privacy requirements of data subjects, data owners, and data consumers. The key benefits for these three roles are summarized in Table 3.

It is important to note that the data mesh concept does not explicitly address the issue of privacy. It is regarded as one of several security goals, rather than a primary concern. Consequently, there is currently no established process for enforcing the rights and requirements of data subjects when dealing with data products. It is, therefore, the responsibility of each and every data owner to develop and implement strategies for establishing privacy. PROTON addresses this issue by introducing systematic processes and techniques that emphasize privacy. However, in the absence of a de facto standard for the privacy-aware handling of data products, it is not possible to evaluate our approach against a baseline.

As backed by our findings, the methodology presented in this paper is capable of adequately addressing the needs of data subjects, data owners, and data consumers. Therefore, it can be concluded that the adoption of PROTON as a reference point for the handling of data products is preferable to the status quo, which lacks comprehensive support and guidance on compliance with privacy regulations.

Currently, there are no agreed upon standards for the technical implementation of data mesh or data products, which would have been needed as a basis for an implementation of PROTON. Therefore, PROTON is an extension of the existing concepts and a guideline on how to handle privacy when implementing a data mesh or data products. We aim to address the definition of the missing standards together with our industry partner. Once such standards are established, PROTON can be implemented in this context, which in turn will allow for a more comprehensive technical evaluation.

7 CONCLUSION

In the context of the rapidly evolving landscape of data mesh, the importance of a trustworthy exchange of data cannot be overstated. Data products frequently constitute the backbone of this data sharing. Common procedures for handling data products inadequately address privacy considerations, thereby failing to establish trust.

To address this issue, we introduce PROTON, a novel privacy-by-design approach to data product management. In PROTON, privacy policies are linked to data products in the form of metadata. Thereby, privacy requirements of data subjects are always evident when data products are handled, thus enabling data owners to identify and apply all relevant policies in an effective manner. We enhanced existing data product creation and access processes to integrate privacy policy enforcement and verification, thereby ensuring trusted data sharing. Our assessment, based on an industrial application scenario, confirms the practicality of PROTON.

For researchers and practitioners engaged within the field of trust in data mesh, PROTON provides a scalable and adaptable solution that bridges the gap between privacy concerns and the necessity for seamless data sharing, thereby reinforcing the integrity and reliability of data-driven ecosystems.

REFERENCES

Ahmadian, A. S., Jürjens, J., and Strüber, D. (2018). Extending model-based privacy analysis for the industrial data space by exploiting privacy level agreements. In *Proceedings of SAC '18*.

Ahmadian, A. S., Strüber, D., and Jürjens, J. (2019). Privacy-enhanced system design modeling based on privacy features. In *Proceedings of SAC '19*.

Alshugran, T. and Dichter, J. (2014). Extracting and modeling the privacy requirements from HIPAA for healthcare applications. In *Proceedings of LISAT '14*.

Blohm, I., Wortmann, F., Legner, C., and Köbler, F. (2024). Data products, data mesh, and data fabric: New paradigm(s) for data and analytics? *Business & Information Systems Engineering*, pages 1–10.

Bode, J. et al. (2024). Towards Avoiding the Data Mess: Industry Insights from Data Mesh Implementations. arXiv:2302.01713v4 [cs.AI].

Borovits, N., Kumara, I., Tamburri, D. A., and van den Heuvel, W.-J. (2024). Privacy Engineering in the Data Mesh: Towards a Decentralized Data Privacy Governance Framework. In *Proceedings of ICSOC '23 Workshops*.

Chee, C. W. and Sawade, C. (2021). HelloFresh Journey to the Data Mesh. HelloFresh Engineering Blog, HelloTech.

Dehghani, Z. (2019). How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh. martinFowler.com.

Dehghani, Z. (2022). *Data Mesh: Delivering Data-Driven Value at Scale*. O'Reilly Media, Sebastopol, CA, USA.

Deshkar, P. A. et al. (2023). Studies on the Use of Various Noise Strategies for Perturbing Data in Privacy-Preserving Data Mining. *International Journal of Intelligent Systems and Applications in Engineering*, 12(8s):281–289.

Driessen, S., Monsieur, G., and van den Heuvel, W.-J. (2023a). Data Product Metadata Management: An Industrial Perspective. In *Proceedings of ICSOC '22 Workshops*.

Driessen, S., van den Heuvel, W.-J., and Monsieur, G. (2023b). ProMoTe: A Data Product Model Template for Data Meshes. In *Proceedings of ER '23*.

Eichler, R. et al. (2021). Enterprise-Wide Metadata Management: An Industry Case on the Current State and Challenges. In *Proceedings of BIS '21*.

Falconi, M. and Plebani, P. (2023). Adopting Data Mesh principles to Boost Data Sharing for Clinical Trials. In *Proceedings of ICDH '23*.

Goedegebuure, A. et al. (2024). Data Mesh: A Systematic Gray Literature Review. arXiv:2304.01062v2 [cs.SE].

González-Velázquez, R. et al. (2024). Smart Factory Hub – Towards a Data Mesh in Smart Manufacturing. *Procedia Computer Science*, 232:2709–2719.

Hasan, M. R. and Legner, C. (2023). Understanding Data Products: Motivations, Definition, and Categories. In *Proceedings of ECIS '23*.

He, Q. and Antón, A. (2003). A Framework for Modeling Privacy Requirements in Role Engineering. In *Proceedings of REFSQ '03*.

Houser, K. A. and Bagby, J. W. (2023). The Data Trust Solution to Data Sharing Problems. *Vanderbilt Journal of Entertainment and Technology Law*, 25(1):113–180.

Huang, G. et al. (2015). A Data as a Product Model for Future Consumption of Big Stream Data in Clouds. In *Proceedings of SCC '15*.

Jeffar, F. and Plebani, P. (2024). Federated Data Products: A Confluence of Data Mesh and Gaia-X for Data Sharing. In *Proceedings of ICSOC '23 Workshops*.

Joshi, D., Pratik, S., and Rao, M. P. (2021). Data Governance in Data Mesh Infrastructures: The Saxo Bank Case Study. In *Proceedings of ICEB '21*.

Lei, B. et al. (2022). Data Mesh — A Data Movement and Processing Platform @ Netflix. Netflix Technology Blog, Netflix Technology Blog.

Machado, I., Costa, C., and Santos, M. Y. (2021). Data-Driven Information Systems: The Data Mesh Paradigm Shift. In *Proceedings of ISD '21*.

Majeed, A. and Lee, S. (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, 9:8512–8545.

McSherry, F. D. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of SIGMOD '09*.

Miyazaki, S., Mead, N., and Zhan, J. (2009). Computer-Aided Privacy Requirements Elicitation Technique. In *Proceedings of APSCC '08*.

Murmann, P., Reinhardt, D., and Fischer-Hübner, S. (2019). To Be, or Not to Be Notified: Eliciting Privacy Notification Preferences for Online mHealth Services. In *Proceedings of IFIP SEC '19*.

Pearson, S. and Casassa-Mont, M. (2011). Sticky Policies: An Approach for Managing Privacy across Multiple Parties. *Computer*, 44(9):60–68.

Podlesny, N. J., Kayem, A. V. D. M., and Meinel, C. (2022). CoK: A Survey of Privacy Challenges in Relation to Data Meshes. In *Proceedings of DEXA '22*.

Quach, S. et al. (2022). Digital technologies: tensions in privacy and data. *Journal of the Academy of Marketing Science*, 50(6):1299–1323.

Reiberg, A., Niebel, C., and Kraemer, P. (2022). What Is a Data Space? Technical report, Gaia-X Hub Germany.

Stach, C. (2023). Data Is the New Oil–Sort of: A View on Why This Comparison Is Misleading and Its Implications for Modern Data Administration. *Future Internet*, 15(2):1–49.

Stach, C. et al. (2022). Demand-Driven Data Provisioning in Data Lakes: BARENTS—A Tailorable Data Preparation Zone. In *Proceedings of iiWAS '21*.

Stach, C., Gritti, C., and Mitschang, B. (2020). Bringing Privacy Control Back to Citizens: DISPEL — A Distributed Privacy Management Platform for the Internet of Things. In *Proceedings of SAC '20*.

Stach, C. and Steimle, F. (2019). Recommender-based privacy requirements elicitation – EPICUREAN: an approach to simplify privacy settings in IoT applications with respect to the GDPR. In *Proceedings of SAC '19*.