

Krista Mätäsniemi

# **SOLVING INTER-ORGANIZATIONAL DATA SHARING CHALLENGES WITH GAIA-X**

Master of Science Thesis  
Faculty of Information Technology and Communication Sciences  
Examiners: David Hästbacka  
Petri Kannisto  
September 2023

## ABSTRACT

Krista Mätäsniemi: Solving inter-organizational data sharing challenges with Gaia-X  
Master of Science Thesis  
Tampere University  
Master's Programme in Information Technology  
September 2023

---

While data collection has become more common, organizations and individuals have identified possibilities of value derived from collected data. To get maximum value from data, it must be shared with others. However, there are multiple unsolved challenges especially in data sharing between organizations in multi-actor environments. In this thesis, these challenges are identified based on systematic literature research. Furthermore, Gaia-X data infrastructure is considered as a solution to the challenges. The main purpose of the thesis is to study, how Gaia-X solves existing inter-organizational data sharing challenges in multi-actor environments.

Firstly, the results of the literature review show there are several types of data sharing challenges that can be divided into three main categories. The business-related challenges are the most critical when organizations consider joining a data sharing ecosystem and its benefits to their organization. However, most of the identified challenges are included in the data governance category. The last category contains interoperability challenges which play an important role in enabling efficient data sharing between organizations.

Secondly, the analysis of how Gaia-X data infrastructure addresses the challenges shows that it focuses on improving the availability of data and not so much on the actual data exchange process. Furthermore, the distributed nature of Gaia-X architecture leaves data owners in charge of many factors affecting the security of data sharing and the quality of the data. As the biggest surprise, the architecture lacked the mechanisms for the enforcement of usage policies and contracts attached to a shared data resource.

Keywords: data ecosystem, data infrastructure, data sharing, Gaia-X

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

# TIIVISTELMÄ

Krista Mätäsniemi: Organisaatioiden välisen tiedon jakamisen haasteiden selvittäminen Gaia-X:n avulla  
Diplomityö  
Tampereen yliopisto  
Tietotekniikan DI-ohjelma  
Syyskuu 2023

---

Samalla kun tiedon kerääminen on yleistynyt, organisaatiot ja yksilöt ovat tunnistaneet kerätystä tiedosta saatavan arvon mahdollisuudet. Jotta tiedosta saataisiin mahdollisimman suuri arvo, sitä on jaettava muiden kanssa. Tiedon jakamisessa on kuitenkin useita haasteita erityisesti organisaatioiden välillä monitoimija ympäristöissä. Tässä opinnäytetyössä kyseiset haasteet on tunnistettu systemaattisessa kirjallisuuskatsauksessa. Lisäksi Gaia-X tieto infrastruktuurin soveltumista haasteiden päihittämiseksi on mietitty. Työn päätarkoituksena on selvittää, kuinka Gaia-X ratkaisee olemassa olevia organisaatioiden välisen tiedon jakamisen haasteita monitoimija ympäristöissä.

Ensiksi, kirjallisuuskatsauksen tulokset osoittavat, että tiedon jakamisen haasteita on monia erilaisia ja ne voidaan jakaa kolmeen pääkategoriaan. Liiketoimintaan liittyvät haasteet ovat kriittisimpiä, kun organisaatiot pohtivat liittymistä tiedon jakamisen ekosysteemeihin ja niiden hyötyjä organisaatiolle. Suurin osa tunnistetuista haasteista liittyy kuitenkin tietohallinto kategoriaan. Viimeinen kategoria pitää sisällään yhteentoimivuuden haasteita, jotka ovat merkittävässä roolissa tehokkaan tiedon jakamisen mahdollistamisessa organisaatioiden välillä.

Toiseksi, analyysi Gaia-X tieto infrastruktuurin kyvykkyyksistä vastata haasteisiin osoittaa, että infrastruktuuri keskittyy enemmän parantamaan tiedon saatavuutta eikä niinkään varsinaista tiedonvaihto prosessia. Lisäksi Gaia-X arkkitehtuurin hajautettu luonne jättää tiedon omistajat vastaamaan monista tiedon jakamisen turvallisuuteen ja tiedon laatuun vaikuttavista tekijöistä. Suurimpana yllättävänä asiana arkkitehtuurista puuttuivat mekanismit jaettuun tietoresurssiin liitettyjen käyttöehtojen ja sopimusten toimeenpanemiseksi.

Avainsanat: tieto ekosysteemi, tieto infrastruktuuri, tiedon jakaminen, Gaia-X

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -ohjelmalla.

## **PREFACE**

The research work of this thesis is a part of a project called Future Electrified Mobile Machines (FEMMa). The project is a collaborative project between Tampere University and VTT that aims to find solutions for electrified mobile machines and enable new value from data.

I would like to thank my supervisors David Hästbacka and Petri Kannisto for this unique opportunity to learn and develop myself in this field. The working environment they have offered me has made this project possible. Their excellent guidance and support have been priceless during the project and helped me to achieve my goals.

I would also like to express my gratitude to my family, friends and supportive co-workers who have supported and encouraged me as the project progressed in facing a wide range of challenges and emotions. Each of them has helped me to successfully complete this project. I appreciate every comment and development suggestion I have received from them for my thesis.

Tampere, 27th September 2023

Krista Mätäsniemi

## CONTENTS

1.	Introduction . . . . .	1
2.	Research Approach . . . . .	3
3.	Background. . . . .	5
	3.1 Data Autonomy and Sovereignty. . . . .	5
	3.2 Data Ecosystem . . . . .	6
	3.3 Data Federation and Data Spaces . . . . .	7
	3.4 Data Governance . . . . .	8
	3.5 Data Sharing and Trust . . . . .	10
4.	Gaia-X Architecture . . . . .	11
5.	Data Sharing Challenges . . . . .	15
	5.1 Business . . . . .	15
	5.1.1 Cooperation . . . . .	15
	5.1.2 Competition . . . . .	16
	5.2 Data Governance . . . . .	17
	5.2.1 Data Quality . . . . .	17
	5.2.2 Data Ownership . . . . .	19
	5.2.3 Security. . . . .	20
	5.3 Interoperability. . . . .	22
	5.3.1 Syntactic Interoperability . . . . .	22
	5.3.2 Semantic Interoperability . . . . .	23
	5.4 Summary of Data Sharing Challenges . . . . .	24
6.	Solving Data Sharing Challenges with Gaia-X . . . . .	26
	6.1 Data Sharing in Gaia-X Compliant Data Ecosystems. . . . .	26
	6.2 Supply of Gaia-X Federation Services . . . . .	30
	6.3 Example Data Sharing Case Scenario . . . . .	34
7.	Discussion . . . . .	40
8.	Conclusions . . . . .	42
	References. . . . .	43

## SYMBOLS AND ABBREVIATIONS

API	application programming interface
B2B	business-to-business
B2C	business-to-consumer
DCAT-3	Data Catalog Vocabulary Version 3
DID	decentralized identity
EU	European Union
GDPR	General Data Protection Regulation
GXFS	Gaia-X Federation Services
IAM	identity and access management
IDM	identity management
IDS	International Data Spaces
IoT	internet of things
IRI	Internationalized Resource Identifier
JSON-LD	JavaScript Object Notation for Linked Data
ODRL	Open Digital Rights Language
RDF	resource description framework
REST	representational state transfers
SoS	system of systems
SSI	self-sovereign identity
UCON	usage control
VC	verifiable credential
VLAN	virtual local area network
W3C	World Wide Web Consortium
XACML	eXtensible Access Control Markup Language

## 1. INTRODUCTION

While data collection has become more common, the importance of data as a resource in value creation and innovation has increased. Data-driven value can be realized in data ecosystems where data is shared among its participants. Data sharing between the participants increases the availability of data and enables wiser decision-making which improves development of data-driven services and products as well as competitive innovations. However, there are some challenges to be overcome before the full potential value of data can be realized in such multi-actor environments.

There are several articles describing a single data sharing challenge, but they do not provide an overall picture of the challenges. Gelhaar and Otto [1] discovered that the fear and the lack of trust cause a lower level of openness than characteristics of a data ecosystem describes. Furthermore, authors of [2] examined the adoption factors of data platforms and identified the relevance of data governance. Nokkala and others [3] pointed out that varying interests of organizations affect the requirements of data governance in shared platforms. Moreover, interoperability is a challenge in multi-actor environments, and it is hard to get all actors to comply with common processes and standards [4]. To overcome the challenges, a well-designed data infrastructure is needed [5].

Gaia-X Association provides a federated and secure data infrastructure based on European values, which increases transparency, interoperability and sovereignty across data and services in any existing cloud or edge technology stack. The infrastructure contains a data sharing architecture which facilitates data sharing between different actors, such as organizations regardless their industry. The components of architecture are developed by utilizing existing standards and widely used technologies in order to standardize data sharing and facilitate the adoption of the infrastructure. Furthermore, Gaia-X provides a set of open-source reference implementations of software components called federation services that aim to improve interoperability across data ecosystems. [6]

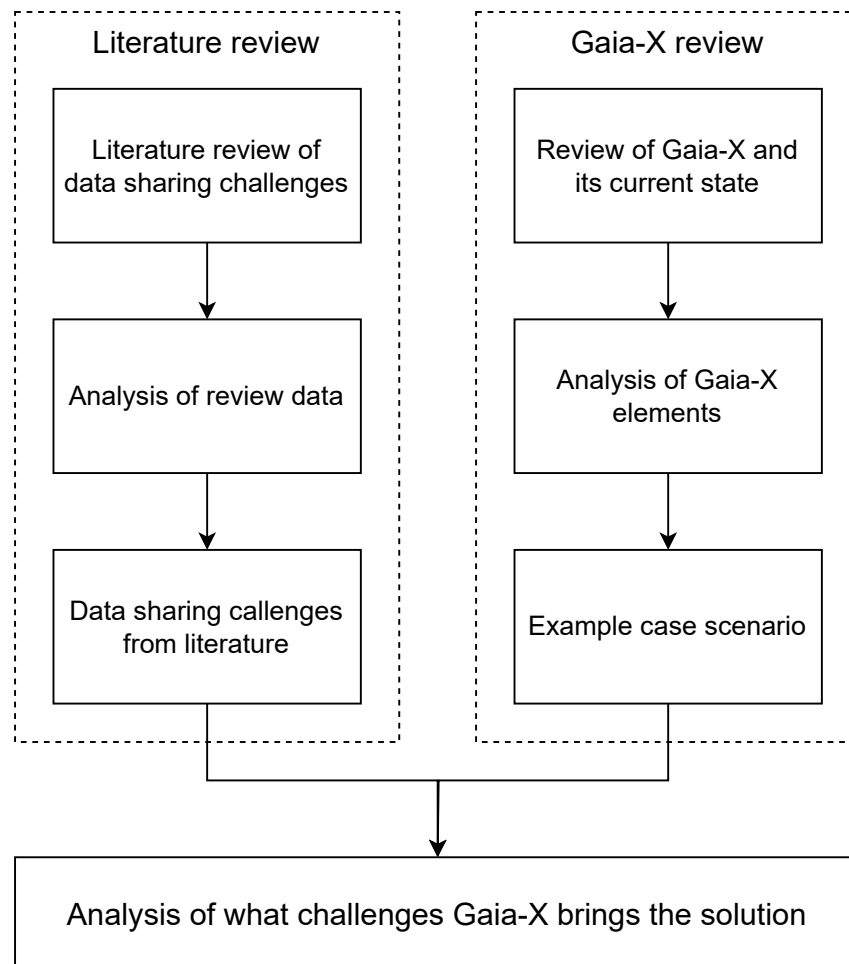
This thesis aims to identify what the existing inter-organizational data sharing challenges are in multi-actor environments and to analyze how Gaia-X solves the challenges and what are the gaps of the data infrastructure. The identification of data sharing challenges supports the development of the infrastructure by defining the requirements. Furthermore, they can be used as a base for organization and case specific analysis of chal-

lenges. On the other hand, the analysis of Gaia-X supports organizations to find suitable solutions to their challenges and hopefully contributes to the standardization of data sharing.

The thesis is structured as follows. Chapter 2 describes the performed research approach. Chapter 3 defines the terms and topics that are required to understand the following chapters of the thesis. Chapter 4 describes an overview of Gaia-X architecture. Chapter 5 contains the literature review of the challenges of data sharing. Challenges are considered especially from the viewpoint of data sharing in multi-actor environments. Chapter 6 has focus on analyzing Gaia-X architecture components as solutions to the challenges and gives an example case scenario how some of the components can be utilized in a data ecosystem. Chapter 7 summarizes and discusses the findings of this thesis and evaluates the research approach. Chapter 8 answers shortly to the research questions and considers areas of future work.

## 2. RESEARCH APPROACH

This thesis is divided into two parts. The first part is a literature review which identifies current data sharing challenges in inter-organizational data sharing. The second part analyses the capabilities of Gaia-X infrastructure to address the challenges. The overview of the research approach is shown in Figure 2.1 below. In the figure, the parts of the thesis are divided into smaller sub-processes leading to the results and outcomes of this thesis. The results answer the research questions specified in the introduction.



**Figure 2.1.** Research approach

The literature review is focused on finding the existing challenges of data sharing. The literature search was conducted systematically and carried out on two different databases:

ACM and IEEE Xplore. The results were limited to the years 2017-2022 and only English-language results were accepted. Out of all the results, only a few articles were selected for detailed review based on their abstracts and titles. Also, by looking at the resource references of the results of the database search, some articles were found to be included in the review. See column "Selected (indirect)" at Table 2.1. Selected articles focus on addressing data sharing from a perspective of organizations. Numbers of articles in each phase of the search process are shown in Table 2.1.

Used search statements differ depending on a database. The search statement in IEEE was ("data shar\*" OR "data exchang\*") AND ("ecosystem" OR "inter-organization\*" OR "across organization\*" OR "between organization\*" OR "dataspace\*" OR "data space\*"). Based on this, the search statement for ACM was formed to be ("data sharing" OR "data exchanging") AND ("data ecosystem" OR "inter-organizational" OR "across organizations" OR "between organizations" OR "dataspace" OR "data space"). The second statement is more specific in order to limit the number of search results more.

Database	Search Results	Based on Abstracts	Selected (direct)	Selected (indirect)	Total
IEEE	257	31	12	5	17
ACM	244	28	5	2	7

**Table 2.1.** *The number of articles selected from each database for the review.*

The data sharing challenges identified in the literature were analysed and divided into categories. The challenges were a base for the analysis in the second part of this thesis. In the analysis, the capabilities of Gaia-X infrastructure as a solution to the identified challenges were considered. The purpose of the analysis was to look at what solutions Gaia-X architecture offers to the challenges of data sharing and which challenges are still unresolved. Furthermore, an example case scenario of data sharing in a Gaia-X compliant way was described to provide a more concrete view.

### **3. BACKGROUND**

In this chapter, the central terms of the thesis are defined. The terms help to understand the context of the performed research, and the definitions aim to clarify the meaning of the terms in the context of this thesis.

#### **3.1 Data Autonomy and Sovereignty**

At a general level, autonomy is individuals' ability to have self-control of their choices. Individuals are not born with the ability because it requires psychological capacities that develop as they grow up. Consequently, all individuals are not equally autonomous and a point at which an individual becomes autonomous is unidentifiable. Furthermore, autonomy can be lost due to changes of individual's capacities or an environment. [7]

In the context of data sharing, autonomy helps to find a balance between data protection and data being accessible and shareable. Data autonomy is defined as the ability of data owners to control their data, for example, they can decide whether to use the data or destroy it. Furthermore, a data owner has the power to control data sharing. However, the concept of data owner is not clear and in many situations identifying and defining the owner of data is not easy. Additionally, the ownership alone is not enough to ensure autonomy, therefore data access rights are integral. To realize data autonomy, capacities are needed and for example, property rights are well-known and widely used in this case. [8]

In relation to the definition of autonomy, sovereignty is a wider concept. Sovereignty may be a familiar term when talking about independent states. In case of independent states, sovereignty means the state's constitutional independence which is seen as a social norm. Definition of sovereignty depends on the context and there is no essential definition for the term. However, there are a common understanding that the context of sovereignty involves rights and responsibilities. [9]

Similarly, in the context of data sharing, the notion of data sovereignty is not clear and its definition varies, which makes it harder to identify its nature. However, it can be noted that the term of data sovereignty is related to the expressed meaningful control and ownership of data. Some see data sovereignty as a right while others think it is an ability. From

both viewpoints, the unifying idea is that the owner can retain the control over their data. Therefore, values like control and power, deliberation and inclusion, and privacy are seen as part of data sovereignty. However, the unclear notion makes it difficult to identify concrete mechanisms to achieve data sovereignty. Therefore, data sovereignty is usually considered when designing system architectures or laws applicable to data processing that can be applied to individuals as well as larger units like society or country. [10]

## 3.2 Data Ecosystem

Data ecosystems are socio-technical networks [11] that aim to unlock the potential value of data by enabling secure data sharing cooperation between different actors. They avail the actors of an ecosystem by offering innovative services [1] and data. Overall, a data ecosystem is a sum of three elements: data, actors and an infrastructure.

The value of an ecosystem comes from data sources that are the main resources of data ecosystems. In data ecosystems, data is seen as a strategic resource which has an economic value in terms of supporting the business [12]. To make the business more profitable, data can be used, for example, to optimize operations or support decision-making. However, various kinds of other resources are complementary and crucial for a data ecosystem functioning [13]. For example, a common platform, software components and services are essential for providing and processing data. Usually, in distributed ecosystems, available resources are heterogeneous which requires knowledge of accessing and consuming these resources [11].

Data ecosystems are data-centered digital ecosystems that built on collaboration between various stakeholders. Actors are independent entities, such as individuals, organizations or machines, who interact with each other. Actors are characterized by having something in common, for instance organizations may be from the same industry. They join an ecosystem in order to achieve benefits or social and economic gain through cooperation. [14]

Relationships between actors can vary from a simple supply chain to a complex network. The actors are connected to each other by a common interest or business models. The relationships between actors determine an organizational structure of a data ecosystem. [13] Collaboration can be vertical or horizontal. Vertical collaboration refers to, for example, supply chains. Horizontal collaboration concentrates on developing data marketplaces and data sharing environments. In addition, collaborating participants can represent different domains and industries. [1]

Data ecosystems can be closed, or open environments based on how they are created. Closed ecosystems are often maintained by a single actor while in more open ecosystems anyone can join in easily [12]. Regardless of their type, data ecosystems form infrastruc-

ture for secure inter-organizational data sharing in a way that all participants benefit from it equally. An environment is transparent and it supports interaction between actors by offering services and value creating functions [14]. By utilizing data and services provided by data ecosystems, organizations can create new business opportunities, innovations, and value [13].

Moreover, data ecosystems can be non-commercial or commercial depending on whether data is shared for free or in the form of donations, or whether the ecosystem is a marketplace where data is traded [12]. The price of data can be determined, for example, by use. On the other hand, data providers assume that they benefit from sharing data in other ways if they share their data for free. For example, applications and services developed with shared data can bring benefits to the data provider.

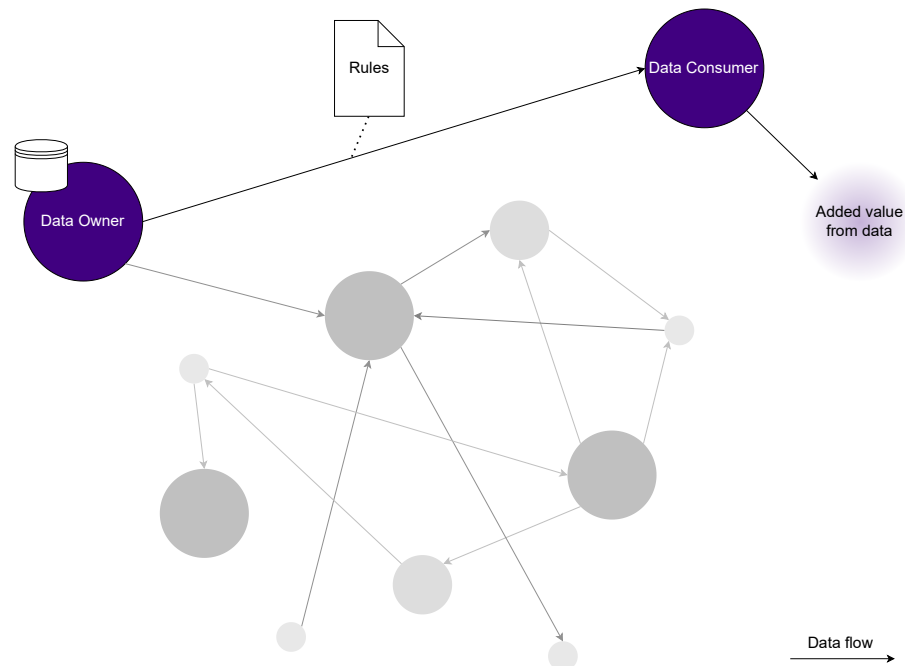
### **3.3 Data Federation and Data Spaces**

In most organizations, data from several internal and external heterogeneous data sources is used. It requires understanding about different data sources in order to access, combine and use data from them. Data integration supports data usage by providing a unified view to several heterogeneous data sources in which users can manage and combine data as if it resided in a single data source. However, integration is usually an expensive solution due to different constraints regarding an organization and data sources. [15, pp. 13–30]

Data federation meets the challenges of data integration by providing a more lightweight solution which does not require the full integration of data but satisfies the need of data integration in organizations. Data federation uses on-demand data integration, in which data is accessed and integrated only when a data consumer asks data [16, pp. 1–26]. It is also known as distributed access, and it creates an illusion that federated data sources are a single integrated data source. The illusion of access to a single data source is experienced via a federated database server. The server is the only server the end user will access directly, and utilizing a database management system uniform access to heterogeneous data sources is provided. [15, pp. 13–30]

Data virtualization is often considered as a synonym of data federation. In data virtualization, heterogeneous data sources are presented as one integrated data source by using an intermediate layer between data sources and data consumers. The layer hides all technical details of data sources, which allows a data consumer to work with a simpler interface. However, data federation is just one aspect of data virtualization. [16, pp. 1–26]

Other distributed data integration concept, data space, is based on a distributed software infrastructure. See Figure 3.1. The infrastructure provides a functionality which supports data sharing and ensuring data sovereignty. Data sovereignty is a requirement to ensure



**Figure 3.1.** Data space overview.

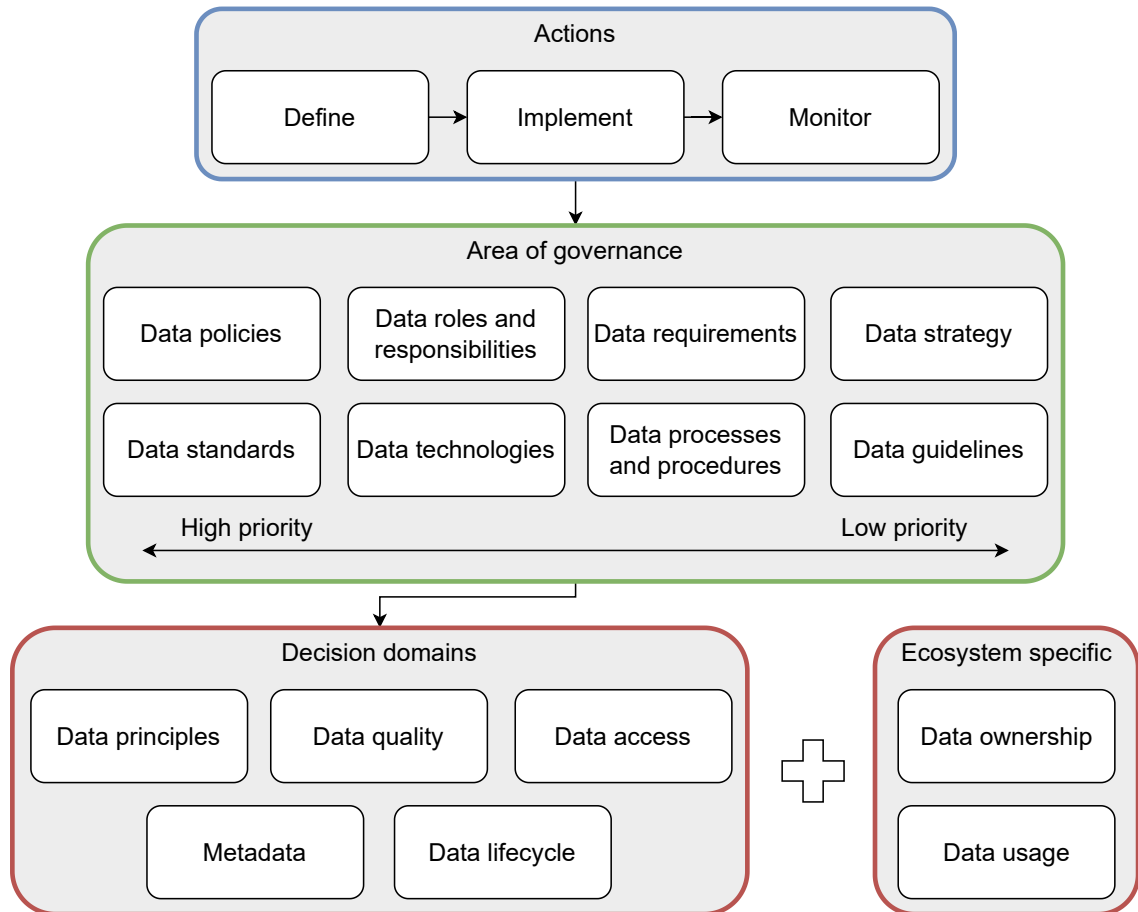
fairness in data sharing. [17]. However, data spaces are not a data integration approach but rather a data co-existence approach which provides base functionality for all heterogeneous data sources [18]. They differ from classical data integration approaches in terms of data storage. In data spaces, there is not a central data store and data is stored at its source. Common database schemas are not required, and data is exchanged directly between two participants of data space. [17]

### 3.4 Data Governance

Definition of data governance is not clear and easy to confuse with data management and it varies depending on the application context. However, roles and responsibilities, policies, metrics, definitions and the lifecycle of data are the focus of data governance [19] and these characteristics arise also from the application specific definitions. One way to define data governance in the context of ecosystems is “Arranged institutions and structures to ensure that individuals behave in line with the collective goals, conflicts between individuals are prevented or resolved, and the effective and fair use of collective resources within the inter-organizational collaboration” written by Van den Broek and Van Veenstra [20, p. 2]. Some sources refer to data governance as data management but in fact data management is the technical implementation of data governance [19].

Organizations benefit from a good data governance in many ways. Such benefits are, for example, improved data quality and data-driven decision-making, increased data accessibility and availability, decreased data management costs and better business per-

formance. Furthermore, successful data governance reduces the risks associated with the data and improves compliance with regulatory and other requirements. Despite these benefits, the main purpose of data governance is to improve the management of data and support to deliver the value to the organization. However, insufficient data governance will prevent an achievement of the benefits and in the worst-case scenario, it may cause negative consequences. [19]



**Figure 3.2.** Data governance activities model based on [21] with ecosystem specific domains from [22].

In practice, data governance is a set of activities, policies and processes to be followed through the data lifecycle. Alhassan and others [21] proposed an activities model of data governance illustrated in Figure 3.2. The model summarizes different areas of data governance to be defined, implemented, and monitored in organizations to ensure successful data governance. The model also prioritizes the areas and organizations should pay more attention to the high priority areas. The areas are applied in each decision domains. For example, an organization should implement policies and roles for data quality and data access. Lee and others [22] discuss data governance from the point of view of ecosystems, and they identify two more decision domains: data ownership and data usage. These ecosystem context specific decision domains should be also considered

when data is shared and managed across different organizations and the context of this thesis.

### 3.5 Data Sharing and Trust

Data sharing is defined as a data exchange process where open and available data formats, standards and known process patterns are used. Participating organizations and individuals who have the access right to a data asset can use its metadata and data. [23] The difference between data sharing and data exchanging is that data sharing is wider term and data exchange is a part of it. Data sharing is a collaborative use of data by multiple parties in order to achieve a shared goal [17]. Data must be trusted in order to maximize its social and economic value [5]. Likewise, if data owners don't trust that their data will be used appropriately, they do not allow it to be used or shared.

In the context of socio-technical systems such as data ecosystems, trust can be viewed from two different perspectives: user and system trust. User trust is seen as the "subjective expectation an agent has about another's future behavior" [24, p. 75]. System trust is defined as "the expectation that a device or system will faithfully behave in a particular manner to fulfil its intended purpose" [25].

Trust is gained via interaction between the entities of a system, and it may also be recommended trust which refers to trust based on third-party interaction, not direct one-to-one interaction between two entities. Trust is highly subjective, and therefore different entities may consider trustworthiness differently. Furthermore, trust has a dynamic nature. The level of trust can change over time based on gained experiences of interaction between entities. It is also possible that trust disappears completely with time. Most recent experiences have usually greater impact on the perceived level of trust than earlier ones. Additionally, the context affects the level of trust. [26]

To increase the level, the following issues need to be considered.

- Who can access the data?
- Who can use the data?
- How the benefits from the data and its usage are distributed? [5]

In practice, this means following ethical practices in collection, sharing and usage of data in the whole ecosystem. Legislation aims to define the base guidelines for the practices. The European Union General Data Protection Regulation (GDPR) is an example of legal obligations supporting ethical data sharing and increasing trust in ecosystems. However, complying with the legislation does not necessarily ensure ethical action. [5] In addition to the ethical aspect, data security and interoperability are required to ensure trust among participants in data sharing [17].

## 4. GAIA-X ARCHITECTURE

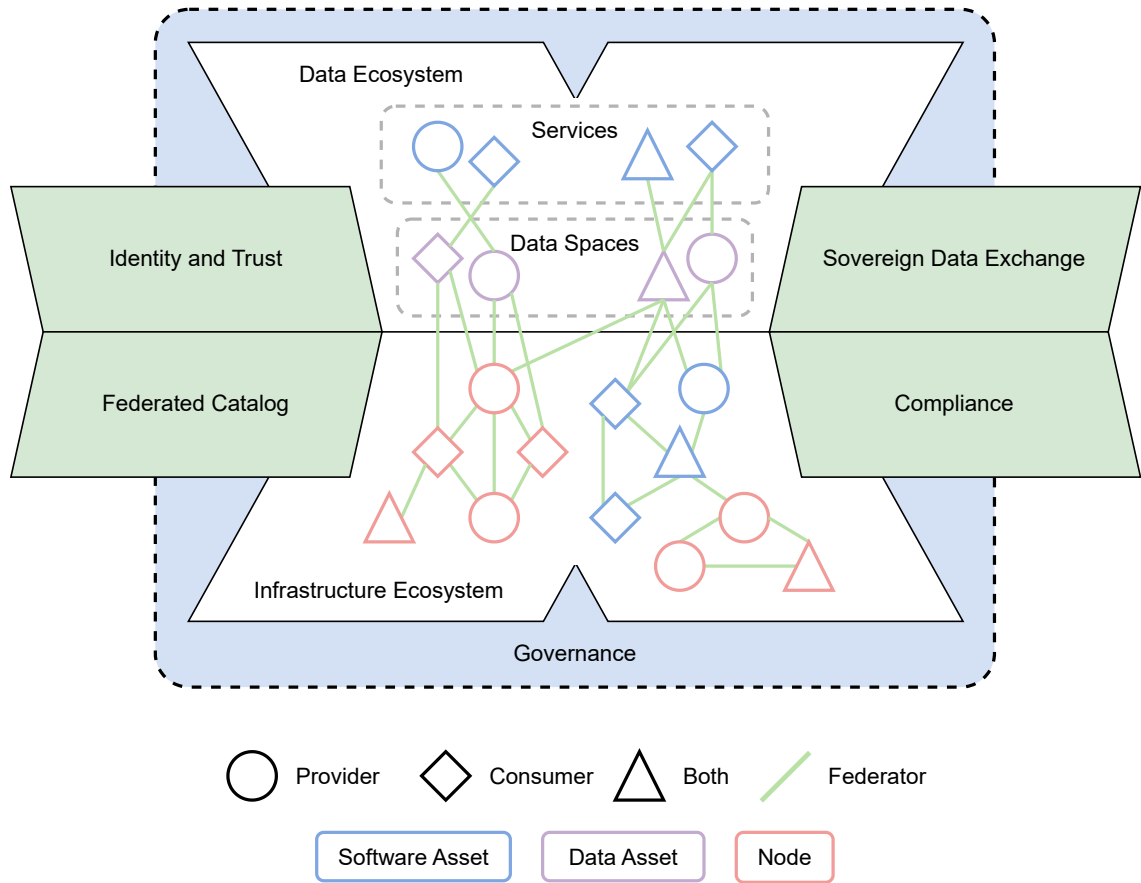
In order to get maximum benefit from data and create and improve products and services, a well-designed infrastructure is needed [5]. This chapter introduces Gaia-X architecture which aims to create a digital infrastructure for secure, transparent and sovereign data sharing. Gaia-X is a federated open data infrastructure provided by Gaia-X association to obtain transparency, sovereignty and interoperability [6].

Sovereignty, interoperability, portability, security, and trust are requirements of Gaia-X architecture addressed by defining common standards, practices and governance mechanisms for data sharing [6]. Furthermore, the data infrastructure needs to be as open as possible and protect privacy and security at the same time to maximize the value gained from data. To increase level of openness, open standards are proper building blocks of the data infrastructure. They speed up value creation because the data can be processed with easily accessible tools. [5]

Gaia-X architecture describes the components of infrastructure and the relationships between them. An overview of the architecture is illustrated in Figure 4.1. By utilizing the architecture and conforming to Gaia-X requirements, individual self-determined ecosystems, the so-called federations, become a part of Gaia-X ecosystem. In the Gaia-X ecosystem, data ecosystem and infrastructure ecosystem are connected via federation services. [6]

Participants of Gaia-X ecosystem can be either natural persons or legal ones [27]. Participants of an ecosystem have different roles. A consumer and a provider are the core roles of Gaia-X ecosystem. Provider offers resources as services defined in service offering which includes terms and conditions as well as technical policies regarding the use of the offered resources. A provider is not necessarily the same entity as the actual owner of resources, they just have a legal contract with the owner to provide resources. A consumer searches service offerings and enables them for end-users. Consumers can act as providers by offering e.g., processed data. A federator is the role enabling business-to-business interaction between a consumer and a provider by offering federation services. [6]

The infrastructure ecosystem of Gaia-X ecosystem consists of computing, storage and interaction elements that are described as nodes, interactions and software resources



**Figure 4.1.** Overview of Gaia-X Architecture redrawn based on [6, p. 7]

with Gaia-X concepts [6]. Nodes are defined as computational entities like e.g., containers or virtual machines. Connections between nodes are detailed with interaction. Interaction can be, for example, an optical fibre or a virtual local area network (VLAN). Resources of Gaia-X ecosystem are divided into two subclasses. The first subclass is a physical resource. Physical resources, such as data centres, have a position and weight in physical space and its location and owner can be defined. The second subclass contains virtual resources. Virtual resources can be either data resources, such as data sets, or software resources. Unlike physical resources, virtual resources do not have a physical location. [27] A set of resources published as a single entry in a catalog by provider forms service offering [6].

The other half of Gaia-X ecosystem is a data ecosystem which has data spaces as a foundation for a non-restricted inter-organizational data sharing network. The data ecosystem provides services to enable data sovereignty and encourage the acceleration of data sharing. It has connected to the infrastructure ecosystem via Gaia-X federation services. Federation services are connecting elements between participants and different ecosystems, and they are needed for the federation of resources and participants. The intention of federation services is to ensure interoperability and portability of resources and data sovereignty for data sharing in a distributed ecosystem environment. They are divided into

four groups of services: identity and trust, federated catalog, sovereign data exchange and compliance. These services enable searching, discovering, and finding resources and ensure trust among participants. [6]

A federated catalog provides an interface to search, filter and query service offerings so that consumers can find the best-matching one. In practice, a federated catalog is a repository containing self-descriptions of participants and their service offerings. [28] All entities, such as participants, resources and service offerings must be described with self-descriptions that are machine-readable descriptions of entities to enable finding and comparing them. Self-descriptions are realized by utilizing the Resource Description Framework (RDF) and JavaScript Object Notation for Linked Data (JSON-LD). [6] Moreover, Gaia-X uses vocabularies in their self-descriptions, for example, data resources extends W3C Data Catalog Vocabulary Version 3 (DCAT-3) [29]. To increase trust in an ecosystem, self-descriptions are world Wide Web Consortium (W3C) verifiable presentations protected with cryptographic signatures signed by participant and self-description issuer [6]. A verifiable presentation describes a subject by aggregating information from verifiable credentials that contain a set of claims about the subject [30]. The signed self-descriptions are immutable, therefore the changes of self-description are released as a new version with new signatures. The structures of self-descriptions are determined by self-description schemas. They ensure a unified representation of the self-descriptions within the federation by defining a set of attributes to describe the entities. The schemas can be extended for a specific application domain of federation. Relationships between self-descriptions are formalized as a self-description graph which enables advanced queries across individual self-descriptions. [6]

Identity and trust fill the trust gap by offering services regarding an authentication and authorization of participants in a federation [28]. Entities described with self-descriptions must have a Gaia-X compliant unique identifier. The identifier and attributes describing an entity within a given context compose an identity of the entity [6]. In practice, decentralized identifiers (DID) are used to express globally unique identifiers, addresses for entities. Gaia-X ecosystem applies Self-Sovereign Identity (SSI) architecture and principles to ensure security and trustworthiness in decentralized manner. SSI concept allows entities manage their digital identities and credentials without a central identity management system (IDM). It increases autonomy of entities regarding the personal data and a level of trust in the whole ecosystem. One of the SSI principles is to give the identity and control of data back to the owner. Entities control and store their digital identities and verifiable credentials (VC) locally in SSI wallet. [31]

Sovereign data exchange services ensure full informational self-determination for all participants by creating transparency and enabling data usage control. These services help participants to determine and keep track of data usage and by that, participants can retain sovereignty when their data is exchanged, shared or used. Sovereign data exchange ser-

vices expand existing usage control concepts, such as traditional access control. They fill existing gaps by considering usage control with requirements regarding data usage patterns. [28]

Gaia-X ecosystem has shared governance that is operationalized by the compliance service and the registry. The compliance services enable the verification of compliance for entities and are a part of the trust framework and implements a set of rules. The Gaia-X policy rule document defines general policies, which are the basis for resource-specific policies. [6] These rules are the minimum set of policies to apply to be part of the Gaia-X ecosystem, and they ensure a common governance across federations. The rules are applied to all self-descriptions of all entities in the ecosystem utilizing Rego or the Open Digital Rights Language (ODRL). [29] Additionally, domain specific governance rules can be defined to meet the requirements of each federation. Policy governs activities of participants by defining the correct or expected behavior of entities. [6]

Furthermore, Gaia-X has labels to ensure a common level of interoperability and transparency. Labels make achieving the decided level of trust easier by hiding complex verification processes and grouping compliance criteria. They are issued for service offerings only to support regulatory or business decisions. Gaia-X defines three basic labels by default. These labels are based on different clusters of compliance criteria regarding security, transparency, data protection, portability and flexibility. Higher level labels extend the requirements defined on lower levels. [32]

## **5. DATA SHARING CHALLENGES**

The potential of data and data sharing has been identified by organizations, but there are still unsolved challenges. Challenges and opportunities of data sharing are widely researched, and many technological approaches are explored to address the identified challenges in different industries. In this chapter, the findings of systematic literature review about inter-organizational data sharing challenges are described, and some existing solutions are presented. The challenges and the solutions are divided into three main categories: business, data governance and interoperability.

### **5.1 Business**

Business related challenges are separated into cooperative and competitive challenges regarding the stakeholders of a business environment. Cooperative and competitive challenges arise from a need for sharing data in order to gain maximum benefits from it. However, lack of trust and a risk of losing sensitive data to competitors are reducing organizations' willingness to share their data. The concern of losses caused by data sharing is creating a need to protect data and, at the worst, it is preventing data sharing.

#### **5.1.1 Cooperation**

Cooperating stakeholders work towards a common objective and cooperation may require sharing of sensitive data which is highly protected. Sharing of sensitive data contributes to the achievement of the objective and benefits cooperating organizations. However, organizations are aware of the value of their data and want to protect it. Over protecting is preventing sharing of data and achieving the objective. To increase the willingness of sharing all kinds of data, the benefits of data sharing and engagement in a data ecosystem must be made clear to every participant. Furthermore, guaranteeing sovereignty over data, even after an exchange, reduces a need of over protection. [1]

The unclear value gains of data sharing make it look unprofitable and do not attract organizations to join data ecosystems. Organizations expect to benefit from data ecosystems and data sharing, and data ecosystems must create business value to the participating organizations. Therefore, the authors of [3] underlined the importance of a shared busi-

ness model which increases knowledge of data sharing benefits and willingness to share data to other organizations. [3]

Additionally, new business relationships for data sharing cause trust issues. It is easier to trust business partners with long cooperation than create new relationships from nothing [1]. Therefore, data ecosystems often rely on existing alliances where participants know each other and work together. Knowledge of the benefits of data sharing increases willingness to share data and encourages organizations to take a risk with new relationships and opportunities. However, more efficient, and reliable way is to build a secure environment that provides means to data owners to retain the control over their data.

Furthermore, data sharing requires skills. Data ecosystems and data sharing may cause the flood of data and organizations lack skills to identify the most useful data from all that is available. On the other hand, from the viewpoint of an organization that shares data, a challenge with skills is the ability to balance privacy and competitive advantage so that neither suffers [12]. The ability is related to identifying valuable data and analyzing possible opportunities of sharing it. Analyzing opportunities is difficult, because benefits of sharing data are not always immediately apparent in an organization, but in a long run, the benefits can be invaluable.

### **5.1.2 Competition**

Competitive challenges refer to the challenges of protecting ideas for data driven services and competitive advantages from competitors. Lack of trust reduces openness in data ecosystems and prevents achievement of the full potential benefits of data sharing. Organizations do not trust each other or technical infrastructure to keep their business secrets and competitive advantages safe. They are afraid that competitors could get their hands on shared data and take the competitive advantages away. Authors of [1] discovered that the fear and the lack of trust cause a lower level of openness than the characteristics of data ecosystems describes. Therefore, a secure and trustworthy environment for data sharing is needed. [1]

Often, the business risks found by the organization affect the willingness to share data. Organizations refuse to share their data if the benefits are smaller than the price it costs. Organizations are afraid to share commercially sensitive data with competitors because it may lead to a loss of competitive advantage or affect an organization's reputation. Data sharing can be seen as an investment that requires time and other resources from organizations. If data sharing is seen as non-profitable and too risky, there is no desire to put an effort into it. [1]

On the other hand, the data sharing process can be found too laborious, in which case the organizations do not participate in data sharing. To address the competition related chal-

lenges, [33] introduces the concept of competitiveness as a driving force of data sharing and proposes a contract theoretic approach to achieve an all-wins situation and secure data sharing. The effectiveness of proposed scheme has been proven and participants can maximize their utilities by using it to design optimal contracts for data sharing. [33]

Overall, the challenges are related to the adoption of a data platform that is the enabler of data sharing and significant part of data sharing environment. In [2], the authors examine the adoption factors of data platforms from the point of view of data providers. As results, they ranked data governance the highest factor out of all factors that platform providers can control directly. Other identified factors besides data governance are the benefits and costs of data sharing, regulation, trust, and platform features such as security, reliability, and scalability. In addition, the readiness of an organization for data sharing, regarding available technology and personnel skills and knowledge, affects the adoption of data sharing platforms and willingness to join data ecosystems. [2]

Once the problems with lack of desire to share data have been overcome, the challenges of technical implementation of data sharing infrastructure can begin to be solved. However, the existence of secure data infrastructure, that enables sovereign data sharing, will reduce risks, and make data sharing more beneficial to organizations.

## **5.2 Data Governance**

In order to benefit and gain competitive advantage from shared data, data governance is a prerequisite. Authors of [3] investigated the differences of data governance used on single organization compared with the governance of a shared digital platform. They concluded that challenges of extending data governance to a multi-actor environment arise from varying interests of participating organizations that need to be considered. [3] The challenges are divided into categories based on the decision domains of data governance activities model 3.2.

### **5.2.1 Data Quality**

In sharing of data, actors in the position of the data consumers assume that they will benefit from the shared data in some way. However, the quality of data contributes to the benefit of data, and thus higher quality ensures better benefits. Unlike the decision domains in 3.2, Geisler and others [4] define data quality as a wider concept which includes data lifecycle and metadata. The three domains are tightly related to each other which justifies their discussion under the same chapter.

Defining a level of data quality to a data resource is challenging, because measurement of data quality is highly subjective and affected by many factors. Quality depends on application and the original purpose of data collection [4]. For example, the original purpose

affects the perceived quality because data fits best for the purpose which it has been initially collected for. In addition to the application, the data type and source, from which data has been collected, affect the quality [12]. Due to variations in level of quality, quality validation is desired in inter-organization data sharing. To ensure the quality of data, requirements for data quality need to be specified, and quality must be assessed. Quality assessment should be performed before data sharing and again during the usage of the data. [12]

Furthermore, data quality requires management that not every organization is capable to do. The authors of [3] interviewed organizations from a European shipbuilding network and identified based on the interviews that data quality is seen as the internal responsibility of each data providing organization. However, the interviewed organizations highlighted that each organization have different levels of capabilities for managing data quality [3]. Sometimes owners with a large amount of data see data sharing as too difficult. This is because a data owner needs to provide data services like storing and cleaning data for exchange [34]. Especially cleaning data to increase its quality is laborious and time-consuming with large amounts of data and all data owners don't have enough resources to do so. To overcome this challenge different roles in a data ecosystem are created. Hence data owners just pass their data to someone who processes data before making it available to the consumers [34].

To make data resources available and discoverable in a data ecosystem, they are described with metadata. Since data is collected from heterogeneous data sources, describing data in a consistent manner is challenging but essential for filtering and comparing data resources. Especially in context of data ecosystems, metadata and its management are critical success factors because every entity and asset are described with metadata. If descriptions are inconsistent in form, it will affect the efficiency of data query and management processes. [12]

Additionally, a distributed nature of data space -based data ecosystems creates challenges to managing data, especially with metadata. Metadata curation covers the lifecycle stages of metadata, such as creation, selection, preservation, and quality assurance. Implementing metadata curation in a distributed data ecosystem, however, requires the creation of a distributed curation environment. The amount of metadata is large in multi-actor environments, and most curation methods do not scale to meet the needs of data ecosystems. However, Louvre, a metadata curation framework, offers a wide range of processes for curating metadata in the context of data ecosystems. [35]

When talking about data lifecycle and data quality, transparency is an aspect to consider because it is crucial for data providers but difficult to achieve. Throughout the lifecycle of data it is necessary to monitor, how and by whom their data is used from its creation to its use. However, a high level of transparency conflicts with an objective of confidentiality of

competition relevant information. [1] Furthermore, achieving data transparency becomes a challenge when more complex networks of data ecosystems are formed, and levels of transparency may differ across the nodes of a network [4]. The challenge is to find suitable means to monitor data usage, and the means should also be applicable in a decentralized environment. Geisler and others [4] suggest distributed ledger technology and secure multi-party computation services as a base for developing a technical support for improving data transparency.

Geisler and others [4] believe that lack of accountability for data transparency is one of the reasons for slow adoption of data ecosystems. The slow adoption is especially evident when alliance-driven business-to-business (B2B) platforms are adopted. On the other hand, keystone-actor-driven business-to-consumer (B2C) platforms and data ecosystems work successfully. [4] The finding indicates that transparency is more valued issue in inter-organizational data sharing than in data sharing between natural persons.

### **5.2.2 Data Ownership**

Concepts of data sovereignty and data ownership cause challenges because they are not defined clearly, for example, by law. Usually, law only considers the aspects of data privacy and doesn't see data as an asset [36]. For this reason, the importance of data governance in inter-organizational data sharing as an enabler is huge. In data governance activities model, in Figure 3.2, data policies are defined as one way to implement data governance. Conceptualizing and implementing policies to manage the usage of data are important in inter-organizational data sharing to ensure sovereignty and retain the owner's control over the data. Zrenner and others [37] list some requirements for usage control policies. The policies define the allowed usage of the data. According to the requirements for data sharing in a multi-actor environment, the policies need to be scalable because different actors have a wide range of requirements regarding them [37].

The policy rules regarding legislation and constraints on data usage are the basis of data sharing agreements that describe how the shared data should be treated. Participants of an ecosystem may come from various countries, in which case legal systems in different countries may have differences in legislation [1]. For example, in the case of General Data Protection Regulation (GDPR), a challenge is also brought by country or state legislation extending the EU regulation. Because the legislation is divided into many levels, it creates a complex environment for data sharing [12].

Furthermore, the division of legislation and conflicts between business needs and privacy considerations lead to challenges in the creation of data sharing agreements. Depending on the application context there are different priorities when it comes to policy rules. These priorities must be considered, especially in conflict situations. Authors of [36] propose an analysis to capture and solve conflicts between rules using priorities. However,

the priority rules are introduced manually in the presented implementation, so an automatization of this process leaves a gap in research. [36]

A decentralized environment creates challenges for the enforcement of contracts and restrictions on data use. It is difficult to keep track of data and the applications that use it [12]. Compliance of data usage control policies must be perceived to ensure trust between organizations and mechanisms to track data usage are required. The technical enforcement of policies supports data exchange between organizations instead of relying only on trust [37]. Therefore, the authors of [37] present several usage control architectures to support data sovereignty. Furthermore, the authors of [38] argue that to ensure sovereignty, a common vocabulary and a usage control policy specification language are needed to define.

Additionally, there are several alternative solutions to reinforce data sovereignty and policy enforcement. Moreno and others [39] propose security pattern called BlockBD to improve the veracity and traceability of data. The proposed pattern uses blockchain as a distributed ledger to register all performed operations on the data, which increases the veracity of the data and the insights obtained from the data analysis will be more reliable [39]. Also, in [38], an architecture for access and usage control between multiple organizations is proposed. The proposed architecture is based on the International Data Spaces (IDS) reference architecture model, the Usage Control (UCON) model, and an extended eXtensible Access Control Markup Language (XACML) reference architecture [38].

The most of existing approaches focus on the sovereignty of data exchange in the context of business ecosystems and limitations of hardware resources are ignored. Therefore, Qarawlus and others [40] implemented two lightweight communication schemes to be able to examine real-time sovereign data exchange in cloud connected internet of things (IoT) devices. One of the schemes was based on request-response and the other on publish-subscribe pattern. Both schemes followed the IDS guidelines by ensuring data sovereignty through usage control policies attached to the shared data. [40]

### **5.2.3 Security**

Centralized data sharing solutions increase the risks of data leakages and misuse of data. Traditionally, data sharing relies on a third-party data center and a consumer has direct access to obtain and process data. This kind of centralized solutions in which encrypted data is stored in one data store maintained by a single authority are more vulnerable to attacks than decentralized solutions [41]. Data attacks may cause data leakages and sensitive data to fall in the wrong hands which violates the privacy and confidentiality of data. Such security concerns are usually associated with data storage in an external

cloud [42]. Furthermore, attacks affect the business and may cause major financial losses to an organization.

Organizations utilize varying mechanisms to protect their data that prevents efficient data sharing. Data confidentiality can be protected from unauthorized access by utilizing encryption and access control mechanisms. For example, data can be encrypted before storing it in a cloud and only organizations with decryption key can decrypt the data for usage. However, there are several alternatives from which an organization can choose the most suitable one for itself. All mechanisms are not directly interoperable with each other and dealing with them causes extra work to share data. [43]

To balance data protection and data sharing, access control systems have a central role. However, the most of existing access control systems are not capable to addressing the requirements of data ecosystems. In context of a data ecosystem, scalability and support for a dynamic and collaborative environment are required features of the access control system. Therefore, authors of [44] propose an intelligent enforcement approach of attribute-based access control policies to address the requirements of complex and dynamic data ecosystems. [44]

Even though, access control systems with authentication mechanisms ensure that data can only be accessed by the entity entitled to it, the systems do not remove the chance that data can be misused by those who have access to it. Michalas and Weingarten [45] propose a protocol based on a Revocable Key- Policy Attribute-Based Encryption scheme for secure health data sharing between multiple clouds hosted by different organizations. The proposed protocol enables a data owner to define a policy related to data sharing to control access rights. [45]

Security issues are a challenge especially in data sharing scenarios in which shared data have a sensitive nature. Very valuable and sensitive data cannot be exposed to a consumer directly. Authors of [46] propose a data sharing model based on blockchain technology that enables data owners to retain full control over their data. The main idea of the model is that the source data is never exposed to a consumer. A data provider uploads a description of data to the blockchain by using smart contracts and consumers can search data with keywords. A data consumer sends an identification of data and data processing model to a provider as a transaction and the provider processes the data and returns the result to the consumer. [46]

Furthermore, privacy data owners and consumers concerns organizations. Privacy of a data owner can also be protected by keeping the identity of a data owner anonymous with encryption [41] or hiding the identity information of a data owner. Furthermore, information about data sharing and data usage actions can be hidden to make the actions untraceable and protect the privacy of data user [47]. Untraceable actions ensure that the target of

data use can not be identified. On the other hand, hiding usage information is a challenge for controlling usage and it violates the principle of data sovereignty in data ecosystems.

Chen and Xue [34] propose a decentralized data exchange solution utilizing blockchain to track data transactions. Unlike [38], in this solution, no third parties are needed because transaction logs and other documents can be recorded with blockchain technology. The blockchain enables logs to be refined by every user and guarantees that the log can not be modified. In this way, the whole data exchange system does not depend on trusting third-party and data owners can audit the data usage. [34]

Developing technologies to meet the changing challenges will help reduce security concerns regarding data sharing. The usage of blockchain technology in data sharing is a widely researched area. It has been applied in many use cases like smart city [48] and healthcare [41] industry. Blockchain technology is considered as a solution to the above problems in untrusted distributed systems. Its flexible scalability makes it easy to expand, distributed nature avoids the risk of a single point of failure [46]. Blockchain technology enables the integrity of data in a decentralized manner and ensures immutability and verifiability of transactions and transparency [41].

Overall, blockchain technology enables the development of new mechanisms to protect data and share it securely. In [48], a blockchain-based approach of exchanging data in smart city scenarios has been described. It enables the safe use of data without accessing the data from another organization. The proposed approach is also applicable in scenarios, such as medical and financial data sharing, where user privacy protection and data sharing access are required [48].

### **5.3 Interoperability**

Data sharing requires interoperability, but in multi-actor environments, it is hard to get all actors to comply with common processes and standards [4]. In context of data, interoperability consists of syntactic and semantic interoperability. The general term of interoperability is seen as systems', networks', devices', applications', or components' ability to communicate, for example, share data with each other in an effective way and make use of shared data [49]. In a multi-actor environment, interoperability challenges often arise from distributed and heterogeneous data sources or the heterogeneity of systems used by different actors. What makes it particularly challenging is that the actors have their own way of doing things and changing these habits can be a difficult and long process [4].

#### **5.3.1 Syntactic Interoperability**

Systems' capability of communicating and exchanging data is called syntactic interoperability that can be realized by utilizing, for example, specified data formats and communi-

cation protocols [49]. Syntactic interoperability challenges arise from differences between information systems used by collaborating parties. Usually, different systems use different technology stacks and do not follow the same standards that make communication and integration between these systems extremely hard or impossible [50].

Authors of [50] introduce distributed ledger technologies as an enabler for seamless collaboration and information sharing in real time. Integration architecture based on blockchain is more scalable and secure than other approaches. Furthermore, the architecture ensures transparency and improves interoperability among collaborating parties. The authors demonstrated a usage of the architecture within a collaboration of public service organizations and got positive feedback from the demonstration. Especially, ability to implement the blockchain layer separately on top of the existing infrastructure was seen as a value adding feature in data sharing. Furthermore, the authors pointed out that the biggest challenge of implementing the distributed ledger technology solution is to get stakeholders together and build trust network, not the technology itself. [50]

### **5.3.2 Semantic Interoperability**

Semantic interoperability is defined as the ability of systems to exchange information with unambiguous meaning. It ensures that individual systems in complex system-of-systems (SoS) understand shared data in a similar way. [49] Semantic interoperability is one step deeper in data interoperability compared to syntactic interoperability. This research showed that there are several ways to build semantic interoperability. However, the existence of alternatives can cause the confusion and difficulties of finding a shared way to act.

Data integration on the semantic level can be achieved through data integration systems or data spaces. A data integration system can be used to integrate data sets from different sources and to provide a unified view of integrated data. It provides a global schema that enables achieving syntactic and semantic interoperability by defining underlying data elements and assigning unambiguous meaning to data. [49] Compared to data integration systems, data spaces do not require a schema [4]. Authors of [49] tested the applicability of data spaces in the energy sector and showed that data spaces enable the integration of energy data and improve the interoperability of energy management applications and services.

Axelsson [51] investigated the use of linked data and ontologies as a means to achieve semantic interoperability within a SoS. Linked data is a generic information representation in which resources are represented by using RDF. Furthermore, International Resource Identifiers (IRI) are used as unique identifiers to represent the described resources. Unique IRIs can be generated in a way that they meet the requirement of distribution in a decentralized environment. To store and exchange RDF graphs between

systems, serialization formats are needed. The RDF linked data presentation creates a terminology to be used within SoS in order to ensure a common understanding of concepts and relationships between them when resources are represented in a specific context. This kind of terminology is also called an ontology. In addition to ontologies, the usage of vocabularies is recommended in order to unify terminology for similar concepts used in the subsystems of a SoS. There are multiple vocabularies available, and a selection of suitable vocabulary depends on the context of use. [51]

## 5.4 Summary of Data Sharing Challenges

As results of the systematic literature review, existing inter-organizational data sharing challenges are identified. The challenges are listed in Table 5.1 by category. All the challenges do not need to be solved in order to share data, and the context of a data sharing case affects their realization and criticality. However, solving as many challenges as possible supports achieving the most efficient data sharing.

The challenge categories are in order by the criticality. Interoperability is the most critical aspect, because it affects the availability of data and makes data sharing possible. Without interoperable means and systems, data sharing in a multi-actor environment is not possible and efficient from the technical point of view. Interoperability challenges prevent data transfers between systems of different actors and hinder easy access to compare and find data from several sources.

Compared with the interoperability challenges, the challenges of data governance do not technically prevent data sharing. However, data governance will affect the willingness of organizations to share their data and how risky they consider data sharing. Most of the identified challenges are included in this category. Moreover, data ownership, one of the data governance challenge subcategories, has the least advanced solutions based on the review. Therefore data ownership issues require the most attention in the bigger picture of inter-organizational data sharing.

Furthermore, the business challenges are tightly related to other two categories of challenges. The business-related challenges are the most critical when organizations consider joining a data sharing ecosystem and its benefits to their organization. If organizations do not see data ecosystems profitable, they are not willing to share their data that limits the availability of data and deducts the achievement of added value from data. By developing solutions to the interoperability and data governance challenges, the number of business challenges can be decreased.

Considering the challenges, individual organizations may have different views of what the most critical challenges are. Divergent views can cause disagreements about the order in which and how challenges should be resolved. Furthermore, the challenges list is

Interoperability	Syntactic	Heterogeneity of systems
	Semantic	Heterogeneity of data
Data Governance	Data Quality	Subjectivity of data quality Varying quality management capabilities Distributed data lifecycle management Data transparency
	Data Ownership	Inadequacy of legislation Division of legislation Enforcement of data usage control policies
	Security	Risks of data leakages and misuse Varying mechanisms to protect data Mechanisms do not meet requirements of data ecosystems Privacy of data owners and data consumers
Business	Cooperation	Protection of valuable data Unclear value gains Lack of skills to identify useful data
	Competition	Lack of trust Risk of losing competitive advantage Laborious process

**Table 5.1.** Data sharing challenges by category.

generic, and depending on the context, there may be more challenges and their criticality may vary. Therefore, Table 5.1 should be used as a guide when the context-specific challenges are identified and analysed.

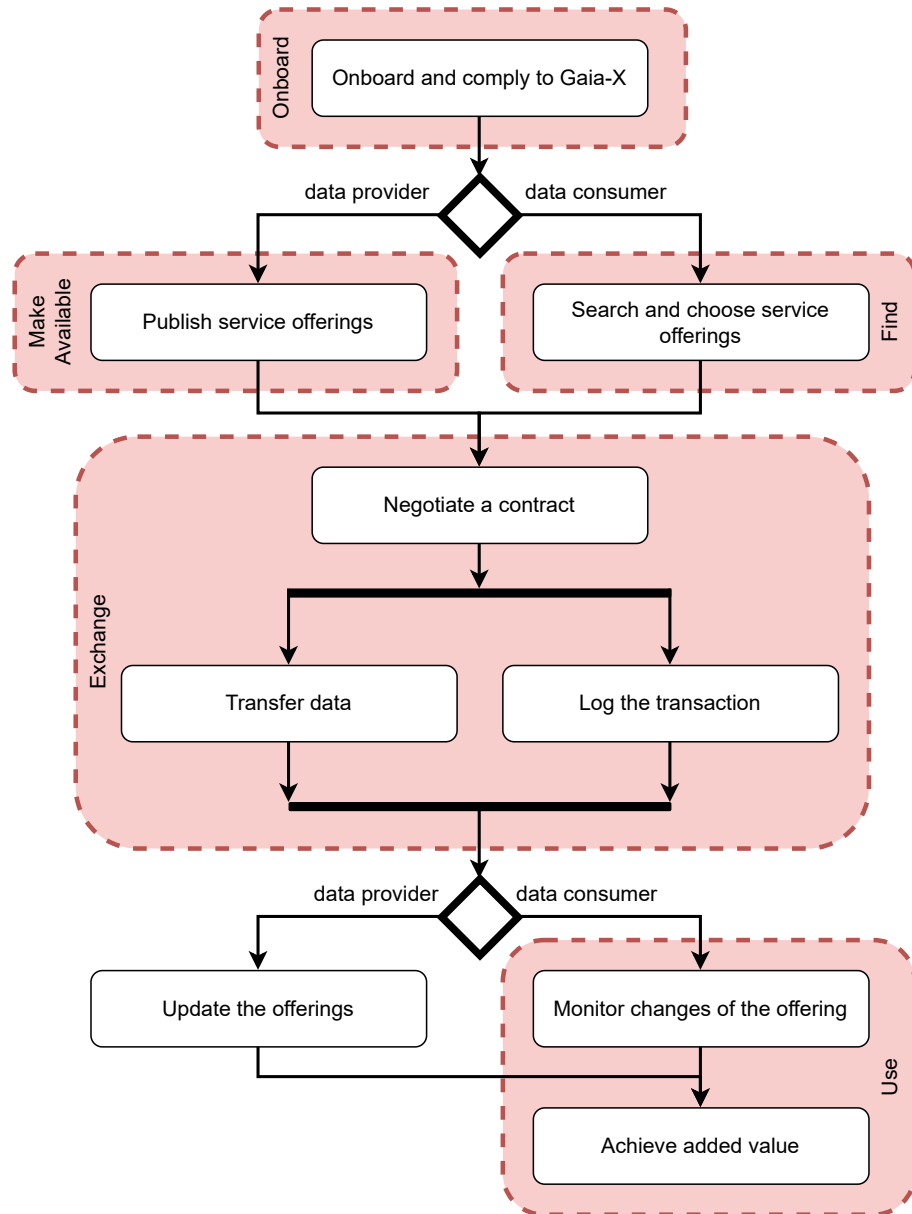
## **6. SOLVING DATA SHARING CHALLENGES WITH GAIA-X**

Once the challenges of data sharing are identified, a solution to facilitate data sharing and meet the challenges can be discussed. In this chapter, the capability of Gaia-X architecture to address the challenges is analysed. The analysis focuses on the architecture components and specifications of reference implementations of Gaia-X Federation Services (GXFS). The aim of the analysis is to identify lack of Gaia-X in relation to the challenges. Furthermore, an example case scenario is described to support understanding of usage of Gaia-X federation services in data ecosystems.

### **6.1 Data Sharing in Gaia-X Compliant Data Ecosystems**

Data sharing is a process which aims to offer added value to data owners and data consumers. To enable data sharing, organizations need to join data ecosystems that provide means for finding, transferring, and consuming data. However, the literature review showed that organizations are concerned about the costs and risks involved in data sharing, and they do not see it as profitable activity. Sensitive nature of business-related data requires trust, security and privacy to be considered in data sharing especially in larger data ecosystems where participants do not have direct business cooperation relationships, and they may be even competitors. To face these challenges, Gaia-X aims to build an infrastructure which reduces risks and facilitates equal data sharing between organizations [6]. Figure 6.1 summarizes the functioning of a data ecosystem from the viewpoint of data sharing and highlights five main phases: onboard, make available, find, exchange, and use.

In the onboarding phase, organizations adopt an ecosystem's practices and commit to shared rules. Gaia-X trust framework defines criteria that every ecosystem's participant must follow. According to the criteria, each entity must be described with Gaia-X self-description and must have a Gaia-X compliant unique identifier. Unique identifiers enable self-descriptions to refer to each other which enables performing more complex queries and improves the availability of data and the capabilities of participants to find the most useful data. Cryptographically verified identifiers also increase the level of trust in an ecosystem by ensuring participants are who they claim to be. [6]



**Figure 6.1.** Activity diagram of data sharing.

Self-descriptions describe heterogeneous entities in the consistent manner and increase interoperability of resources. The Gaia-X trust framework specifies mandatory attributes for self-descriptions in order to ensure that each data resource is described with sufficient precision [29]. The attributes are expressed as self-description schemas which can be extended by each federation for their application domain needs [6]. The schemas ensure all self-descriptions follow a common structure and semantics that makes entities described with Gaia-X self-descriptions discoverable and comparable. Self-descriptions comply with RDF, which has been shown to improve semantic interoperability [51]. Furthermore, the interoperability and deployment of self-descriptions have been improved through use of existing standards and widely used technologies, such as RDF and JSON-LD [6].

Moreover, the operation of Gaia-X compliant ecosystems is based on common policy rules that tackle the challenge of divided legislation. The policy rules describe common ground rules and principles for collaboration and participation in Gaia-X ecosystem considering legal and market requirements and covering privacy and security policies [6]. However, it is extremely hard to define common policies aligned with the divided legislation [12]. Therefore, Gaia-X focuses on harmonizing data sharing only within the European Union where legislation is already uniform [52]. The policy rules can be extended with additional rules by each individual ecosystem, and they form the base for more detailed service offering specific policy to ensure the equal level of privacy and security to all participants of an ecosystem [6]. Ensuring equal privacy for participants, regardless of their own capabilities, will help to increase the willingness of sharing data.

The privacy of data owners and consumers are equally important, but their realization can not be implemented identically to be able to comply with the principle of sovereignty. To protect data owners, their identity can be kept anonymous [41]. In the case of data users, the privacy issue is not so simple if the principle of sovereignty is respected. Hiding information about data usage ensures that actions cannot be traced [47], and information of data user or the target of use cannot be identified. According to the data sovereignty principle, data owners should know who is using their data and how the data is used [6]. To be able to track data usage the data user cannot be anonymous. However, the data owner does not need to know the purpose of use if it can be ensured that the policies for the data usage are followed.

After onboarding, data owners have a Gaia-X role called data provides which enables them to make their data available to the rest of the ecosystem. A data provider can tie multiple data resources together and publish them as one service offering in a federated catalog [6]. Service offerings require their own self-descriptions which include terms and conditions as well as technical policies regarding the use of the offered resources [29]. Furthermore, data providers can attach labels to their offerings that describe different levels of criteria regarding security, transparency, data protection, portability and flexibility [6]. The labels indicate the quality level of data but do not consider its subjective nature.

On the other hand, onboarding enables data consumers to gain access to available service offerings published in a federated catalog by data providers. Gaia-X federated catalogs provide filtering options that help data consumers to find best-matching service offering to their purposes [53]. Because service offerings are described with human and machine-readable self-descriptions and labels[6], data consumers can automate processes, such as filtering and monitoring, which makes it easier to all participants to find and identify the most useful data. Furthermore, self-descriptions describe the context of data which helps participants to understand the purpose of each data resource and what kind of use cases they are suitable for.

When the proper service offering is found, the data consumer and the data owner negotiate a contract that restricts the usage of agreed resources. Contract negotiation enables data owners to choose their customers and set the price to their service offerings which reduces the financial risks of data sharing. The negotiation process results an agreement about the service offering and regulations regarding its usage [6]. Data owners' ability to regulate data with data usage policies improves their control over their data and increases the level of trust in an ecosystem. In Gaia-X ecosystems, negotiating participants can utilize a data contract service which saves the resources of organizations by supporting automatic and semi-automatic negotiation processes [54].

Data exchange can begin when the contract is finalized and have been signed by both parties. In the data exchange phase, a data set is transferred to the data consumer or some other access to the data is provided. Gaia-X do not define how a data transfer should be performed or an access provided. The parties involved in an exchange can agree access to data or transfer mechanisms and formats on a data sharing contract. On the other hand, Gaia-X encourages participants to log data exchange actions via a data exchange logging service. The logging service provides evidence about successful data exchange which can be utilized for clearing or billing [55].

Utilizing access control systems enables secure data sharing and reduces the risk of misuse. In context of distributed data ecosystems, such as Gaia-X compliant data ecosystems, scalability and support for a dynamic and collaborative environment are required features of an access control system [44]. An access control system with authentication mechanism makes operating in data ecosystems more secure by reducing the misuse of data and ensuring that the data can only be accessed by the entity entitled to it. However, Gaia-X does not provide an access control mechanisms for data, just for their federation services [54]. Because Gaia-X compliant ecosystems are distributed, and a data owner stores the data in the way which meets best their requirements, data providers are responsible for secure access to their resources [6]. This means that data providers do not need to trust third-party provided centralized data storage, that reduces the risk of data leakages.

Moreover, data can also be misused by those who have access to it, so the enforcement of usage policies is essential. The blockchain-based data sharing model proposed in [46] reduces the risk of misuse, because the original data is not exposed to a data consumer and the data owner is only who can access and process the data. However, in Gaia-X context the data sharing is based on data transactions and the original data set is transferred to the consumer for processing [6]. Therefore, mechanisms for enforcement of contracts and restrictions should be provided. At this point of development of Gaia-X, federation services are unable to provide the necessary evidence of a violation of usage policies [55].

Once the data consumer has received the data, they can utilize it and achieve added value through data processing. Gaia-X architecture is focused on improving the availability of data and does not provide any means for data processing. However, the architecture enables sharing of all kinds of physical and virtual resources, so participating organizations can offer data processing algorithms and software to each other [29]. This reduces the differences in data processing capabilities between different organizations, at least on some level, even if it does not prevent the existence of the differences.

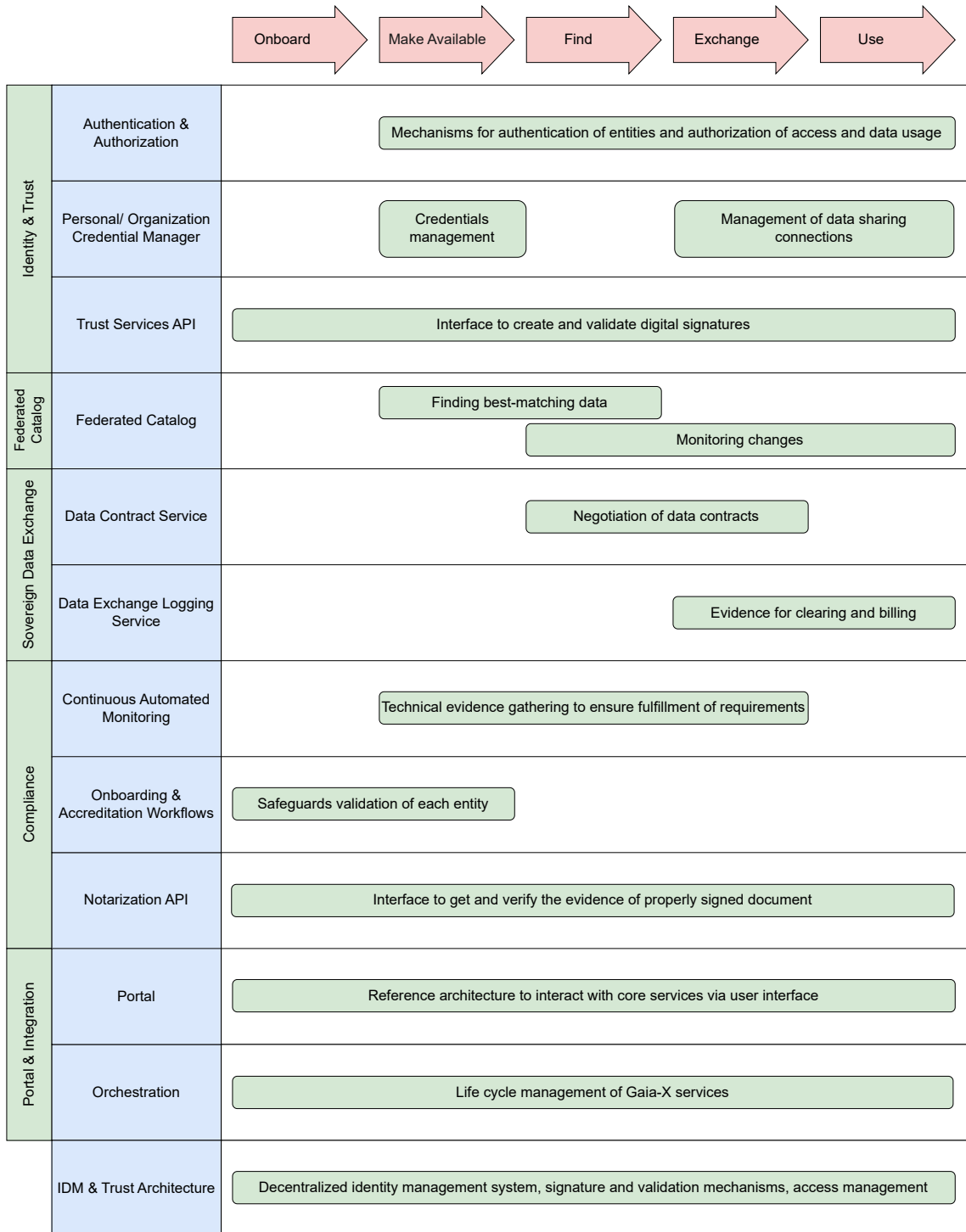
In addition to data consumers, data owners expect to get value from data sharing which may require payment service. However, possible value gains are unclear to organizations and data sharing is not seen profitable operation [3]. Data sharing contracts enable data owners to define a price for their data, but ecosystems need means to realize such payments. At this point of development of Gaia-X, the infrastructure does not contain any billing service and data providers need to offer the service by themselves if they need one [54].

As a summary, Gaia-X architecture advances the creation of data ecosystems that enable equal opportunities for the participating organizations to benefit from data sharing. It facilitates the joining into an ecosystem as a new participant and improves interoperability at several levels across ecosystems. Existing Gaia-X architecture components make data from heterogeneous sources findable and comparable to all participants and enable finding the best-matching data for their purposes. However, the transaction of data and any billing services required are under the responsibility of the contracting party. Furthermore, data sovereignty, one of the main principles of Gaia-X, also requires work and further development of the architecture. The architecture enables data owners to define the usage policies of data and decide what information related to a data set is shared and who can access the data. However, the enforcement of the policies is still a mystery. Additionally, data owners are responsible for safeguarding data confidentiality in terms of data storage and access control systems by themselves.

## 6.2 Supply of Gaia-X Federation Services

In addition to the architecture, Gaia-X provides reference implementations of federation services as open-source code. By providing the reference code, Gaia-X facilitates the adoption of the architecture in data ecosystems and improves inter-organizational data sharing. Figure 6.2 lists federation services and their current reference implementations, and describes their purposes from the point of view of data ecosystem functions based on the product specifications. Implementing all services is not necessary, and data ecosystems can choose a limited set of products that meet their specific needs.

*Authentication and authorization* makes data ecosystems secure and reduces the risk of data sharing. It is a product that provides access management functions, in addition to



**Figure 6.2.** Scope of Gaia-X federation services.

the authorization and authentication functionalities, to enable the usage of SSI concepts [56]. Furthermore, the product is integrative with basic identity and access management (IAM) systems [56]. The integrative nature of the product enables ecosystems to utilize their existing IAM systems and save on costs. Low costs of data ecosystems will decrease the financial risks of data sharing and make data ecosystems more profitable for

organizations. Additionally, offering layers to be added on existing software will facilitate the adoption of Gaia-X compliant systems.

*Credential managers* act as interfaces that support communication between an ecosystem's participants. Organizational and personal credential managers allow organizations to manage their self-descriptions and connections to other parties in trustful and secure way [57], [58]. In accordance with the sovereignty principle, data providers can decide which information regarding their resources and service offerings are published in a catalog by forming verifiable presentations from credentials [6]. The freedom of data providers to choose which information is published within an ecosystem contributes the achievement of data sovereignty. In more detail, a personal credential manager provides an interface for natural persons to interact with an ecosystem and store their verifiable credentials in one place [58].

*Trust services API* ensures a consistent level of trust in an ecosystem by providing an interface to communicate with trust services. All other architecture components and federation services can utilize trust services API to sign and validate digital signatures [59]. The creation and validation of signatures enable trust chains where all participants do not need to trust each other directly. It is enough for the participants to rely on the trust anchors. Furthermore, the product improves trust between an ecosystem's participants by ensuring trust chains and building policy driven trust by defining functions for policy evaluation [59]. Trust services are provided by qualified certificate authorities called trust service providers who belong to Gaia-X trust anchors [6].

*Federated catalog* forms a uniform view to heterogeneous resources and makes them comparable and searchable. It acts as storage for self-descriptions and a self-description graph which contains relationships of resources and enables complex queries across them [53]. Functionality of federated catalog enables data consumers to find best-matching data to their purposes and monitor changes in offerings while protecting valuable data. Because federated catalog does not provide access mechanisms to resources directly [53], data can be stored safely in distributed manner. Distributing supports data protection and gives data owners the freedom to decide on the security features of their data storage. When a data consumer shows interest in a service offering, data consumer and data provider negotiate how the data can be accessed.

*Data contract service* increases the control of data owners and releases human resources of an organization for other purposes. Data providers can choose their customers and personalize the price and usage terms of an offering depending on the customer [54]. The ability of data owners to specify service offering specific policies strengthens control over their data and follows the sovereignty principle of data ecosystems. Furthermore, data contract service supports automatic and semi-automatic contract negotiations to create legally waterproof contracts [54]. Automatic contract negotiations do not require as much

human resources, in which case, organizations do not need to put so many resources to data sharing.

*Data exchange logging service* increases the traceability of data but lacks the required level of capabilities to enforcement data usage policies. The technical enforcement of policies supports data sharing between organizations instead of relying only on trust [37]. Gaia-X prefers trust creation through technical means and offers a reference implementation to data exchange logging service to record data transactions and to provide the evidence of enforcement and violation of data usage policies [55]. The Logging service improves the traceability of data even though the service does not provide required measures for policy enforcement. Lack of means to enforce data usage policies builds an obstacle to the full realizations of data sovereignty.

*Continuous automated monitoring service* improves the security and interoperability of data ecosystems. The service evaluates continuously whether service offerings published in a catalog adhere Gaia-X principles [60]. The evaluation is based on Gaia-X general policy rules that cover security, privacy, transparency and interoperability during onboarding and service delivery [6]. The monitoring service ensures that an ecosystem's participants and their service offerings follow the rules and meet the necessary criteria for forming a united and functional ecosystem. Furthermore, the evidence of compliance gathered by continuous automated monitoring service is essential to ensure secure data sharing in cases that involve dealing with sensitive data [60].

*Onboarding and accreditation workflows* increase transparency and improve the security of ecosystems. The workflows ensure that each participant and asset in a federated catalog fulfills the Gaia-X requirements that foster secure and reliable participation in data ecosystems. During the onboarding workflow, relevant onboarding data, such as self-descriptions of participants and assets are generated, and accreditation agreement and the terms and conditions are signed. The accreditation workflow verifies onboarding information and creates verifiable credential utilizing Notarization service. Onboarding and accrediting data assets are handled by data contract service. The workflows provide transparency about the compliance of assets. [61]

*Notarization API* establishes digital trust by providing interfaces to make any credentials verifiable in standard format. Notarization service allows an authorization officer to attest given master data and transform it to tamper-proof credential to enable the verification of identities and documents, such as the claims of resources. The implementation of notarization service will provide digital support for existing certification bodies. However, each individual ecosystem does not implement notarization service themselves, but there will be one European wide implementation. The implementation includes REST APIs to facilitate integration with other software. [62]

*Portal* facilitates data sharing by gathering the main functionality of a data ecosystem into one place. It contains the core functionalities that a minimal viable product of Gaia-X requires [63]. A portal makes Gaia-X infrastructure easy to approach, which contributes to its adoption in ecosystems. The core functionalities include displaying and searching the content of Gaia-X federated catalog through a web-based user interface [63]. The interface eases exploring of available resources in data ecosystems and makes ecosystems more user-friendly to their participants.

*Orchestration* provides an API for lifecycle management of service offerings improving the quality of the offerings. The standard API enables ecosystems to utilize existing deployment and management technologies, and therefore an implementation of methods is not described in the Gaia-X specifications [64]. However, methods to manage the lifecycle of data are required in data ecosystems and have an impact on the quality of data [4]. The lifecycle management of Gaia-X orchestration service is limited to service offerings, and it can not be utilized to manage the actual data resources. In other words, the service does not improve the quality of data resources.

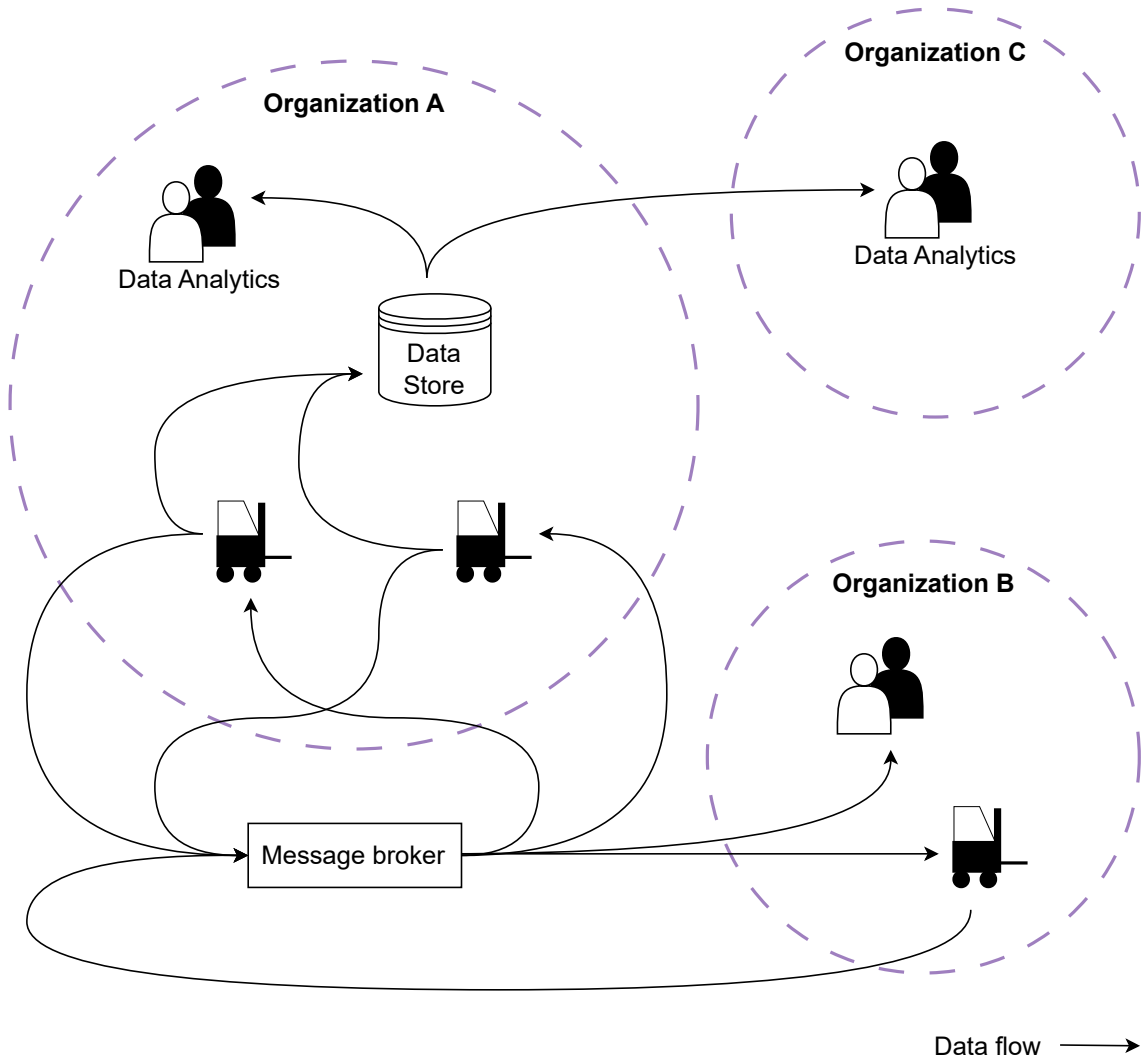
*Identity management and trust architecture* supports self-sovereignty, security and trust in data ecosystems and reduces costs and improves interoperability by utilizing existing standards. The architecture is based on SSI and W3C standards for DID and VC that enables the creation of a secure and self-sovereign identity management and trust mechanisms [65]. Use of existing standards reduces costs of participating in an ecosystem and increases interoperability across different ecosystems. Furthermore, the architecture has a decentralized nature, and any government authority is not able to control the Gaia-X ecosystem [65]. Decentralization allows ecosystems to operate more autonomously.

Several federation services are focused on creating trust while other important aspects, such as security, data sovereignty and quality remain in the background. Ensuring secure data exchange is responsibility of interacting actors and data owners should provide access control for their data resources. Similarly, data quality and data lifecycle management are in the hand of data owners. Furthermore, contrary to expectations, federation services do not support the enforcement of data sharing contracts and usage policies with technical means, which would be important in terms of sovereignty and trust. However, the overview in Figure 6.2 will support organizations to choose needed federation services in their ecosystems.

### **6.3 Example Data Sharing Case Scenario**

This example case scenario describes how Gaia-X principles and federation services can be applied in the operations of a small data ecosystem. The example does not include all federation services because some of them are not relevant in this scenario. However, if the example ecosystem wants to scale their data sharing with other ecosystems, some

of the federation services will become relevant to enable data integration between two or more individual ecosystems.



**Figure 6.3.** Data sharing example between multiple organizations.

In the example case ecosystem, there are three different sized organizations with distinctive needs for data sharing and usage. Additionally, the organizations share real-time and historical data, so the ecosystem requires an infrastructure to support data sharing of both types of data. Figure 6.3 illustrates the ecosystem and the data flows between the participating organizations within the ecosystem. The data shared in the ecosystem is derived from mobile machines owned by Organization A and Organization B. Each machine has a state information containing a location data of a machine and a code of the task the machine is performing. The machines are sharing the data in real time through a message broker. Furthermore, Organization A stores the historical data of routes and tasks performed by their machines. The organizations do not have a shared data store, so data sharing occurs in distributed manner.

Organizations want also to restrict the availability of data to ensure only specific actors can access the data shared with them. The real time location data of each machine is available to all. However, the ongoing tasks are only available to a specific team in Organization B. The team manages task priorities and makes decisions regarding which mobile machine is going to perform a task and when. The collected historical data is used by data analytics teams of Organization A and Organization B. Only the parties, to whom the data is useful and needed, can access it. The ecosystem must provide means to restrict the access and usage of the data. However, the organizations included in the case cooperate towards a shared goal, and they do not try to achieve financial profit from data sharing, so any payment services are not needed.

Because Gaia-X do not provide federation services as products, but reference implementations, each individual ecosystem must implement the federation services they need by utilizing the reference implementations or just the federation service specifications. In this example case, Organization C acts as a federator who provides the federation services for the ecosystem. The federation services of this case are: all compliance services as well as authentication and authorization services, federated catalog, data exchange logging service and portal.

Each organization needs its own Gaia-X compliant identifier and self-description which are achieved by going through the onboarding and accreditation workflows. In the example use case, Organization A and Organization B act as data providers and consumers so their identity contains both roles. However, Gaia-X trust framework [29] does not yet describe attributes for different roles. As data providers, organizations write claims about its resources and sign the claims utilizing a trust services API. After signing, the claims are approved through a notarization API. The compliance of the verifiable credentials is verified by a continuous automated monitoring service when they are published in a federated catalog. As data consumers, organizations can explore offering from the catalog and verify the claims and identities of other participants through the notarization API.

Considering the example use case, Organization A wants to offer the real-time location data of all its mobile machines as a single offering. Organization A creates a service offering and describes it with verifiable credentials. The service offering is a verifiable presentation which goes through the same process as the verifiable credentials of a participant and a resource. An example verifiable presentation is described in Figure 6.4. After the verification, Organization A can decide, in a self-sovereign manner, which information they want to publish in the federated catalog. Utilizing a credential manager, Organization A chooses the information describing the offering and publishes it to the catalog.

Once Organization A has published the self-description of the service offering, it is available to Organization B. Organization B finds the service offering by searching for alter-

```

{
  "@context": {
    "cred": "https://www.w3.org/2018/credentials/v1",
    "gx": "https://docs.gaia-x.eu/policy-rules-committee/trust-framework/22.10/"
  },
  "@id": "did:example:OrganizationA/machines/all/locations",
  "@type": ["VerifiablePresentation"],
  "gx:providedBy": "https://www.example.org/participants/OrganizationA",
  "gx:termsAndConditions": [{
    "gx:URL": "http://www.example.org/terms_and_conditions",
    "gx:hash": "7dbfc7d1639da6f849f66a3c358292b19e247b0a527d62142a2dedc62a5418c6"
  }],
  "gx:policy": [{
    "@default": "allow"
  }],
  "gx:dataAccountExport": [{
    "gx:requestType": "email",
    "gx:accessType": "digital",
    "gx:formatType": "application/json"
  }],
  "verifiableCredential": [{
    "@id": "did:example:OrganizationA/machines/1",
    "@type": ["VerifiableCredential"],
    "cred:issuer": "did:example:issuer",
    "cred:issuanceDate": "2022-06-12T19:38:26.853Z",
    "cred:credentialSubject": {
      "@id": "did:example:OrganizationA/machines/1/location",
      "@type": "gx:DataResource",
      "gx:copyrightOwnedBy": ["Organization A"],
      "gx:license": ["CDLA Permissive-2.0"],
      "gx:policy": [{
        "@default": "allow"
      }],
      "gx:producedBy": "http://www.example.org/participants/OrganizationA",
      "gx:exposedThrough": ["http://www.example.org/services/MessageBroker"],
      "gx:containsPII": false
    },
    "cred:proof": {
      "@type": "JsonWebSignature2020",
      "cred:verificationMethod": "did:example:issuer#key",
      "cred:created": "2022-06-12T19:38:26.853Z",
      "cred:proofPurpose": "assertionMethod",
      "cred:jws": "z2iiwEyyGcqWLPMDXnjEdQU4zGzWs6cgjrmXAM4LRcFxn1PpZ44EBuU6o2EnkXr4uNMVJcMbaYTLBg3WYLbev3S"
    }
  }], {
    "@id": "did:example:OrganizationA/machines/2",
    "@type": ["VerifiableCredential"],
    "cred:issuer": "did:example:issuer",
    "cred:issuanceDate": "2022-06-12T19:38:26.853Z",
    "cred:credentialSubject": { ... },
    "cred:proof": { ... }
  }],
  "proof": {
    "@type": "JsonWebSignature2020",
    "cred:verificationMethod": "did:example:issuer#key",
    "cred:created": "2022-06-12T19:38:26.853Z",
    "cred:proofPurpose": "assertionMethod",
    "challenge": "1f44d55f-f161-4938-a659-f8026467f126",
    "domain": "4jt78h47fh47",
    "cred:jws": "z2iiwEyyGcqWLPMDXnjEdQU4zGzWs6cgjrmXAM4LRcFxn1PpZ44EBuU6o2EnkXr4uNMVJcMbaYTLBg3WYLbev3S"
  }
}

```

**Figure 6.4.** Example self-description of a service offering.

natives from the catalog using filters and keywords through a portal. Self-descriptions allow Organization B to compare different service offerings and to find the best-matching offering. The same service offerings are also discoverable to Organization C. Additionally, Organization C can search service offerings that match its interest, such as offerings considering the historical data of performed tasks.

The ecosystem's organizations have negotiated a data sharing contract at the general level and have an agreement of which data is shared and with whom. They have also agreed that shared data can only be utilized to achieve the shared goal. Therefore, the organizations do not need to negotiate a contract each time when data is shared. The data provider just defines the identities to whom a service offering is intended and at-

taches additional usage policies to the offering, if needed. For example, when Organization B requires service offering of mobile machine location data from Organization A, the access to the data can be automatically granted and Organization B will see the service in its credential manager. The example contract in Figure 6.5 follows an example contract agreement of IDS [66], because Gaia-X has not defined clear specifications regarding data sharing contracts.

```
{
  "@context": [
    "http://www.w3.org/ns/odrl.jsonld",
    {
      "ids": "https://w3id.org/idsa/core/",
      "idsc": "https://w3id.org/idsa/code/"
    }
  ],
  "@type": "ids:ContractAgreement",
  "@id": "https://w3id.org/contract/policy-access",
  "ids:provider": "did:example:organizationA",
  "ids:consumer": "did:example:organizationB",
  "ids:profile": "http://example.org/odrl:profile",
  "ids:permission": [{
    "ids:action": "use",
    "ids:target": "did:example:OrganizationA/machines/all/locations",
    "ids:constraint": [{
      "ids:leftOperand": "purpose",
      "ids:operator": "eq",
      "ids:rightOperand": [{
        "@type": "xsd:anyURI",
        "@value": "http://example.org/purpose/OperationOptimization"
      }]
    }]
  }]
}
```

**Figure 6.5.** Example data sharing contract.

When data is exchanged or access provided, organizations use a logging service to log transactions. Firstly, Organization A sends notification to the logging service. The notification indicates that data exchange has been started. The logging service responds to the notification with a corresponding notification identifier. Organization A sends the notification identifier to Organization B with the access instructions to the data. In the context of this example, the exchanged data is online data which has continuous updates so it can not be exchanged as a single data set. In this kind of cases data is exchanged through other channels, such as providing access to the message broker. When Organization B gains the access to the data, it sends notification to the logging service. The notification must contain the notification identifier to form a log entry with the previous notifications.

Organization A utilizes authentication and authorization service to provide access to its resources. The service is also used with the federation services of the ecosystem. When Organization B accesses to the data, it does not need different identity than in federa-

tion services. Through the access to the message broker, Organization B promotes the ecosystem to reach its goal by processing the data in accordance with agreements.

This case scenario is quite simple, and the full potential value of federation services might be achieved in larger and more complex data ecosystems. However, by adopting the compliance services, that are mandatory for all Gaia-X compliant ecosystems, such small ecosystems create a sustainable foundation for a potentially expanding ecosystem. Furthermore, following the Gaia-X principles makes it possible to share data across individual ecosystems in the interoperable way.

## 7. DISCUSSION

The thesis summarized the current data sharing challenges based on the literature review. The review showed that in addition to challenges related to a technical implementation, there are challenges related to human behavior and attitudes and to conflicting business needs. The challenges decrease the willingness of organizations to share their data and join in data ecosystems. Without data sharing, the potential value of data is not utilized, and organizations may miss their opportunities to grow their businesses and create new innovations. However, organizations are aware of the value of their data resources which causes a lower level of openness and higher level of protection in data sharing [1].

The literature review was conducted utilizing two databases, which gives a narrow sample of articles discussing data sharing challenges. However, it can be assumed that the same challenges arise in different databases, so including several databases in the review would not have brought significant value to it or changed the results. On the other hand, the search statements could still have been iterated more making the search results more accurate and reducing manual filtering of articles. Furthermore, the ideal would have been to use the same statements in both data bases. The research statement in ACM needed to be more specific to obtain a manually processable number of search results. However, the same statement would have been too narrow for IEEE and given too limited results.

In this thesis, the suitability of Gaia-X infrastructure as a solution to the identified challenges is analyzed. The analysis showed that Gaia-X has a wide range of components which aim to facilitate data sharing between organizations. However, the components are more focused on to solve challenges such as interoperability and data heterogeneity, which improves data availability, than data exchange between organizations. The actual data exchange process is responsibility of acting parties, and they must negotiate what are the channels of data transfer as well as the usage policies. From the perspective of sovereignty, which is one of the data ecosystem principles, Gaia-X architecture is deficient when it comes to the enforcement of data resource specific rules and contracts and gathering evidence of their compliance. At this point of development of Gaia-X architecture, the components support only the technical enforcement of general rules defined by Gaia-X Association.

Furthermore, the distributed nature of Gaia-X architecture gives freedom of choice to data owners when it comes to data storage. Because a third-party hosted central data storage in data sharing ecosystems concerns data owners [41], distribution feels more secure solution. With a distributed architecture, data owners can be sure that their data is stored the way they want, and they are more aware of its location. This reinforces data owners' sense of control over their data and increases willingness to join data ecosystems and share data with others. However, distribution increases the responsibility of data owners from the viewpoint of the security and quality of data, because mechanisms, such as access control and data lifecycle management remain the responsibility of the owner.

The ongoing development of Gaia-X architecture has made the research process of this thesis difficult. The continuous changes of the architecture rapidly age the results of the research. It is difficult for someone from outside of the developer team to keep up with the changes, and the information about the changes comes always late because it takes time to write documents about them. On the other hand, the development of Gaia-X may fulfill the current gaps of the architecture and address the challenges better in the future.

Hopefully, the results of this thesis will help infrastructure developers identify the need and requirements of an infrastructure and its current shortcomings. Additionally, the results support organizations to perceive their own location in the field of data sharing and encourage to participate in data ecosystems. The thesis provides a good basis for organizations to identify and solve their own challenges and gives an easily approachable overview to the offerings of Gaia-X. Raising awareness of Gaia-X and data sharing benefits will hopefully help the development of the infrastructure, thus creating a more efficient and value-generating operation from data sharing for organizations.

## 8. CONCLUSIONS

In this thesis, the challenges of inter-organizational data sharing in multi-actor environments are identified. The challenges are divided into three main categories: business challenges, data governance and interoperability. More detailed challenges are presented in Table 5.1. They define the need for an infrastructure supporting fair and secure data sharing between organizations and enabling achievement of maximum benefits from data. Considering the challenge categories, business challenges are in the key role when organizations are encouraged to be involved in data sharing. However, most of the identified challenges are related to data governance. Among the data governance challenges, there are several existing solutions to security-related challenges, while data quality and ownership issues are more challenging to solve. Lastly, the interoperability challenges have the greatest impact on the efficiency of data sharing. They are the most critical challenges from the viewpoint of the technical implementation of data sharing.

Furthermore, the suitability of Gaia-X data infrastructure under development for solving the challenges is analyzed and an example case scenario of data sharing utilizing Gaia-X is described. The analysis showed that Gaia-X is more focused on making heterogeneous data available and comparable in data ecosystems and not so much facilitating the actual data exchange process. Additionally, Gaia-X architecture components support data owners to retain the control over their data by enabling automatic and semi-automatic data contract negotiation processes and by providing means to the data owner to define usage policies for their data. However, the biggest weakness of infrastructure compared to the challenges is lack of technical enforcement of usage policies. Data users can violate the usage policies without getting caught because data owners can not monitor the compliance of contracts at the required level.

In the future, when Gaia-X architecture has developed further, the example data sharing case scenario could be implemented. With the implementation, Gaia-X architecture components can be reanalyzed more detail that providers more reliable results about their suitability to address the data sharing challenges between organizations. Furthermore, the implementation will demonstrate the gaps of Gaia-X in relation to the challenges and provide a concrete example for organizations how to utilize Gaia-X architecture and how to adopt the components in their own data ecosystems.

## REFERENCES

- [1] J. Gelhaar and B. Otto, “Challenges in the emergence of data ecosystems.”, in *Twenty-Third Pacific Asia Conference on Information Systems*, 2020, p. 175.
- [2] F. De Prieëlle, M. De Reuver, and J. Rezaei, “The role of ecosystem data governance in adoption of data platforms by internet-of-things data providers: Case of dutch horticulture industry”, *IEEE Transactions on Engineering Management*, vol. 69, no. 4, pp. 940–950, 2020. DOI: 10.1109/TEM.2020.2966024.
- [3] T. Nokkala, H. Salmela, and J. Toivonen, “Data governance in digital platforms”, in *25th Americas Conference on Information Systems (AMCIS 2019)*, 2019.
- [4] S. Geisler, M.-E. Vidal, C. Cappiello, *et al.*, “Knowledge-driven data ecosystems toward data transparency”, *ACM Journal of Data and Information Quality (JDIQ)*, vol. 14, no. 1, pp. 1–12, 2021.
- [5] L. Dodds and P. Wells, *Data Infrastructure*. The Open Data Institute, 2019. DOI: 10.5281/zenodo.2677811.
- [6] *Gaia-x architecture document 22.04 release*, Gaia-X European Association for Data and Cloud AISBL, 2022.
- [7] A. Sneddon, *Autonomy*. New York: Bloomsbury Publishing Plc, 2013, pp. 11–46.
- [8] C. Fracassi and W. Magnuson, “Data autonomy”, *Vand. L. Rev.*, vol. 74, pp. 327–383, 2021.
- [9] L. Glanville, “Sovereignty”, in *The Oxford Handbook of the Responsibility to Protect*, Oxford University Press, Jun. 2016. DOI: 10.1093/oxfordhb/9780198753841.013.9.
- [10] P. Hummel, M. Braun, M. Tretter, and P. Dabrock, “Data sovereignty: A review”, *Big Data & Society*, vol. 8, no. 1, p. 2 053 951 720 982 012, 2021.
- [11] M. I. S. Oliveira, L. E. R. Oliveira, M. G. R. Batista, and B. F. Lóscio, “Towards a meta-model for data ecosystems”, in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 2018, pp. 1–10.
- [12] C. Cappiello, A. Gal, M. Jarke, and J. Rehof, “Data ecosystems: Sovereign data exchange among organizations (dagstuhl seminar 19391)”, in *Dagstuhl Reports*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, vol. 9, 2020.
- [13] M. I. S. Oliveira and B. F. Lóscio, “What is a data ecosystem?”, in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, 2018, pp. 1–9.

- [14] E. Chang and M. West, "Digital ecosystems a next generation of the collaborative environment", *Information Integration and Web-based Application and Services*, vol. 214, pp. 3–24, 2006.
- [15] C. Ballard, N. Davies, M. Gavazzi, M. Lurie, and J. Stephani, *IBM Informix: Integration through data federation*. IBM, International Technical Support Organization, 2003.
- [16] R. Van Der Lans, *Data Virtualization for business intelligence systems*. Elsevier, 2012.
- [17] B. Otto, "The evolution of data spaces", in *Designing Data Spaces*, Springer, Cham, 2022, pp. 3–15.
- [18] A. Halevy, M. Franklin, and D. Maier, "Principles of dataspace systems", in *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2006, pp. 1–9.
- [19] J. Reichental, "Data governance for dummies", in John Wiley & Sons, 2023, ch. Defining Data Governance. [Online]. Available: <https://learning.oreilly.com/library/view/data-governance-for/9781119906773/c01.xhtml>.
- [20] T. Van den Broek and A. F. Van Veenstra, "Modes of governance in inter-organizational data collaborations: Complete research", *Proc. 23rd Eur. Conf. Inf. Syst.*, 2015.
- [21] I. Alhassan, D. Sammon, and M. Daly, "Data governance activities: A comparison between scientific and practice-oriented literature", *Journal of enterprise information management*, vol. 31, no. 2, pp. 300–316, 2018.
- [22] S. U. Lee, L. Zhu, and R. Jeffery, "Data governance for platform ecosystems: Critical factors and the state of practice", *Proc. 21st Pac. Asia Conf. Inf. Syst.*, pp. 1–12, May 2017.
- [23] D. Puspasari, A. N. Hadiyanto, and S. Setiawan, "Inter-organizational data sharing: What issues should be considered?", in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, IEEE, 2021, pp. 181–186.
- [24] L. Mui, "Computational models of trust and reputation: Agents, evolutionary games, and social networks", Ph.D. dissertation, Massachusetts Institute of Technology, 2002.
- [25] D. Moreland, S. Nepal, H. Hwang, and J. Zic, "A snapshot of trusted personal devices applicable to transaction processing", *Personal and Ubiquitous Computing*, vol. 14, no. 4, pp. 347–361, 2010.
- [26] J. Byabazaire, G. O'Hare, and D. Delaney, "Data quality and trust: Review of challenges and opportunities for data sharing in iot", *Electronics*, vol. 9, no. 12, p. 2083, 2020.

- [27] A. Strunk and C. Lange, *Self-description of resources, service offerings and participants within gaia-x ecosystems*, Gaia-X European Association for Data and Cloud AISBL, 2022.
- [28] *Gaia-x ecosystem kickstarter*, Gaia-X European Association for Data and Cloud AISBL, 2022.
- [29] *Gaia-x trust framework 22.04 release*, Gaia-X European Association for Data and Cloud AISBL, 2022.
- [30] D. Burnett, M. Sporny, D. Longley, B. Zundel, G. Noble, and K. D. Hartog, “Verifiable credentials data model v1.1”, W3C, W3C Recommendation, Mar. 2022, <https://www.w3.org/TR/vc-data-model/>.
- [31] B. Maier and N. Pohlmann, *Gaia-x secure and trustworthy ecosystems with self sovereign identity*, Gaia-X European Association for Data and Cloud AISBL, 2022.
- [32] *Gaia-x labelling framework*, Gaia-X European Association for Data and Cloud AISBL, 2021.
- [33] B. Guo, X. Deng, Q. Guan, and J. Tian, “A competitiveness-driven and secure incentive mechanism for competitive organizations data sharing: A contract theoretic approach”, in *2018 1st IEEE International Conference on Hot Information-Centric Networking (HotICN)*, IEEE, 2018, pp. 30–36.
- [34] J. Chen and Y. Xue, “Bootstrapping a blockchain based ecosystem for big data exchange”, in *2017 IEEE international congress on big data (bigdata congress)*, IEEE, 2017, pp. 460–463.
- [35] M. I. S. Oliveira and B. F. Lóscio, “Louvre: A framework for metadata curation in data ecosystem”, in *Proceedings of the XV Brazilian Symposium on Information Systems*, 2019, pp. 1–8.
- [36] E. Karafilli and E. C. Lupu, “Enabling data sharing in contextual environments: Policy representation and analysis”, in *Proceedings of the 22Nd ACM on Symposium on Access Control Models and Technologies*, 2017, pp. 231–238.
- [37] J. Zrenner, F. O. Möller, C. Jung, A. Eitel, and B. Otto, “Usage control architecture options for data sovereignty in business ecosystems”, *Journal of Enterprise Information Management*, 2019.
- [38] A. Munoz-Arcenales, S. López-Pernas, A. Pozo, Á. Alonso, J. Salvachúa, and G. Huecas, “An architecture for providing data usage and access control in data sharing ecosystems”, *Procedia Computer Science*, vol. 160, pp. 590–597, 2019.
- [39] J. Moreno, E. B. Fernandez, E. Fernandez-Medina, and M. A. Serrano, “Blockbd: A security pattern to incorporate blockchain in big data ecosystems”, in *Proceedings of the 24th European Conference on Pattern Languages of Programs*, 2019, pp. 1–8.
- [40] H. Qarawlus, M. Hellmeier, J. Pieperbeck, R. Quensel, S. Biehs, and M. Peschke, “Sovereign data exchange in cloud-connected IoT using international data spaces”, in *2021 IEEE Cloud Summit (Cloud Summit)*, IEEE, 2021, pp. 13–18.

- [41] A. G. M. Alzahrani, A. Alenezi, A. Mershed, H. Atlam, F. Mousa, and G. Wills, "A framework for data sharing between healthcare providers using blockchain", 2020.
- [42] M. Ali, R. Dhamotharan, E. Khan, *et al.*, "Sedasc: Secure data sharing in clouds", *IEEE Systems Journal*, vol. 11, no. 2, pp. 395–404, 2015.
- [43] L. Rao, Q. Xie, and H. Zhao, "Data sharing for multiple groups with privacy preservation in the cloud", in *2020 International Conference on Internet of Things and Intelligent Applications (ITIA)*, IEEE, 2020, pp. 1–5.
- [44] M. Anisetti, C. A. Ardagna, C. Braghin, E. Damiani, A. Polimeno, and A. Balestrucci, "Dynamic and scalable enforcement of access control policies for big data", in *Proceedings of the 13th International Conference on Management of Digital EcoSystems*, 2021, pp. 71–78.
- [45] A. Michalas and N. Weingarten, "Healthshare: Using attribute-based encryption for secure data sharing between multiple clouds", in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2017, pp. 811–815.
- [46] S. Wang and J. Liu, "Blockchain based secure data sharing model", in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, 2021, pp. 464–469.
- [47] M. A. Will, R. K. Ko, and S. J. Schlickmann, "Anonymous data sharing between organisations with elliptic curve cryptography", in *2017 IEEE Trustcom/BigDataSE/ICSS*, IEEE, 2017, pp. 1024–1031.
- [48] Y. Qian, Z. Liu, J. Yang, and Q. Wang, "A method of exchanging data in smart city by blockchain", in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, IEEE, 2018, pp. 1344–1349.
- [49] V. Janev, M. E. Vidal, K. Endris, and D. Pujic, "Managing knowledge in energy data spaces", in *Companion Proceedings of the Web Conference 2021*, 2021, pp. 7–15.
- [50] A. Shahaab, I. Khan, R. Maude, and C. Hewage, "A hybrid blockchain implementation to ensure data integrity and interoperability for public service organisations", in *2021 IEEE International Conference on Blockchain (Blockchain)*, IEEE, 2021, pp. 295–305.
- [51] J. Axelsson, "Experiences of using linked data and ontologies for operational data sharing in systems-of-systems", in *2019 IEEE International Systems Conference (SysCon)*, IEEE, 2019, pp. 1–8. DOI: 10.1109/SYSCON.2019.8836909.
- [52] *Policy rules document 22.04 release*, Gaia-X European Association for Data and Cloud AISBL, 2022.
- [53] *Software requirements specification for gaia-x federation services federated catalogue core catalogue features*, GAIA-X European Association for Data and Cloud AISBL, 2021.

- [54] *Software requirements specification for gaia-x federation services sovereign data exchange data contract service*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [55] *Software requirements specification for gaia-x federation services sovereign data exchange data exchange logging service*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [56] *Software requirements specification for gaia-x federation services authentication/ authorization*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [57] *Software requirements specification for gaia-x federation services organization credential manager*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [58] *Software requirements specification for gaia-x federation services personal credential manager*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [59] *Software requirements specification for gaia-x federation services trust services api*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [60] *Software requirements specification for gaia-x federation services compliance continuous automated monitoring*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [61] *Software requirements specification for gaia-x federation services compliance onboarding and accreditation*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [62] *Software requirements specification for gaia-x federation services compliance notarization api*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [63] *Software requirements specification for gaia-x federation services integration and portal portal*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [64] *Software requirements specification for gaia-x federation services integration and portal orchestration*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [65] M. Binzer, C. Bormann, P. Kowalik, *et al.*, *Gxfs - idm and trust architecture overview*, GAIA-X European Association for Data and Cloud AISBL, 2021.
- [66] S. Steinbuss, A. Eitel, C. Jung, *et al.*, "Usage control in the international data spaces", International Data Spaces Association, Position paper, 2021, <https://doi.org/10.5281/zenodo.5675884>.