Speaking the same language or automated translation? Designing semantic interoperability tools for Data Spaces

Keywords: Data Spaces, Semantic Interoperability, Design Principles, Data Ecosystem

Abstract: This paper tackles the challenge of semantic interoperability in the ever-evolving data management and sharing landscape, crucial for integrating diverse data sources in cross-domain use cases. Our comprehensive approach, informed by an extensive literature review, focus-group discussions and expert insights from seven professionals, led to the formulation of six innovative design principles for interoperability tools in Data Spaces. These principles, derived from key meta-requirements identified through semi-structured interviews in a focus group, address the complexities of data heterogeneity and diversity. They offer a blend of automated, scalable, and resilient strategies, bridging theoretical and practical aspects to provide actionable guidelines for semantic interoperability in contemporary data ecosystems. This research marks a significant contribution to the domain, setting a new design approach for Data Space integration and management.

1 Introduction

In today's digital era, data is a critical asset driving innovation and economic growth. Recognizing this, the European Commission introduced the European Data Strategy (European Commission, 2020), aiming to create a single market for data within Europe. This strategy emphasizes the importance of inter-organizational data sharing to foster a competitive and innovative digital economy through seamless and secure data exchange. It supports the development of new products and services, enhances decision-making, and contributes to societal benefits such as improved healthcare and sustainable development (Hutterer et al., 2023; Guggenberger et al., 2024).

Data Spaces and Data Ecosystems are central to this strategy. Data Spaces are federated platforms designed to facilitate the sovereign and secure exchange of data between organizations, providing the necessary infrastructure for diverse data interactions. Data Ecosystems integrate multiple Data Spaces, creating a comprehensive environment that supports data-driven innovation across various domains and industries.

Our research addresses the critical challenge of achieving semantic interoperability within and across Data Spaces. We aim to develop tools that support specific ontologies and data structures within individual domains while facilitating their integration across different domains. This is essential to ensure that Data Spaces evolve into interconnected networks rather than isolated silos, supporting a wide array of applications and use cases (Otto, 2022). Semantic interoperability plays a crucial role in data integration, ensuring that different systems can correctly interpret and utilize exchanged data. Without semantic interoperability, other layers (technical, organizational, and legal) remain ineffective. Additionally, there is a significant research gap in semantic interoperability compared to other layers, which are relatively well-researched and standardized. By addressing this gap, our paper contributes new insights and solutions to this critical aspect of data interoperability. Before introducing the research questions, it is essential to clarify the significance of the three desirable attributes of semantic interoperability: automatable, scalable, and resilient.

Automatable semantic interoperability reduces manual intervention, minimizes errors, and increases efficiency by enabling seamless data integration and transformation across various systems. Scalability ensures that a system can handle increasing amounts of data and a growing number of participants without compromising performance. Scalable solutions support the expansion of data ecosystems, accommodating additional load and complexity, which is necessary for evolving and adapting to new requirements and larger datasets. Resilience means maintaining system functionality and performance despite variations in data quality, formats, and sources. A resilient system effectively manages challenges such as data heterogeneity, inconsistencies, and errors, ensuring robust and reliable data exchange even amid disruptions or changes in the data landscape.

The goal of our research is to develop tools that support the integration, management, and interconnection of data across various domains. The research question we aim to explore is:

> *RQ: How can tools be designed for automatable, scalable, and resilient semantic interoperability within and across Data Spaces?*

To investigate this, our approach involves a two-pronged strategy. First, we conduct a thorough literature review and engage in expert interviews to gather and analyze existing knowledge, establishing a set of meta-requirements (MR). Following this, we use these MRs as a foundation to derive design principles (DPs) for a tool that encapsulates the desired qualities of automation, scalability, and resilience, essential for fostering semantic interoperability between Data Spaces (Curry et al., 2022).

Addressing the challenge of automated interoperability from both practical and scientific standpoints, our approach aims to harmonize disparate data models and standards while significantly lowering barriers to data sharing and integration. The successful development and implementation of these principles promise to streamline data integration processes across various domains, paving the way for a unified and efficient data ecosystem. Such a transformation would unlock new potentials for innovation and value creation, revolutionizing the data landscape in Europe and setting a global benchmark for data interoperability and integration (Jabbar et al., 2017; Ouksel and Sheth, 1999).

The paper is structured as follows. In Section 2, we lay the foundational knowledge base, delineating the literature streams that form the groundwork for developing MRs. Section 3 details our research methodology. The MRs, derived from expert interviews, are systematically presented in Section 4. Building upon these MRs, Section 5 elaborates on the formulated DPs. The paper culminates in Section 6, where we discuss the broader implications of our findings, acknowledge the study's limitations, and highlight potential avenues for future research, concluding with a summative overview.

Main Contribution This paper significantly advances semantic interoperability in heterogeneous data ecosystems and data spaces. The main contributions are:

- **Conceptual Clarity:** Clear differentiation between data spaces and traditional database systems, enhancing the understanding of their unique roles within data ecosystems.
- **Meta-Requirements and Design Principles:** Identification of key meta-requirements for services promoting semantic interoperability, derived from a structured literature review, focus groups, and expert interviews. These meta-requirements form the basis for novel design principles ensuring automation, scalability, and resilience in data exchange processes.
- **Methodological Rigor:** Comprehensive methodological framework detailing each study stage, including the literature review, focus group discussions, and expert interviews, providing a robust basis for the study's conclusions.
- **Timely and Relevant Research:** Addressing contemporary issues within the European Data Strategy, aligning contributions with strategic objectives to foster a unified data market in Europe, with practical implications for policy and industry stakeholders.
- **Innovative Approach:** Dual focus on meta-requirements and design principles to tackle semantic interoperability challenges, providing actionable guidelines for developing tools supporting data integration and management across diverse domains.

In summary, the paper bridges critical gaps in the literature by offering a theoretically and empirically grounded framework for advancing semantic interoperability in data spaces, thus supporting the broader goal of creating interconnected and efficient data ecosystems.

2 Theoretical Background

This chapter delineates the theoretical underpinnings of dataspaces and semantic interoperability, providing the foundational knowledge crucial for the subsequent derivation of DPs.

2.1 Dataspaces

Originally conceptualized by Franklin and Halvey (Franklin et al., 2005; Halevy et al., 2006), the notion of dataspaces has evolved as a viable alternative to traditional relational databases. Subsequent definitions have expanded upon this initial concept, often emphasizing specific characteristics or adapting the concept for particular applications or domains. Table 1 presents a curated selection of these diverse definitions.

Presently, numerous dataspace approaches exist, each referencing different reference architectures and

Definition	Source
”Dataspaces are not a data integration approach; rather, they are more of a data co-existence approach. The goal of dataspace support is to provide base functionality over all data sources, regardless of how integrated they are.”	(Halevy et al., 2006)
“A dataspace system processes data, with various formats, accessible through many systems with different interfaces, such as relational, sequential, XML, RDF, etc. Unlike data integration over DBMS, a dataspace system does not have full control on its data, and gradually integrates data as necessary.”	(Wang et al., 2016)
“Dataspace is defined as a set of participants and a set of relationships among them.”	(Singh and Jain, 2011)

Table 1: Extract of definitions of datspaces - the complete overview is shown in (Curry, 2020b)

incorporating distinct core components. Notable examples include Gaia-X, Catena-X, IDS, FAIR dataspaces, and SOLID. These initiatives, while distinct, demonstrate efforts towards technical interoperability (European Commission, 2020). Yet, full technical compatibility among these initiatives remains to be achieved. An examination of various reference architectures (e.g., Gaia-X, Catena-X, IDS) and literature on dataspace components across sectors and domains (Curry, 2020a; Curry et al., 2022; Otto et al., 2022; Theissen-Lipp et al., 2023) reveals core components essential for controlled and secure data exchange:

1. Providing and Accessing Data (Connector):
   This component is tasked with managing data in accordance with defined usage policies, ensuring data sovereignty.
2. Intermediation Services (Metadata broker, the App Store, etc.):
   The Resource Catalog, a fundamental service, enumerates available offers, characteristics, and conditions of use.
3. Identity Management and Secure Data Exchange:
   This facet ensures participant identity verification and transaction security.
4. Management Components:
   These components are integral for daily operations, managing participant activities such as registration, deregistration, revocation, suspension, and monitoring.

Dataspaces confer various benefits to business (e.g., leveling the playing field for industrial data sharing, enhancing access to vast, heterogeneous data ecosystems), individuals (e.g., empowering control over personal data, expanding personal data monetization opportunities), science (e.g., increasing the socioeconomic impact of research data across domains and borders), and governance/public sector (e.g., establishing data commons for improved government services, facilitating evidence-based policymaking) (Curry et al., 2022).

Dataspaces are engineered to provide federated and self-determined interoperability for executing specific use cases (Otto, 2022). Typically driven by domain or use case specifics, examples like Catena-X for the automotive industry and Mobility Dataspace (MDS) for the mobility sector epitomize the need for integration across dataspaces. The European Commission envisages a singular European dataspace (Theissen-Lipp et al., 2023), and the concept of interoperable dataspace extends beyond enterprise boundaries, encapsulating distributed, federated, and decentralized data systems. When considering the interoperability challenge across dataspaces (dataspace mesh), it becomes more complex in terms of scalability, efficiency, and governance (Drees et al., 2021).

2.2 Semantic Interoperability

> ”Semantic interoperability ensures that these exchanges make sense—that the requester and the provider have a common understanding of the “meanings” of the requested services and data.” - (Heiler, 1995)

The term ’semantic interoperability’ has been recognized since the definition given above was published, emphasizing the importance of meaningful data and service exchange through a shared understanding. This core aspect remains relevant, with (Ouksel and Sheth, 1999) proposing the categorization of interoperability into system, syntax, structure and semantics levels.

In this classification, syntactic heterogeneity pertains to differences in machine-readable data representations, while structural interoperability concerns data modeling constructs. Schematic heterogeneity, especially prevalent in structured databases, is also a facet of structural heterogeneity. Despite advancements in systems, syntactic, and structural/schematic interoperability, comprehensive solutions for semantic interoperability (i.e., unified understanding of ”meaning”) are still elusive (Ouksel and Sheth, 1999).

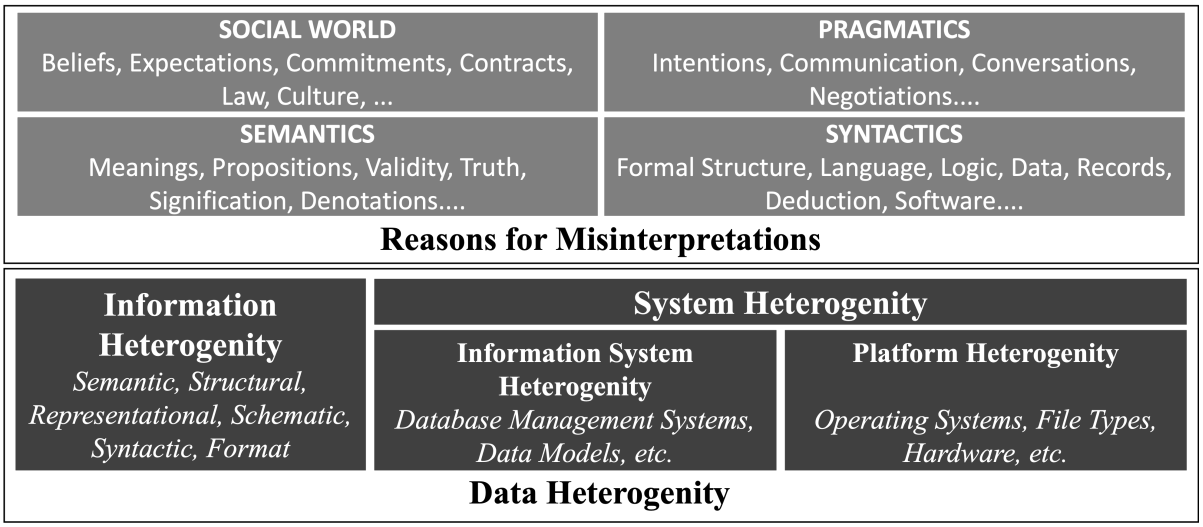

Figure 1: Heterogeneity in information systems and reasons for misinterpretations of data according to (Wenz et al., 2021)

A significant obstacle is the assumption in much web services technology of semantic homogeneity, implying a universal vocabulary and mutual understanding of data across systems (Uschold and Gruninger, 2004). Historical attempts to integrate systems under a single vocabulary have largely failed (Haslhofer and Klas, 2010). Recognizing and accommodating semantic heterogeneity is pivotal for achieving seamless system connectivity (Uschold and Gruninger, 2004). Figure 1 illustrates the challenges of heterogeneity in information systems, including resultant misinterpretations and the implications for semantic interoperability solutions. The "Data Heterogeneity" section characterizes the physical variances in data, often stemming from architectural decisions made during development. In contrast, the "Reasons for Misinterpretations" section highlights individual and subjective sources of error in data use and interpretation. In the milieu of dataspaces, characterized by large, distributed, autonomous, diverse, and dynamic information sources, accessing relevant and accurate information becomes increasingly complex (Ouksel and Sheth, 1999). The integration of semantic interoperability and semantics-based technologies in dataspaces is widely regarded as a fundamental component for their market success and establishment (Theissen-Lipp et al., 2023; Otto et al., 2022; Curry et al., 2022). The fusion of complex systems, from healthcare to finance, smart cities, and industrial automation, necessitates a unified framework for effective communication and understanding among these systems (Boukhers et al., 2023). Thus, the realization of a digitally connected world hinges on the deployment of robust, scalable, and resilient services that guarantee semantic interoperability.

While the necessity for such services is established (Boukhers et al., 2023), concrete implementation proposals or practical tests are yet to be realized. This gap in research underscores the motivation behind our proposal of DPs for such a service.

3 Methodology

In order to develop theoretically and empirically grounded set of DPs for interoperability tools for dataspaces, we follow a sound methodology to provide useful contributions. In the following, we will discuss the data collection, analysis, and foundations of DP generation. First, we conducted a structured literature review to gather the existing knowledge as preliminary design requirements. Second, building on that, we refined our understanding of the problem space and restructured the requirements. Finally, we developed an interview guideline for conducting semi-structured interviews to triangulate our preliminary findings with empiricial data. For this purpose thematic analysis of the focus group discussions is performed to identify key themes and topics. These identified themes are then used to structure the interview guide, ensuring that the questions are tailored to explore the relevant issues in greater depth.

3.1 Literature Review

The first step, that we perform to gather knowledge, is a structured literature review. Following well-established guidelines in information systems re-

search and computer science, we conducted the literature review in multiple steps (Webster and Watson, 2002; Zhang et al., 2011; Levy and Ellis, 2006; Vom Brocke et al., 2015).

We first identify databases covering relevant literature to the topic, i.e. information systems and computer science research. Thus, for our literature collection, we use IEEE Xplore, ACM Digital Library, ScienceDirect, Wiley InterScience and SCOPUS. In each database, we use the filtering functions to include only peer-reviewed english and german publications with full-text access. Furthermore, we exclude journals and conferences that are not related to our research topic "Interoperability in dataspaces". Second, we define the search strings following the procedure model of Schoormann et al. (Schoormann et al., 2018) iteratively. Finally, we performed the search with the parameters in each of the databases according to the specific string specifications, shown in Table 2:

Doing so, we identify 69 distinct publications. After filtering by title, abstract, and full text regarding the defined coarse and narrow focus, 31 publications remained. Following the guidelines of Webster and Watson (Webster and Watson, 2002), we scanned the references of the resulting publications from the main search. Again, these referenced publications are filtered by title, abstract and body according to the inclusion and exclusion criteria. After eliminating duplicates, we added three further studies resulting in a total of 31 publications.

3.2 Focus Groups and Expert Interviews

Informed by the key findings from the literature review, we structured existing knowledge into preliminary requirements, particularly for semantic interoperability tools. These requirements were evaluated and refined in the focus groups, which are formed by the core working group "Semantic Modeling and Interoperability" from a family of projects. The family of projects is funded by the German Federal Ministry for Economic Affairs and Climate Protection, with more than 80 partners from industry, research and the public sector. We used seven regular meetings (remote) in the focus groups for this purpose.

There are currently 16 members in the focus group from all three sectors, with expert knowledge in interoperability, data systems, application programming, semantics, operators, and end users. We developed a semi-structured interview-guideline based on the literature review and preliminary discussion within the focus group. Through semi-structured interviews, we only specify topics, and the experts could still say that it is not needed or so on. Using expert interviews as in-depth conversations for the elicitation of the requirements and DPs. After seven interviews with experts, whose profiles are shown in Table 3, theoretical saturation was reached and no further interviews were scheduled.

3.3 Design Principle Generation

DPs are prescriptive guidelines that codify design knowledge about a specific class of artifacts. DPs are meta-artifacts representing a general solution for this class within defined constraints (Chandra et al., 2015; Gregor, 2006; Baskerville et al., 2018). Within this constraints, DPs guide developers to increase the efficiency of design processes. Besides that, they are recognized as an excellent medium to communicate design knowledge with stakeholders (Chandra et al., 2016; Mcadams, 2003; Hevner et al., 2004). They are an important part of design science research (Sein et al., 2011; Möller et al., 2020), thus, we acknowledge DPs as the nucleus of a design theory as they cover three core components of a design theory: *causa finalis, materialis, formalis* (Jones and Gregor, 2007).

To develop the DPs, we follow the well-established guidelines of Möller et al. (Möller et al., 2020) and use the template of Chandra et al. (Chandra et al., 2015) for documentation. Our research approach is supportive and represents the ultimate starting point of a design science project of developing and implementing a tool for semantic interoperability in dataspaces. We use a literature review, focus group meetings, and expert interviews as the knowledge base to elicit MRs (Möller et al., 2020; Gregor and Hevner, 2013). Based on the literature review, we developed a preliminary list of potential requirements which we discussed in the focus groups. Furthermore, we discussed the definition of the problem spaces and the solution objective, which is defined in the motivation and research question. Resulting from that, we developed the questionnaire for the expert interviews. Finally, we evaluated the DPs argumentatively.

4 Formulating Meta-Requirements

The following section presents the MRs for services that make semantic interoperability in dataspaces automatable, scalable and resilient. These were derived as a result of the literature research and the expert interviews. Table 4 provides an overview of the MRs and the respective basis for their derivation from the expert interviews. In the following, we

	Search Strings
S1	`(semantic* AND (automated* OR resilient OR scalable OR shared OR sharing))`
S2	`(interoperability* OR inter-operability)`
S3	`(dataspace* OR data space OR datenraum)`
S	`S1 AND S2 AND S3`

Table 2: Search Strings

Expert	Occupation	Company / Industry
E1	Data Manager (PhD)	Ministry of Transport and Mobility Transition
E2	Research Associate Data Business	Institute for Software and Systems Engineering
E3	Senior Expert Cyber Physical Systems	Automotive Supplier ($>$ 200.000 employees)
E4	Lead Business Consultant (PhD)	Large consulting company ($>$ 10.000 employees)
E5	Head of Advisory Council	Dynamic Data Economy Foundation
E6	Research Associate Industry 4.0 Innovation	Large Software Company ($>$ 100.000 employees)
E7	Research Associate Data Science and AI	Institute for Applied Information Technology

Table 3: Expert Overview

describe and explain the five MRs for semantic interoperability, using the selected quotations to illustrate their meaning.

Meta-Requirement # 1: *Contextualization and metadata:* The effective use of an artifact for semantic interoperability depends on the appropriate provision of metadata and context. As *E1* and *E4* note, services that provide "metadata or extended metadata" and "semantic models" play a central role in facilitating the accurate interpretation of data. Boukhers et al. (Boukhers et al., 2023) reiterates this and also describes the importance for data consumers to understand the data and determine whether it meets their needs. *E6*'s comment on the need for data to be "semantically and syntactically correct", underpins the requirement for comprehensive metadata. *E2* emphasizes that the visibility of metadata enables a clearer understanding of data provenance and usage. This is important because, according to Curry et al. (Curry, 2020b), a dataspace must support the different data models and the different query languages of the participants with varying degrees of query expressivity. The importance of metadata is also emphasized by *E2* and *E7*, who note that understanding constraints and ensuring similar operational capabilities between parties is crucial for successful data exchange. Overall, contextualization and metadata are seen as the most important MR, as stated by 71,43% of the experts in the interviews. *E1, E2, E4 and E6* criticize the current approaches to describing existing data, where the descriptions are either incomplete (*E2, E4, E6*), incomprehensible (*E1, E2*), insufficient or unusable (*E1, E6*).

Meta-Requirement # 2: *Resilience of data:* The ability of an artifact to handle diverse data qualities and formats is pivotal for achieving semantic interoperability in dataspaces. *E1* highlights that, data quality can vary significantly, requiring a system that can "make the data comparable through automation". Ouksel et al. (Ouksel and Sheth, 1999; Ganzha et al., 2018) see the bridging of quality differences as a decisive factor in building global data ecosystems. Approaches from the fields of ontology matching and ontology alignment show possibilities for overcoming semantic heterogeneity (Otero-Cerdeira et al., 2015; Liu et al., 2021; Ardjani et al., 2015; Uschold and Gruninger, 2004). However, *E1* does not consider the current approaches in this area to be sufficient, as "important specifications are left out at the meta level", which is crucial for achieving automated interoperability. Ensuring the integrity and authenticity of data, as noted by *E5*, is fundamental to ensure that "the data sets you are using are correct". A resilient artifact is one that can inherently handle this diversity and complexity, acting as a robust backbone for dataspaces. Along with scalability, data resilience is regarded as the second most important MR, as noted by 57,12% of the experts.

Meta-Requirement # 3: *Scalability:* For an artifact to foster effective semantic interoperability, it must be accessible and adaptive to users regardless of their technical background. The remarks of *E1* and *E6* on the importance of semantics for interoperability and the need for an automated homogenizing approach highlight that an artifact needs to have the ability to seamlessly expand. *E7* comments on the ease of transforming data "from

MR	Meta-Requirement	Description	Experts	#Experts (%)
1	Contextualization and metadata	The artifact should require the provision of data context and mandatory metadata (specifics to be defined) for effective use	E1, E2, E4, E6, E7	5 (71,43%)
2	Resilience of data	The artifact must be resistant to different data qualities, data types and data formats in order to ensure practical usability	E1, E3, E5, E6	4 (57,12%)
3	Scalability	The artifact should be designed in such a way that it can be automated so that people without specialized knowledge can use it effectively to facilitate scalability in the complex semantic landscape	E1, E2, E5, E6	4 (57,12%)
4	Ease of use and simplicity	To encourage broad engagement, the artifact should be designed for extreme simplicity	E3, E4, E7	3 (42,85%)
5	Community-driven learning	The artifact should be able to continuously learn and improve by taking into account feedback from the community and users	E1, E7	2 (28,57%)

Table 4: Meta-Requirements Overview. In addition to the `Meta-Requirement` and `Description` columns, the `Experts` column lists which experts have named requirements that can be assigned to the respective meta-requirement. Meta-requirements have been ordered by importance, starting with the most important MR.

a label to a label or from a format to a format" reinforce the importance of scalability in dealing with the complexity of data structures. Theissen-Lipp et al. Theissen-Lipp et al. (Theissen-Lipp et al., 2023) describe dataspaces as "providing a scalable way for data exchange between their participants". Therefore, methods and concepts that enable efficient interoperability of heterogeneous systems are being investigated to understand how the interoperability problem should be addressed (Nilsson and Sandin, 2018). 57,12% of the experts named the scalability of the artifact as an important requirement in the context of dataspaces. In the interviews, scalability was also often mentioned together with "automation" and "resilience" (*E1, E2, E7*). *E1, E6, and E7* believe that scaling is only possible through automation.

Meta-Requirement # 4: *Ease of use and simplicity:* The success of an artifact in achieving widespread adoption hinges on its simplicity and user-friendliness. *E4* calls for "intuitive design" and artifacts that are "quite open-source", advocating the necessity for accessible design. In addition to intuitive design, the literature also describes that providing semantic interoperability should not require stakeholders to adapt to major changes in their systems, or the solution should be dependent on their system (Noura et al., 2019). *E3* and *E6* discuss the creation of an ecosystem through the linkage of technical services (*E3*) and the advantages of time-saving, reduced effort, and non-required expertise (*E6*). *E7* mention of "KI and LLMs" providing a natural language interface, which exemplifies the potential for intuitive user interaction, making complex systems more approachable for non-expert users. There are already some research approaches that try to combine new technology trends such as LLMs with knowledegraphs (Wang et al., 2023; Pan et al., 2023) and other semantic technologies (Baek et al., 2023; Trajanoska et al., 2023). Boukhers et al. (Boukhers et al., 2023) identify the following possibilities to create semantic interoperability in dataspaces through artificial intelligence algorithms: *Automatic Metadata Extraction, Ontology and Vocabulary Alignment, FAIRness Evaluation, Data Quality Assessment & Enhancement, Privacy Preserving, Compatibility Improvement*. 42,85% of the experts mentioned the importance of the simplicity of the artifact.

Meta-Requirement # 5: *Community-driven learning:* An artifact that leverages the collective intelligence of its user community has the potential to enhance and evolve over time. According to 28,57% of the experts, this capability is an important characteristic of the artifact. *E1, E5* explicitly mention the need of being able to update new data schemas, ontologies, or structures and to learn individual or domain-specific characteristics (*E6*). Dataspaces are also seen in the literature as a dynamic, constantly changing medium that must be able to cope with the volatility of the data landscape (Curry, 2020a; Drees et al., 2021; Franklin et al., 2005). *E1* suggests a service that "aggregates and analyzes data" to enhance interoperability aligns with this need. Similarly, *E7* emphasizes an artifact acting as an "adapter

or translator" to unite different ontologies embodies the ideal of community-driven evolution, where user feedback leads to continuous refinement and increased effectiveness. Ontology matching and alignment approaches such as those that have been researched for several years can form a starting point for dataspace-specific research (Otero-Cerdeira et al., 2015; Liu et al., 2021; Ardjani et al., 2015; Uschold and Gruninger, 2004).

Each of these MRs is interrelated, creating a cohesive framework for an artifact that enables semantic interoperability. By addressing these core needs, the proposed artifact can serve as a robust, inclusive, and adaptive framework for managing and utilizing data in a semantically interoperable manner. In chapter 5, DPs based on the MRs are derived.

5 Design Principles

In the burgeoning field of dataspaces, where interoperability is essential, the following DPs were formulated on the basis of the MRs to connect the various data sources, enhance data resilience, and promote an inclusive and adaptive environment for data exchange and processing. Figure 2 shows the fulfillment of the MRs above by the DPs (Möller et al., 2020). The seven resulting principles are discussed in the subsection. The format of Chandra et al. (Chandra et al., 2015) is used to present the DP. Subsequently, we describe a preliminary evaluation of them using the framework of Iivari et al. (Iivari et al., 2021).

5.1 Design Principle Description

DP1: Integration Optimization: *Design interoperability artifacts to optimize the seamless integration of diverse data sources, domains, and formats, with an emphasis on scalability and user-friendly automation, catering to users needing robust integration solutions across multiple data platforms.* x1

Rationale: This principle is vital for the establishment of interoperable dataspaces as it ensures a cohesive and seamless integration of heterogeneous data sources. Derived from MR1, MR2 and MR3, it advocates for the design of interoperability artifacts that are not only scalable but also user-friendly, automating the integration process to handle diverse data qualities and formats. The significance of this principle lies in its direct impact on the artifact's scalability, as highlighted by experts, who emphasizes the necessity for data to be "brought from A to B and potentially transformed" (E3) and the benefits of making "data comparable through automation" (E1). By optimizing integration, the artifact effectively supports the practical usability of diverse data, acknowledging the insights from both MRs that underline the need for resilience against data heterogeneity and the provision of rich contextual metadata.

DP2: Data Resilience Promotion: *Equip the system with interoperability artifacts that incorporate mechanisms for data robustness, allowing users to maintain reliable performance with data of varying quality levels, types, and formats. These mechanisms are necessary given the diverse nature of data sources and the requirement for consistent data integrity in varying operational environments.*

Rationale: Reflecting the concerns expressed through MR1, MR2, MR3, and MR5, this principle addresses the core need for an artifact's capability to handle data of varying quality. It encapsulates the notion that data resilience is foundational, ensuring that the interoperability artifacts maintain reliability across disparate data landscapes. This principle is embedded in the understanding that "if you structure the data correctly, then you can at least be sure that the data sets you are using are correct" (E7), which enforces the importance of robustness against different data types and formats. It resonates with the need for simplicity and scalability, ensuring that the artifact is resilient enough to adapt to the evolving data ecosystem, thereby fostering a robust semantic interoperability framework.

DP3: Metadata Enhancement: *Implement interoperability artifacts in the system that require rich metadata and contextual information, thereby enabling users to effectively use and understand data across various domains. This principle is particularly relevant in settings where data from multiple domains must be integrated and understood collectively.*

Rationale: Building upon MR1 and MR2, this principle fortifies the essence of contextualization by mandating rich metadata for data utility maximization. The requirement for "services that provide metadata or extended metadata" (E1) and the call for semantic models "so that they can be interpreted" (E4) are testimonies to the principle's alignment with the essential need for metadata enrichment. This DP thus underpins the effectiveness of interoperable artifacts by ensuring that metadata is not just present but also informative and indicative of the data's context, thus enhancing the semantic interoperability across different systems and domains.

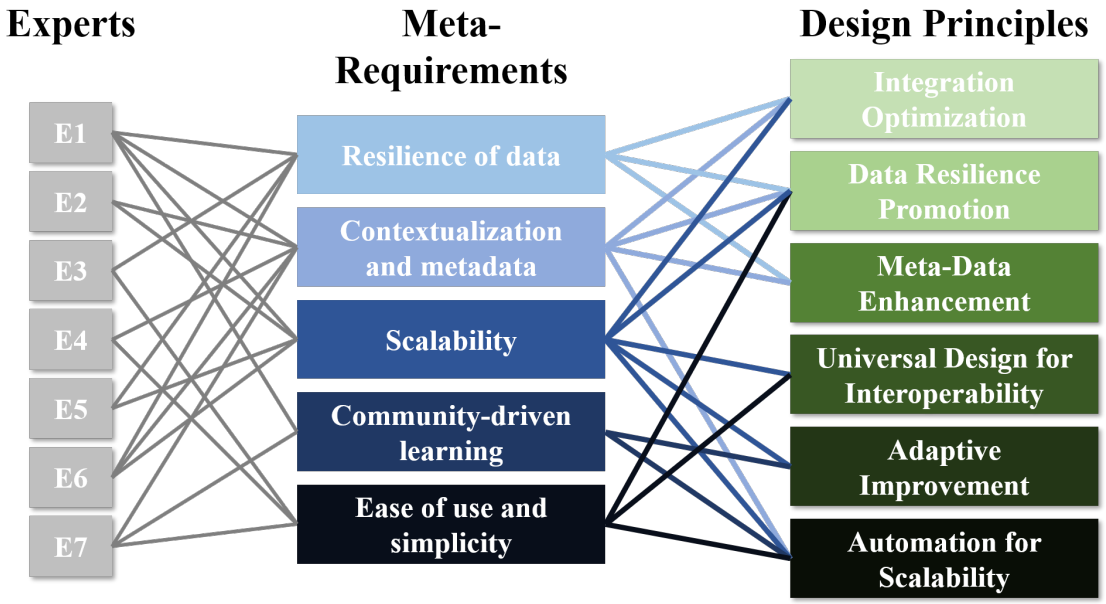

Figure 2: Overview of the dependencies of the experts, the meta-requirements and the design principles. The links between the experts and the meta-requirements show on which interviews the meta-requirements were formulated and between the meta-requirements and the design principles the basis for deriving the design principles from the meta-requirements.

DP4: Universal Design for Interoperability: *Construct interoperability artifacts with a universal design in the system, simplifying interactions to enable a broad range of users, regardless of their technical expertise, to engage with and use the system effectively. This approach is essential in environments where users with varying levels of technical knowledge need to interact with the system.*

Rationale: Corresponding with MR3 and MR5, this principle is instrumental in democratizing the use of interoperability artifacts. The principle leverages the notion that ease of use leads to broader engagement, which is crucial in an environment where "the interface, in which virtually anyone can participate without any expertise, is widely used" (E5). It ensures that the artifact is not just for those with technical acumen but also approachable for lay users, echoing the need for artifacts that are intuitive and simple, thereby reducing barriers to entry and facilitate a wider adoption of the interoperability standards.

DP5: Adaptive Improvement: *Develop interoperability artifacts in the system that support adaptive learning through community feedback, allowing users to contribute to and benefit from continuous improvements and the integration of new data formats. This principle is vital in settings where ongoing community engagement and evolution of data handling capabilities are required.*

Rationale: Aligned with MR3 and MR4, this DP embraces the dynamism of the data ecosystem. It emphasizes the need for interoperability artifacts to be adaptive, learning from community feedback, which is imperative as "a service that aggregates and analyzes data could help improve the situation" (E1). The concept of continuous learning and adaptation guarantees the evolution of artifacts by integrating new data formats and community insights, thereby underpinning the progressive refinement of semantic interoperability mechanisms.

DP6: Automation for Scalability: *Integrate a high degree of automation in interoperability artifacts within the system, enhancing resilience against varying data qualities and formats and establishing a scalable framework. This automation is crucial for users operating in environments with diverse data types and a need for scalable data management solutions.*

Rationale: Reflecting insights from MR2, MR3, MR4, and MR5, automation stands as a cornerstone for enhancing the resilience and scalability of interoperability artifacts. This principle encapsulates the concept that "an automated approach that can homogenize data" (E6), thus the artifact can maintain robustness amidst the fluctuating landscapes of data types and qualities. It underscores the significance of automated processes in managing complexity and fos-

tering scalability, ensuring that the artifact not only serves current needs but is also primed for future expansion and diversification of dataspaces.

Incorporating these principles into the development of interoperability artifacts offers a clear pathway towards the creation of resilient, scalable, and user-friendly dataspaces. Each principle, derived from empirical insights, functions synergistically to ensure that dataspaces can meet the demands of an increasingly interconnected and data-driven world.

5.2 Preliminary Evaluation

This section summarizes the analytical evaluation of the DPs we developed. We evaluate them as a set, because it is the unit of prescriptive knowledge (Iivari et al., 2021). DPs can help developers and operators of interoperability tools in dataspaces with their operations. To increase *accessibility*, we use the language of practitioners and domain experts, respectively. The framework presented by Chandra et al. (Chandra et al., 2015) helps by giving clear guidelines on materiality, action, and boundary conditions. The set of DPs is *important* to practitioners. Interoperability is an important topic within and across dataspaces. We focus on one pillar of the European Interoperability Framework (Commission, 2023). The DPs provide clear guidance on how to develop tools to cope with the challenges of semantic interoperability. The provision of a comprehensive set of DPs is of *novelty* both in research and in practice. While there is some research on semantic interoperability, no publication is yet addressing tools to enable it in dataspaces. As we provide actionable quotes from the experts, these suggestions can be directly implemented. Thus, the DPs are *actable*. Finally, the set of DPs provides *guidance* for developers of semantic interoperability tools. In summary, the argumentative evaluation suggests that the DPs are sufficiently defined and usable for their purpose.

6 Conclusion and Future Work

In the rapidly evolving landscape of data management and sharing, semantic interoperability is proving to be a crucial factor. The integration of different data sources, models and ontologies is a complex but important task.

Contributions. Our study makes a significant contribution to the field of semantic interoperability in dataspaces by developing six novel DPs. These principles integrate extensive conceptual and empirical knowledge and are specifically tailored to the requirements of automatic, scalable and resilient semantic interoperability. The DPs are new to the field of semantic interoperability for dataspaces and to the ability to translate complex theories into practical, actionable guidelines, thus providing significant added value for both academic research and practical application in dataspace management. The development of these principles is based on a careful analysis of 31 professional publications and expert interviews, underlining their relevance and applicability in current and future dataspace integration and management scenarios.

Limitations. While our research provides directional insights, some limitations need to be considered. Research on dataspaces is subject to continuous change, which means that our findings, although current at the moment, may require adjustments in the future. Furthermore, the design principles presented are yet to be practically evaluated in terms of their effectiveness in real-world application scenarios. The qualitative data of our study, obtained through a focus group and expert interviews, offer multiple perspectives but might be shaped by the context of the participants.

Future Work. To address these limitations and further develop our research, several avenues are open. An immediate step is the instantiation of the DPs into a working prototype, allowing for practical evaluation. We are in the process of establishing a conceptual framework for developing this prototype with a small developer group. Prior to development, an empirical evaluation with a broader expert group is planned to ensure the effectiveness and practical applicability of the DPs, particularly focusing on their level of abstraction and guidance for practitioners. Another critical area of exploration is the level of integration of interoperability tools within dataspaces and the extent of their specialization. Our long-term vision is to develop a universal tool akin to ”Translator for data models.” However, the efficiency and feasibility of such a universal tool versus more specialized tools require further investigation.

Conclusion. While our study makes significant strides in the field of semantic interoperability in dataspaces, it also opens up numerous research opportunities. The dynamic nature of dataspaces, the evolving requirements of interoperability tools, and the economic considerations of their implementation all point towards a rich and fertile ground for future research. The development of a practical prototype based on our DPs, followed by empirical evaluation and economic modeling, will be crucial steps in advancing the field and realizing the full potential of semantic interoperability tools in dataspaces.

REFERENCES

Ardjani, F., Bouchiha, D., and Malki, M. (2015). Ontology-Alignment Techniques: Survey and Analysis. *International Journal of Modern Education and Computer Science*, 7(11):67–78.

Baek, J., Aji, A. F., and Saffari, A. (2023). Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering.

Baskerville, R. L., Baiyere, A., Gregor, S. D., Hevner, A. R., and Rossi, M. (2018). Design science research contributions: Finding a balance between artifact and theory. *Journal of The Association for Information Systems*, 19:3.

Boukhers, Z., Lange, C., and Beyan, O. (2023). Enhancing Data Space Semantic Interoperability through Machine Learning: A Visionary Perspective. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1462–1467, Austin TX USA. ACM.

Chandra, Seidel, and Gregor (2015). Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions. In *2015 48th Hawaii International Conference on System Sciences*, pages 4039–4048, HI, USA. IEEE.

Chandra, Seidel, and Purao (2016). Making Use of Design Principles. In Parsons, J., Tuunanen, T., Venable, J., Donnellan, B., Helfert, M., and Kenneally, J., editors, *Tackling Society's Grand Challenges with Design Science*, volume 9661, pages 37–51. Springer International Publishing, Cham.

Commission, E. (2023). The European Interoperability Framework in detail | Joinup. https://bit.ly/3SDMRfP.

Curry, E. (2020a). *Dataspaces: Fundamentals, Principles, and Techniques*, pages 45–62. Springer International Publishing, Cham.

Curry, E. (2020b). *Real-Time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems*. Springer International Publishing, Cham.

Curry, E., Scerri, S., and Tuikka, T., editors (2022). *Data Spaces: Design, Deployment and Future Directions*. Springer International Publishing, Cham.

Drees, H., Pretzsch, S., Heinke, B., Wang, D., and Langdon, C. S. (2021). Data Space Mesh: Interoperability of Mobility Data Spaces.

European Commission (2020). European data strategy. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en.

Franklin, M., Halevy, A., and Maier, D. (2005). From databases to dataspaces: A new abstraction for information management. *ACM SIGMOD Record*, 34(4):27–33.

Ganzha, M., Paprzycki, M., Pawłowski, W., Szmeja, P., and Wasielewska, K. (2018). Towards Semantic Interoperability Between Internet of Things Platforms. In Gravina, R., Palau, C. E., Manso, M., Liotta, A., and Fortino, G., editors, *Integration, Interconnection, and Interoperability of IoT Systems*, pages 103–127. Springer International Publishing, Cham.

Gregor (2006). The Nature of Theory in Information Systems. *MIS Quarterly*, 30(3):611.

Gregor and Hevner (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, 37(2):337–355.

Guggenberger, T. M., Altendeitering, M., and Langdon Schlueter, C. (2024). Design principles for quality scoring-coping with information asymmetry of data products. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*.

Halevy, A., Franklin, M., and Maier, D. (2006). Principles of dataspace systems. In *Proceedings of the Twenty-Fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 1–9, Chicago IL USA. ACM.

Haslhofer, B. and Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, 42(2):1–37.

Heiler, S. (1995). Semantic interoperability.

Hevner, March, Park, and Ram (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1):75.

Hutterer, A., Krumay, B., and Mühlburger, M. (2023). What constitutes a dataspace? Conceptual clarity beyond technical aspects. In *AMCIS 2023 Proceedings*.

Iivari, J., Rotvit Perlt Hansen, M., and Haj-Bolouri, A. (2021). A proposal for minimum reusability evaluation of design principles. *European Journal of Information Systems*, 30(3):286–303.

Jabbar, S., Ullah, F., Khalid, S., Khan, M., and Han, K. (2017). Semantic Interoperability in Heterogeneous IoT Infrastructure for Healthcare. *Wireless Communications and Mobile Computing*, 2017:1–10.

Jones, D. and Gregor, S. (2007). The Anatomy of a Design Theory. *Journal of the Association for Information Systems*, 8(5):312–335.

Levy and Ellis (2006). A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Science: The International Journal of an Emerging Transdiscipline*, 9:181–212.

Liu, X., Tong, Q., Liu, X., and Qin, Z. (2021). Ontology Matching: State of the Art, Future Challenges, and Thinking Based on Utilized Information. *IEEE Access*, 9:91235–91243.

Mcadams, D. (2003). Identification and codification of principles for functional tolerance design. *Journal of Engineering Design*, 14(3):355–375.

Möller, F., Guggenberger, T. M., and Otto, B. (2020). Towards a Method for Design Principle Development in Information Systems. In Hofmann, S., Müller, O., and Rossi, M., editors, *Designing for Digital Transformation. Co-Creating Services with Citizens and Industry*, volume 12388, pages 208–220. Springer International Publishing, Cham.

Nilsson, J. and Sandin, F. (2018). Semantic Interoperability in Industry 4.0: Survey of Recent Developments and Outlook. In *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, pages 127–132, Porto. IEEE.

Noura, M., Atiquzzaman, M., and Gaedke, M. (2019). Interoperability in Internet of Things: Taxonomies and Open Challenges. *Mobile Networks and Applications*, 24(3):796–809.

Otero-Cerdeira, L., Rodríguez-Martínez, F. J., and Gómez-Rodríguez, A. (2015). Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949–971.

Otto, B. (2022). A federated infrastructure for European data spaces. *Communications of the ACM*, 65(4):44–45.

Otto, B., Ten Hompel, M., and Wrobel, S., editors (2022). *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*. Springer International Publishing, Cham.

Ouksel, A. M. and Sheth, A. (1999). Semantic interoperability in global information systems. *ACM SIGMOD Record*, 28(1):5–12.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap.

Schoormann, T., Behrens, D., Fellmann, M., and Knackstedt, R. (2018). Design Principles for Supporting Rigorous Search Strategies in Literature Reviews. In *2018 IEEE 20th Conference on Business Informatics (CBI)*, pages 99–108, Vienna. IEEE.

Sein, Henfridsson, Purao, Rossi, and Lindgren (2011). Action Design Research. *MIS Quarterly*, 35(1):37.

Singh, M. and Jain, S. K. (2011). A Survey on Dataspace. In Wyld, D. C., Wozniak, M., Chaki, N., Meghanathan, N., and Nagamalai, D., editors, *Advances in Network Security and Applications*, volume 196, pages 608–621. Springer Berlin Heidelberg, Berlin, Heidelberg.

Theissen-Lipp, J., Kocher, M., Lange, C., Decker, S., Paulus, A., Pomp, A., and Curry, E. (2023). Semantics in Dataspaces: Origin and Future Directions. In *Companion Proceedings of the ACM Web Conference 2023*, pages 1504–1507, Austin TX USA. ACM.

Trajanoska, M., Stojanov, R., and Trajanov, D. (2023). Enhancing Knowledge Graph Construction Using Large Language Models.

Uschold, M. and Gruninger, M. (2004). Ontologies and semantics for seamless connectivity. *ACM SIGMOD Record*, 33(4):58–64.

Vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., and Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems*, 37.

Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., and Liu, T. (2023). HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledge.

Wang, Y., Song, S., and Chen, L. (2016). A Survey on Accessing Dataspaces. *ACM SIGMOD Record*, 45(2):33–44.

Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii.

Wenz, V., Kesper, A., and Taentzer, G. (2021). Detecting Quality Problems in Data Models by Clustering Heterogeneous Data Values.

Zhang, H., Babar, M. A., and Tell, P. (2011). Identifying relevant studies in software engineering. *Information and Software Technology*, 53(6):625–637.