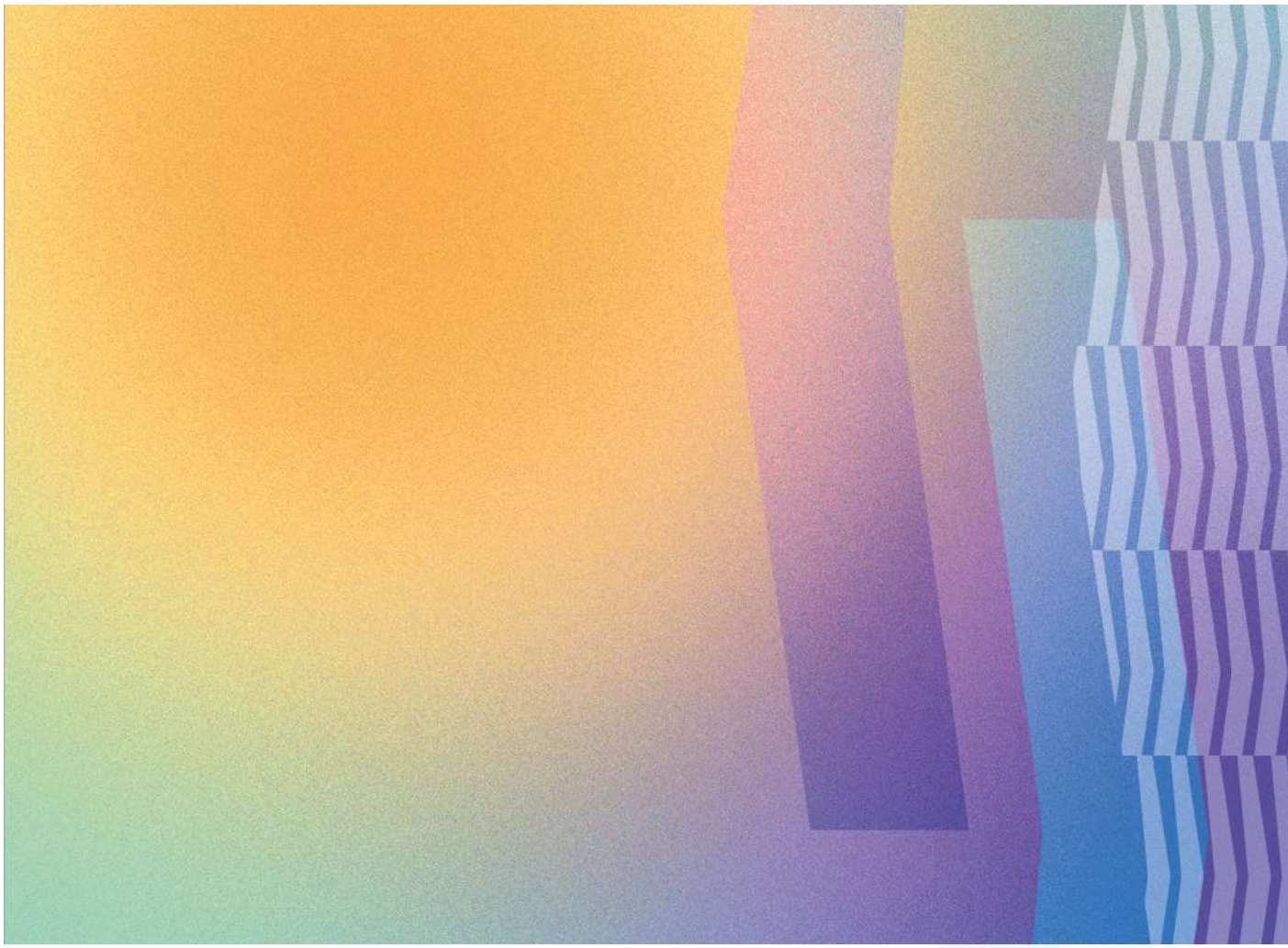
Technology Program Edge Data Economy
German Federal Ministry for Economic Affairs and Climate Action



DATA INFRASTRUCTURES FOR DATA ECOSYSTEMS

A decision aid for designing the system architecture of data infrastructures for data sharing

Imprint

Author

Nils Jahnke

Fraunhofer Institute for Software and Systems Engineering ISST
Dortmund

Publisher

Peter Gabriel
Dr. Nicole Wittenbrink

Begleitforschung Edge Datenwirtschaft
Institute for Innovation and Technik (iit)
within the VDI / VDE Innovation + Technik GmbH
Berlin

Date

April 2025

Layout

PRpetuum GmbH

This guide was commissioned by the German Federal Ministry for Economic Affairs and Climate Action as part of the coordination and support action (Begleitforschung) for the technology program Edge Data Economy (Edge Datenwirtschaft)

This English version has been produced by a machine translation tool with minimal human intervention. We do not guarantee the accuracy of this translated version and accept no liability for any errors. Please visit https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/EDGE-Datenwirtschaft/20250429_Entscheidungshilfe_Dateninfrastrukturen.html for the original publication in German language.

Executive Summary

With the help of innovative data-driven applications, companies are generating extensive economic added value. At the same time, they address some of the major challenges of our time, such as the shortage of skilled workers or the conservation of natural resources. To leverage the potential of data-driven applications, companies are often dependent on data provisioned by external parties. In data sharing constellations, data providers make data available for external parties and in return receive compensation for sharing their data. Compensation can include the increased access to information, compliance with regulatory requirements, or monetary rewards. If the data sharing relationships between the actors have a network constellation, data ecosystems emerge. In data ecosystems, companies collaborate and compete around data as a central resource.

To implement data-driven applications, the data must be available at the right time and in the required quality. This is the task of a data infrastructure. A demand-driven data infrastructure is the foundation for the efficiency and performance of the applications that rely on it, and it supports the scalability of the business models that are being implemented.

This decision aid supports companies and industry consortia in selecting a suitable system architecture for the data sharing data infrastructure based on the demands of the environmental setting and the planned use cases. In particular, the decision aid addresses architecture teams and decision-makers who are involved in the new development or further development of the data infrastructure. The decision aid compares centralized, federated, decentralized and distributed architectures. Based on individually derived decision criteria, the most suitable data infrastructure design option is identified as a result.

Based on established methods from the field of enterprise architecture, the decision aid explains four phases of the decision-making process:

1) Environmental analysis: Identification, description and evaluation of influences and interests in the design of the data infrastructure from the environment
2) Business model analysis: Derivation and characterization of relevant data, data flows and systems from the business model
3) Requirements definition: Identification and prioritization of relevant requirements for the data infrastructure as a foundation for decision-making
4) Preliminary design: Development and evaluation of possible design options and identification of the best possible alternative

The core of this decision aid is a catalogue of criteria for evaluating data infrastructure design options. This catalog of criteria explains the decision factors for data infrastructures, their relevance for specific data sharing constellations and possible interdependencies between the factors. Finally, it also provides an assessment of the extent to which the decision factors are fulfilled by the selected design options.

Contents

1 Introduction

Innovative, data-driven applications enable improvements to processes, products and services and provide economic and social value-add. A study by MIT shows that companies in which processes are data-based, automated and take place in real time achieve more than 50% higher growth compared to their less digitized competitors (Weill et al., 2024). Examples of data-driven applications include the measurement and optimization of energy consumption in production processes or the intelligent integration of decentralized systems to stabilize the power grid. Due to the high complexity of these applications, value is usually not created from data by one actor alone. Instead, several actors join forces to provide services collaboratively.

The central component of this collaboration is the sharing of data. So-called data ecosystems are created through the data exchange relationships between the actors. Data ecosystems are a new, complex form of organization in which autonomous actors with common and conflicting interests collaborate with each other. Data must be made actionable for the implementation of value-adding applications across the boundaries of individual locations and beyond the spheres of influence of individual actors.

Depending on the application, there are different requirements for data sharing. For example, some applications require data to be made available with particularly low latency times, whereas other applications place importance on the protection of personal data during data provisioning.

The basis of data sharing is a reliable and trustworthy data infrastructure that is available to the data exchange partners in the long term and enables the development of a variety of novel applications. Data infrastructures provide technical services, common standards, rules and guidelines as well as trust mechanisms (Dodds & Wells, 2019). Concrete services of data infrastructures are, for example, a participant registry, an identity management system or catalogs for describing and discovering available data and services.

The data infrastructure can be designed and implemented either by a single company or in a collaborative process involving several participating organizations. In order to develop a solution that is suitable for the application purposes, several questions from the areas of planning, design, implementation and maintenance of data infrastructures must be answered (Otto et al., 2022). The design of a data infrastructure must balance the requirements of data providers, data donors and the general public in order to convince as many stakeholders as possible to share their data. An important question is therefore whether data infrastructure services should be designed in a centralized or decentralized form. However, given the novelty of the topic, it is often difficult for those responsible to make an informed decision regarding the design of data infrastructure services. The cross-organizational sharing of data in particular gives rise to complex requirements and framework conditions that deviate from conventional internal company design and development processes. These include, for example, the need to find a balance between the individual and collective interests of the stakeholders.

This document addresses the described challenge. Companies that want to implement an application based on shared data and in collaboration with other actors are provided with a decision aid for the design of the underlying data infrastructure services. The aim of the decision aid is to identify the best possible design option for the data infrastructure based on the external macro-environmental factors and the applications to be implemented. This decision aid takes into

account the concerns of small and medium-sized enterprises and smaller data infrastructure initiatives. Specifically, the decision aid is aimed at system architects and decision-makers who are involved in the implementation of collaborative data-based applications in their companies.

The decision aid first introduces the topic of data infrastructures by describing relevant data infrastructure services and their design options. It then presents a procedure model for selecting a suitable design option for data infrastructure services for data sharing use cases in the data economy. The procedure model includes the four phases "environmental analysis", "business model analysis", "requirements definition" and "preliminary design" and is based on established concepts from the field of enterprise architecture. The central element of this decision aid is a catalog of criteria for evaluating the design options for data infrastructure services. The criteria catalog describes the key decision factors for data infrastructures, their significance for specific data sharing scenarios and explains the extent to which individual design options can support the decision factors.

This decision aid is structured as follows: Section 2 explains the basic services of data infrastructures and provides an overview of possible design options. Section 3 presents a procedure model for evaluating the design options of data infrastructure services. The associated catalog of criteria for data infrastructure services is presented in Section 4. The concluding outlook recommends further literature on individual reference architectures as a starting point for technical implementation (Section 5).

The collaboration with the experts from the early adopter projects of the technology program Edge Data Economy has made a significant contribution to integrating practical findings in the development of data infrastructures into this decision aid and to comprehensively addressing the requirements of different business areas and company sizes. The author would therefore like to take this opportunity to thank all of the interviewees once again for their support in the creation of this decision aid:

- Lars Nolting, Tennet TSO GmbH, project DEER
- Florian Remark, Strategion GmbH, project SECAI
- Bruno Ristok, C&S Computer und Software GmbH, project Careful-Edge-X
- Heinrich Traphöner, Empolis GmbH, project EASY

The responsibility for the content of this decision aid lies exclusively with the authors.

The decision aid was produced as part of the coordination and support action (Begleitforschung) for the technology program Edge Data Economy of the Federal Ministry for Economic Affairs and Climate Action (BMWK). The program comprises ten projects that are developing new forms of edge cloud systems for important sectors of the German economy.

2 Essential Services and Design Options for Data Infrastructures

2.1 Components of data infrastructures

Data infrastructures usually comprise three types of building blocks to enable data sharing applications: Technical services enable data exchange and data integration between physically separate systems and ensure security against unauthorized access or data loss. Standards facilitate the harmonization of data and processes, thereby promoting organizational, semantic, and technical interoperability. Rules and guidelines define a common framework for the use of the data infrastructure and the design of data sharing processes.

Current developments in the field of data infrastructures focus on the development of data infrastructure services (Demchenko et al., 2014; Otto & Burmann, 2021). Due to the high potential of data infrastructures for generating data-driven value, both the European Commission and the German government have provided public funding in recent years to promote the conceptualization and development of data infrastructure services in various initiatives. This has resulted in different private-public associations such as the International Data Spaces Association (IDSA)[1] or Gaia-X[2], public initiatives such as SIMPL[3] or 8ra[4], open-source development work under the umbrella of the Eclipse Dataspaces Working Group[5] or CKAN[6] as well as other private data infrastructure initiatives. The content of the initiatives differs in terms of their focus areas and the services provided (e.g. conceptualization versus implementation). However, common data infrastructure services can be identified across the initiatives. However, the services differ in terms of their respective architectural approaches and technology stacks. These common services include a participant registry, components for identity and access management, and a data and service catalog. In addition, further individual services are conceivable for each initiative, for example to ensure data sovereignty or data integrity.

Participant registry

A participant registry operationalizes the management of the actors involved in data sharing as well as the accessible endpoints, such as smart services or smart devices. It lists basic information about the participants and makes this information findable on a global level. An entry in the participant registry is usually made after the actor has gone through an automated or manual onboarding process. A listing in the participant registry thus creates a fundamental basis of trust between the data exchange partners and is a necessary condition for the use of further data infrastructure services and for participant interaction. The usual requirements for a participant registry are the integrity, availability and inalterability of the stored data.

[1] https://internationaldataspaces.org/
[2] https://gaia-x.eu/
[3] https://digital-strategy.ec.europa.eu/en/policies/simpl
[4] https://www.8ra.com/
[5] https://dataspace.eclipse.org/projects/
[6] https://ckan.org/

Definition: Actors and participants in the context of data infrastructures

In this document, an actor is defined as any organization or natural person with a connection to the realization of added value based on shared data. In the area of data sharing, actors are usually legal entities. Once an organization has successfully completed the necessary onboarding processes, it becomes a participant in a data infrastructure. The participants in a data infrastructure can take on one or more roles depending on the use case. As data providers, they provide available data via the data infrastructure and receive monetary or non-monetary compensation. As data consumers, they consume the data of the data providers via the data infrastructure in order to realize added value. In return, they provide compensation for other participants. In addition, participants can take on the role of operators of data infrastructure services and thus ensure the functionality of the data infrastructure. Monetary incentives are usually created for the operation of data infrastructure services.

Identity and access management

An identity is understood as the sum of properties of an actor that allows for its identification within a specific context. An identity includes a unique identifier and additional attributes . Depending on the type of actor, these additional attributes can be of different types. For natural persons, for example, name, address, age or organizational affiliation are typical attributes. For organizations, for example, organization name, legal form, legal address and certifications received are conceivable. The identity of an actor is required to determine which applications and services may be used, and which privileges the actor has during their use. Possible attributes for uniquely identifying an actor must therefore be defined for the respective context.

Important functions of identity and access management are the authentication and authorization of actors. Authentication ensures that the identity is the actor it claims to be. Authorization involves determining which permissions an actor receives based on its attributes. Trust into the correctness of the attributes is therefore a prerequisite. Thus, in some cases, the identity management component is also referred to as "identity & trust" to emphasize its important role in establishing trust between the actors. Identity and access management is essential to manage access to other data infrastructure services, to allow for the retrieval and provisioning of data, and to enable the use of intelligent services.

Data and services catalog

The catalog service lists information about data and services offered by participants in a data infrastructure using metadata. Metadata is often described as "data about data" and comprises structured information about technical properties, the relation to organizational concepts, and the origin and use of the data. Metadata is therefore usually less sensitive than the actual data used for value creation.

The metadata describing the data and services offered are usually listed in the catalog by the data providers. Participants can view the metadata within a catalog if they meet the access guidelines defined by the data provider and implemented by the catalog operator. Catalogs therefore fulfil important functions along the data value chain. Data catalogs bring together supply and demand in data sharing through the possibilities of inventorying and discovering offers. In addition, available data and services can be analyzed and evaluated before they are used. Based on the information curated in the catalog, a usage request or retrieval of the offered data or services can also take place.

Additional services

Depending on the context and application area of the data infrastructure, further data infrastructure services are conceivable. These include directory services for standardized semantic data models. Semantic data models define which concepts are described by data, how these concepts are grouped and classified and what relationship they have to each other. Semantic data models therefore contribute to a better understanding of the data.

Data connectors can be used as standardized endpoints to the systems of data providers and data consumers as additional services. Data connectors are able to receive and provide data and metadata, check terms of use during data exchange and, in some cases, enforce them with technical means. Data connectors thus abstract the technical complexity of the data exchange between the participants in a data infrastructure.

Orchestrator services represent another conceivable category of additional services. These serve as coordinators to distribute and aggregate tasks across different data infrastructure services. Orchestrator services therefore help to seamlessly execute workflows across the various services of a data infrastructure with as little manual interaction as possible.

2.2 Design options for data infrastructures

A fundamental decision in the design of a system is the definition of the system architecture. The system architecture describes the arrangement of functional elements, the realization of the functional elements through physical components and the interfaces between the components. For data infrastructure services, amongst others, the following questions need to be answered: "How often do I need a service?", "Who provides this service?" and "What is the relationship between these services?". Answering these questions requires consideration of the objectives of the data infrastructure and the environmental constraints.

The data infrastructure services described above can be realized through various design options (Schleimer et al., 2023). A design option is a specific possibility for achieving the desired functionality. The following paragraph characterizes four basic design options for data infrastructure services.

Fully centralized design

A centralized design of a data infrastructure service implies that it is provided exactly once for the entire data infrastructure (see Figure 1). The distinctiveness of the service is defined by the perspective of the data infrastructure users. Functions such as data storage can be invisibly replicated to enhance resilience and recoverability. The delivery of this service may be conducted by a dedicated neutral party or, in exceptional instances, by the entity with the highest stake in the operational integrity of the data infrastructure. To populate the data infrastructure service, data or metadata are transferred to the data infrastructure service by the involved actors. The operator of the data infrastructure service is trusted to use the data in the interests of the data provisioning actors. This is particularly relevant for services that mediate data between the participants in a data infrastructure. For example, a catalog service should act neutrally, and not present data offers to users on a preferential basis.

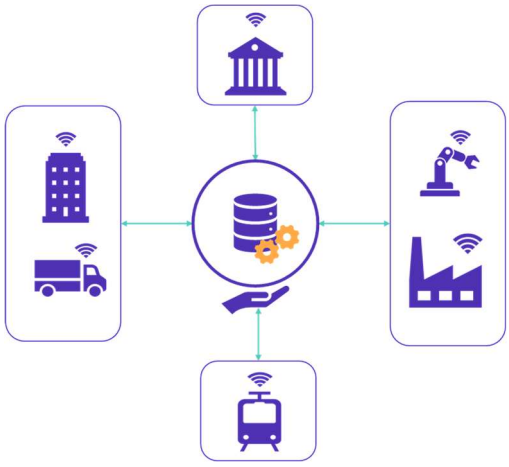


Figure 1: Schematic representation of centralized data infrastructure services

Example: A centralized data and service catalog

A single operator provides the data and service catalog. The actors transfer their metadata to the central catalog service. This makes the data and services offered by the actors transparent and searchable for other actors. The catalog operator manages the visibility and access to the metadata. In the event of requests for updates or deletions of data or service offerings, the central service must be contacted by the actors. An example of a centralized data catalog that is implemented once for a data infrastructure is the "Metadata Broker" in the IDSA reference architecture model for data spaces[7]. The SIMPL initiative also envisions a centralized catalog[8].

Fully decentralized design

The fully decentralized design of data infrastructure services (see Figure 2) implies that a service is implemented multiple times for the same data infrastructure without services of the same type being able to communicate with each other. This can be the case, for example, if there are different requirements and expectations about a service or if actors wish to refrain from transferring data to a centralized service. As multiple executions of the same service are present, different provision options are also possible. For example, operators of the data infrastructure can provide the data infrastructure service as a dedicated service for one participant or as a shared service for several participants. In addition, individual participants can also manage the service themselves for their smart devices or services to decouple themselves from dependencies (see middle part of Figure 2). A fully decentralized design means that potentially all data infrastructure services of the same type must be queried to obtain a complete picture of the data infrastructure. It can therefore be expected that this constellation will lead to a competitive situation between services of the same type, which usually results in service consolidation (Bader et al., 2023).

[7] https://github.com/International-Data-Spaces-Association/metadata-broker-open-core
[8] https://simpl-programme.ec.europa.eu/book-page/simpl-open-architecture

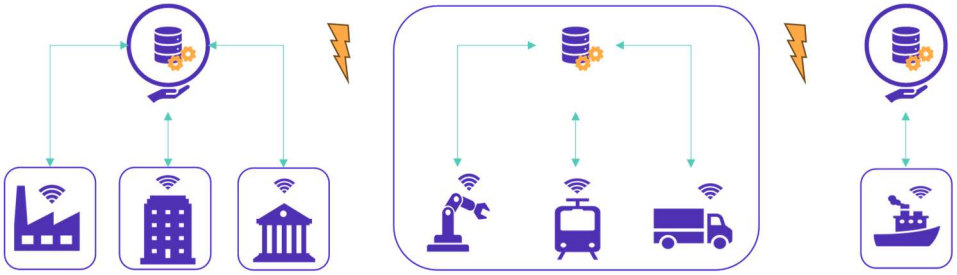


Figure 2: Schematic representation of decentralized data infrastructure services

Example: A fully decentralized data and service catalog

Several unconnected catalog services are available. Before registering its data and service offerings, an actor must therefore decide whether it wants to join an existing catalog service or provide its own catalog (possibly by contracting a service provider). The actor transfers its metadata to the relevant catalog or uses its own catalog. Potential data consumers must either know where potentially relevant data offerings can be found or search through many available catalogs. The IDSA Metadata Broker and the SIMPL catalog can also serve as examples in the case of a decentralized design of the catalog service. Both initiatives potentially provide for multiple, parallel uses of the catalogue service in their architectural documents.

Federated design

The federated implementation of a data infrastructure service (see Figure 3) basically consists of a large number of independent data infrastructure services of the same type that are interoperable with each other and cooperate by exchanging (meta) data. Interoperability is ensured by common standards and protocols that are supported by the individual services. Each service can decide autonomously which metadata it shares with which other service. It further determines for itself how the metadata available should be presented to the actors. Accordingly, each service presents an individual view of the data infrastructure. It is also possible for individual services to flexibly join or leave the federated services system, if required. As in the fully decentralized design, the individual data infrastructure services of the same type can be operated by the actors themselves or by a service provider contracted by the actors. The highest degree of autonomy and self-determination is achieved if the actors operate the service themselves.

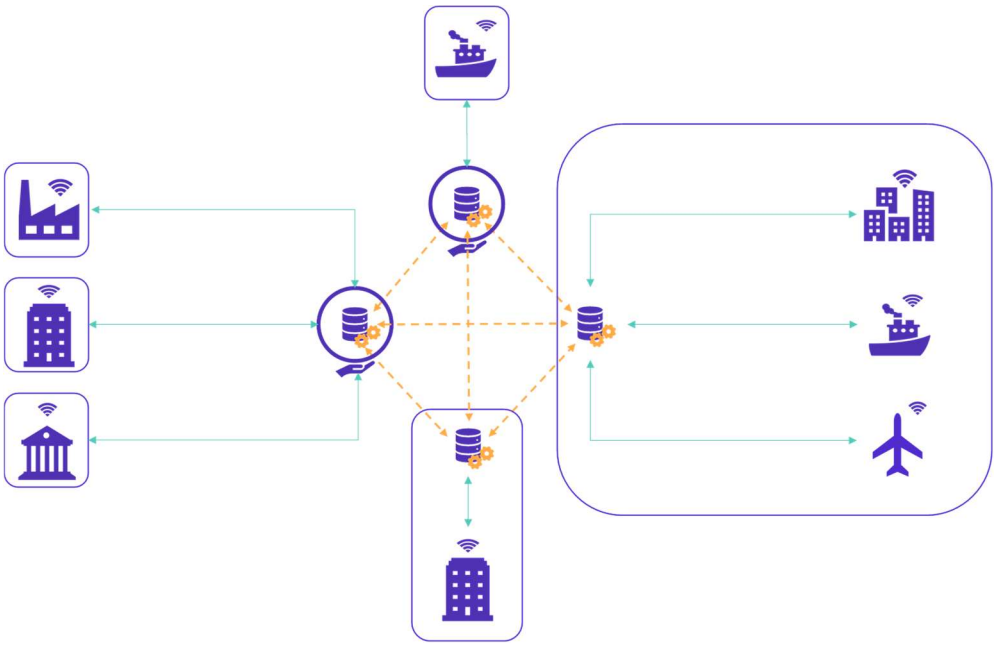


Figure 3: Schematic representation of federated data infrastructure services

Example: A federated data and service catalog:

An actor sets up its own catalog service to participate in the federated catalog system. The catalog must follow the defined standards and protocols. Ideally, an open reference implementation of the catalog is available for easy implementation. In the first step, the actor describes the data and service offerings in its own catalog. In the next step, depending on the synchronization method, the offers of other catalogs are requested. If these catalogs give their consent, their offers can be retrieved and displayed in the actor's own catalog. Conversely, the other catalogs can also list the actor's offers, provided they meet the predefined criteria. Each of the federated catalogs receives an individual view of the available data and service offers by replicating the (meta) data received. To initiate a removal or update of the offers, the external catalogs must be notified and implement the request. For example, a federated implementation of a data catalog is part of the Eclipse Dataspace Components (EDC)[9].

Distributed design

There are various concepts and subtypes for implementing a distributed design of data infrastructures. These types can be characterized according to their topology. The selected topology influences the properties and behavior of the data infrastructure, such as the synchronization behavior of the individual data infrastructure services of the same type. The following table summarizes a selection of possible topologies for designing distributed systems.

[9] https://eclipse-edc.github.io/

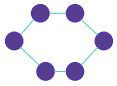
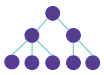
Topology	Description	Advantages	Disadvantages
 Ring	Connecting a node to two other nodes results in a closed ring-shaped data transmission path. Data travels along the path until it reaches its destination.	- Equality of the nodes - Simple error detection	- Susceptibility to failure - Limited scalability
 Line	Each node is successively linked to another node, but a ring closure is not performed, resulting in open ends.	- Simple implementation - Easy to extend	- Susceptibility to failure - Data collisions
 Tree	Hierarchical arrangement of individual nodes in the form of branches of a tree with ascending degrees of centralization per hierarchy level.	- Simple administration - Good scalability	- High connection costs - Failure of the central node affects the entire network
 Star	There is a central node that acts as the orchestrator of the distributed system. All nodes are connected to this orchestrator and exchange data with it.	- Simple troubleshooting - High reliability - Good scalability	- High connection costs - Failure of the central node affects the network
 Fully connected	Each node is connected to every other node. The nodes usually have equal functionality.	- High reliability - Direct connections - High redundancy	- Very high connection costs - Complex implementation and management

Table 1: Topologies of distributed architectures in comparison

This decision aid focuses on distributed ledger technology (DLT) as a particular type of a distributed system (see Figure 4). A distributed ledger is essentially an information store that is shared across the nodes involved in the system. A consensus mechanism ensures that the information can be kept synchronized across the nodes. Accordingly, each node of a DLT system always holds the same set of information. Like a land register, DLT systems only append information. A DLT system can also implement business logic based on smart contracts. Due to their data replication and use of consensus mechanisms, DLTs are particularly relevant for creating transparency and trust in data infrastructures. They are currently extensively tested and used in the areas of finance and supply chain management.

To implement a data infrastructure service based on DLT, data infrastructure participants, contracted service providers and other stakeholders (e.g. public organizations) each operate a node of the distributed system. The data pool and the business logic of the smart contracts are replicated by the participating nodes. If another actor wishes to operate a node and participate in the system after its setup, a consensus must be reached among the existing nodes. However, actors can also choose to interact with the data infrastructure via an existing node, provided this is permitted by the node provider. If individual participants in the DTL system decide to terminate their nodes, the service is maintained by the other nodes and the existing data remains available.

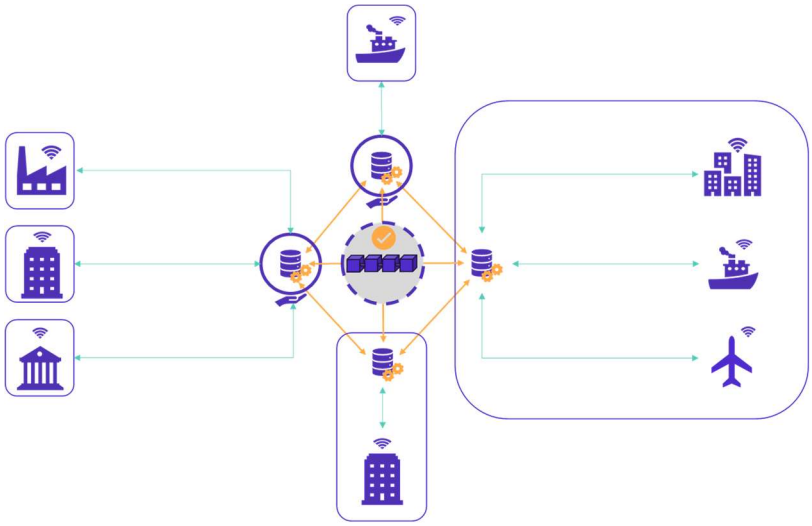


Figure 4: Schematic representation of distributed data infrastructure services based on distributed ledger technology

Example data and service catalog:

To implement the catalog service, several actors and service providers join forces, each operating a node of the distributed system. Each node is identical and behaves according to the jointly defined rules. To list its offers, an actor makes the metadata of its data and services available to the distributed data infrastructure service via its own node or via the node of other network participants. If a consensus on inclusion is reached between the nodes, the metadata is stored by the distributed catalog service. Each of the nodes receives a synchronous image of the listed offers. Once written to the distributed ledger, the offers cannot be removed. Instead, they are labelled as inactive or expired. If one of the nodes fails, a current image is still available to all other nodes. An example of the implementation of a data and service catalog based on DLT is implemented as part of Ocean Protocol[10]. In Ocean Protocol, data offers are represented through non-fungible tokens.

Comparison of the characteristics of the design options

A comparison of the design options is shown in the following table. Services of the same type can be available either once or several times within a data infrastructure. The services can be provided either by dedicated service providers with no interest in participating in data sharing or by the data providers and data consumers themselves. In the case of multiple provision of services of the same type, the services can be completely isolated from each other, replicate the data completely among themselves or decide autonomously which data is made available to which service.

[10] https://docs.oceanprotocol.com/

Design option	Number of services of the same type	Provision of services	Relationship between services of the same type
Fully centralized	1	Service provider	-
Fully decentralized	>1	Service providers and/or data providers/data consumers	Fully insulated
Federated	>1	Service providers and/or data providers/data consumers	Self-determined data sharing
Distributed	>1 (virtual 1)	Service providers and/or data providers/data consumers	Fully replicated

Table 2: Comparison of the design options

In addition to the architectural options outlined, there exist other hybrid forms or specific feature variations that are not covered in this decision aid to abstract complexity. Various properties and framework conditions can influence the decision to adopt or reject a specific design option. In any case, the appropriate architectural option should be chosen for the specific application domain to provide optimal support for the intended applications and to encourage broad stakeholder participation in the planned applications.

3 A Procedure Model for the Design of Data Infrastructure Services

The procedure model presented in this section aids companies in designing data infrastructure services by outlining the necessary activities and outcomes required to make an informed decision regarding the system architecture. It is based on established methods in the field of enterprise architecture. To this end, concepts from the Generalized Enterprise Reference Architecture (ISO, 2019) and the Business Motivation Model (Object Management Group, 2015) are adapted to the design of data infrastructure services. The procedure focuses on setting up a new infrastructure between the actors involved in data sharing (greenfield approach). The outcome of the procedure model is the selection of one of the previously described design options for data infrastructure services.

The starting point for applying the procedure model is the existing specification of one or more business models based on data sharing mechanisms. Elements such as the value proposition, the target group, the data sharing partners, the key resources and possible revenue mechanisms should be known before the data infrastructure design. These aspects are used below to derive concrete boundary conditions and requirements, which in turn form the basis for a well-considered design decision.

Figure 5 shows an overview of the life cycle of data infrastructure services. Each block represents a life cycle phase, spanning from environmental analysis to the decommissioning of the data infrastructure. The respective phases are characterized by a logical dependency. Preceding phases provide a mandate in terms of standards, requirements or boundary conditions for the subsequent phases. Conversely, the preceding phases may provide feedback regarding the actual feasibility or usefulness of the previously defined concepts. As part of a continuous improvement process, these feedback loops enable the architecture to be continuously adapted based on internal or external changes.

In this decision aid, we focus on the first four lifecycle phases to decide on the high-level architecture of the data infrastructure and its services:

- Environmental analysis: In this initial phase, the relevant stakeholders and influencers and their expectations are identified and evaluated.
- Business model analysis: This phase identifies the relevant data and systems and outlines the data flows required for service provision based on the business model.
- Requirements definition: This phase derives the requirements for the data infrastructure from the environmental analysis and business model analysis and prioritizes them.
- Preliminary design: This phase works out the possible design options for implementing the data infrastructure and identifies the best possible alternative based on the previously defined requirements.

The phases of "Detailed design", "Implementation", "Operation and maintenance" and "Decommissioning" are not part of this decision aid due to their limited generalizability.

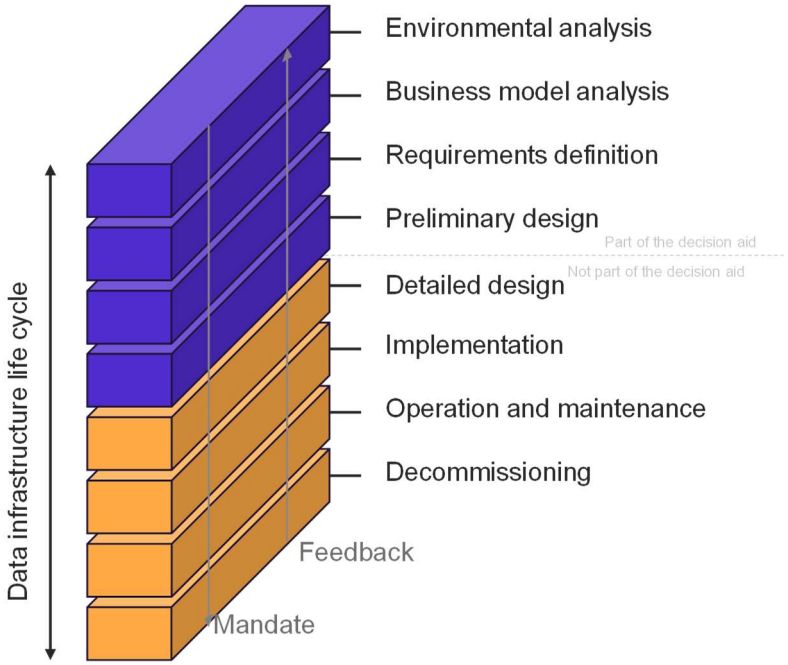


Figure 5: Procedure model for data infrastructure services

3.1 Phase 1 - Environmental analysis

Every application supported by the data infrastructure, and consequently the data infrastructure itself, exists within an environment that imposes specific expectations and influences on its design. Conversely, the data infrastructure itself causes environmental effects that arise through direct interaction with the environment or indirectly during its life cycle (e.g., energy consumption). The aim of the environmental analysis is to identify the relevant stakeholders situated in the environment and to derive further factors influencing the design of the data infrastructure. A stakeholder is a person, organization or group that has a claim or interest in the resulting data infrastructure. Stakeholders attempt to influence the design of the data infrastructure. Influencing factors basically include all aspects that can exert an influence on the design of the data infrastructure without the direct exercise of coercion or power (Object Management Group, 2015). Influencing factors can have technological, business, organizational, political, regulatory, ecological or social backgrounds, amongst others.

Identifying and outlining the environment is relevant in many respects. Through the environmental analysis, overarching social expectations of the solution can be derived, which in turn form the basis for the influence of various stakeholders. Particularly in the development of solutions with a broad scope that impact numerous stakeholders, factors related to long-term social and environmental responsibility become paramount interests. The environmental analysis identifies specific stakeholder groups that can steer the decision-making process in a particular direction that deviates from the optimal solution. These groups include Non-governmental organizations (NGOs), for example.

As part of the environmental analysis, three consecutive topics need to be addressed. The starting point is the identification of relevant stakeholders. Subsequently, the stakeholders' expectations and influences are described. Finally, the impact of these expectations is assessed.

The stakeholders of a data infrastructure encompass both actors who influence its design and those whose actions are affected by their design. The environmental analysis focuses on external stakeholders. According to the acronym PESTEL, the external environment is usually analyzed with regards to political, economic, social, technological, ecological and legal factors.

The interests of a stakeholder are expressed in the form of demands on the development of the data infrastructure. These requirements originate from the stakeholder's perspective and are based on their own knowledge and experience, their responsibility and their mission. The requirements can be of a general nature or relate explicitly to design aspects of the data infrastructure, such as the supported standards. A stakeholder can have one or more claims. The same claim can also be asserted by different stakeholders. The identified stakeholders, their claims and possible influences must not only be named, but also described in a way that is comprehensible to other parties involved. The tone of the description should be as neutral as possible to separate the claims description from their assessment. The identified stakeholders, expectations and influences can be summarized in a table format. An overview of possible stakeholders and their expectations and influences in the development of data infrastructures is summarized in the following table.

Stakeholders	Exemplary interests, claims and influences
Regulation	Compliance with regulatory obligations, particularly in the areas of data law, antitrust law and product liability
Users	Long-term availability of the solution; Use of established standards and guidelines; Achievement of relevant certifications
Competitors	Benchmarking of the developed solution; Interoperability and standardization; Coopetition
Technology partners	Utilization of resources, technologies and service offerings available on the market; Consideration of current trends and facilitation of future improvements
Shareholders and partners	Maintaining competitiveness; Growth of the company; Achieving positive social and sustainability indicators
NGOs	Reducing the environmental impact of digital infrastructures; Ensuring digital rights and privacy; Implementing fair business practices; Meeting additional domain-specific concerns (e.g., about consumer and patient protection, employee rights or industry interests)
Research and development	Insight into solutions used by researchers; Use of new research findings as part of the implementation of the data infrastructure

Table 3: Stakeholders and their exemplary expectations and influences in the context of data infrastructures

Finally, the extent to which the stakeholder can legitimately influence the design of the data infrastructure is evaluated. This essentially determines whether the claim or influence asserted by the stakeholder must be considered in the further phases. The assessment is based on the role of a stakeholder and their financial, social or legal status. An assessment can be made in terms of risks or opportunities (Object Management Group, 2015). According to this approach, the financial benefits and risks associated with addressing or ignoring a stakeholder's concern are evaluated against each other, resulting in a prioritization of claims based on monetary criteria.

3.2 Phase 2 - Business model analysis

A data infrastructure is set up to support at least one application and the underlying business model. Accordingly, it is important that the requirements of a data infrastructure are clearly motivated from the underlying business model. To achieve this, it is essential to establish a fundamental understanding of the business: First, one must address the question of "What needs to be done?" before determining how it can be accomplished.

Identifying the characteristics of the business models is relevant to map the requirements resulting from the business model to the technology and to avoid a discrepancy that could hinder the success of the planned applications. For example, the business model can provide information about expected profits and savings. These monetary key figures then have an influence on whether the costs of a data infrastructure are a relevant influencing factor. The relationship between the characteristics of the business model and the requirements for the data infrastructure is visualized in Figure 6.

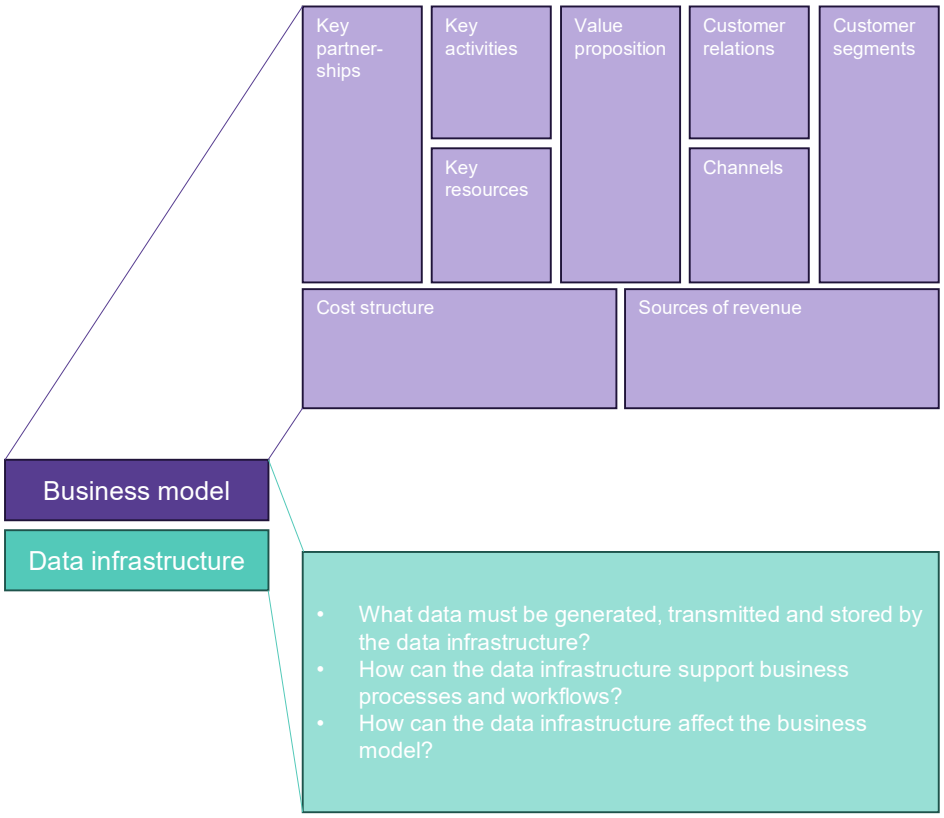


Figure 6: Relationship between business model and data infrastructure based on Osterwalder et al. (2010)

The aim of the "business model analysis" phase is therefore to derive relevant characteristics of the business models that have an influence on the design of the data infrastructure. To this end, the core components of the business model are initially identified. Subsequently, the required data is characterized in detail as a value-added resource, and the data flows between the source and target systems are described.

Established tools such as the " Business Model Canvas" (Osterwalder et al., 2010) can be used to describe the business model. The canvas helps to identify and structure the key building blocks of a business model and to generate a common understanding of the business model.

Based on the Business Model Canvas, other similar tools with the same concept emerged in recent years. One example is the "Data Cooperation Canvas" (DS4SSCC, 2023), which takes up the general concepts of business models and expands them to include specific aspects of data sharing such as the interoperability standards and data demands.

Regardless of the specific business model analysis tool, it is important to pay attention to data as a central strategic resource in data sharing business models in order to be able to derive influences on the data infrastructure. For example, data update intervals provide an indication of whether the data infrastructure needs to be real-time capable. The following data-related characteristics should therefore be described in more detail:

- Metrics and results to be achieved through data integration and data analysis
- Required data and their data holders (own data, data from business relationships, open data, purchased data)
- Relevant data sources (data-holding systems) and data sinks (e.g. the analysis services to be used)
- Other characteristics of the required data (frequency, quality, sensitivity, formats, other metadata)

Figure 7 depicts a template for the description of data resources. The data providers field describes individual natural or legal persons or groups of these entities that provide data for the realization of value-added services. These actors are characterized by their negotiating power, which results, for example, from the importance of their data for the use case or the legal framework, such as obligations to disclose data. The bargaining power of the data providers is an essential characteristic for assessing the relevance of their concerns. The characteristics of the data provided summarize properties of the data from which requirements for the data infrastructure can be derived. Data sharing rewards are assessed by the reward type and the perceived value for the data provider under concern.

Actor							
Name	...						
Type	Natural Person			Organization			
Bargaining Power	Strong		Medium		Low		
Characteristics of provided Data							
Data Sources	...						
Personal Data	Yes			No			
Data Type	Structured	Semi-structured	Data Stream	Text	Audio	Image	Video
Data Update	Real-time		Interval		One time transfer		
Real World Entity	Company	Person	Asset	Process	Patent	Location	Others
Data Sharing Rewards							
Reward	Money	Virtual Assets	Data	Service	Reputation	Regulatory Compliance	None
Reward Value	High		Medium		Low		

Figure 7: Template for characterizing relevant data

Further, additional boundary conditions relating to the data infrastructure can be derived from the business model. These include, for example, the number and characteristics of future users,

the budget restrictions, the possible technology partners and the remuneration mechanisms for data provisioning.

Once the relevant data, the sources and the data sinks have been defined, they can be visualized in the form of data flow diagrams. Data flow diagrams support system design, optimization and documentation. Thanks to their visual form, data flow diagrams support communication between individual participants and make the system boundaries visible. Data flow diagrams also show in which context which support services are required by the data infrastructure. A classic method for describing data flows is the notation according to Gane & Sarson (1979). An extended version of this notation covering relevant elements for the visualization of data flows in data sharing applications is shown in Figure 8.

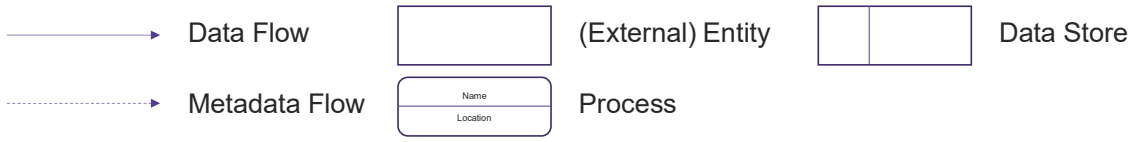


Figure 8: Data flow notation adapted to data sharing applications according to Gane & Sarson (1979)

3.3 Phase 3 - Requirements definition

In this phase, the requirements for the data infrastructure are derived and evaluated based on the characterization of the environment, the business model and the data processes. This procedure allows the defined requirements to be traced back to specific environmental conditions and influencers. It further enables traceability of requirements at later stages of the data infrastructure implementation process. In addition, this procedure offers the possibility of adapting the requirements structures in a controlled manner in the event of changing influences, such as changes in the regulatory framework or adjustments to the business model.

The result of this phase is a catalog of requirements. The requirements catalog includes relevant decision criteria and their prioritization. The first step is to identify the requirements. Afterwards, the factors are evaluated.

As with other software sytems, the requirements for data infrastructures can be divided into functional and non-functional requirements. Functional requirements describe the desired behavior of the system. The system behaviour is realized through services, tasks or functions. Functional requirements describe the "what" of the data infrastructure. Non-functional requirements are criteria that describe the quality and characteristics of a system without referring to specific functions or behaviors. They describe the "how" of the data infrastructure. While the functional requirements therefore relate to the functions of the data infrastructure services to be implemented, the non-functional requirements correspond to the quality and decision criteria relevant for the selection of the design option. Accordingly, the non-functional requirements are the focus of this phase.

The central element of this decision aid is a detailed catalog of criteria that explains relevant decision criteria and shows under which circumstances special consideration of a criterion is advisable. Further, the catalog of criteria lists, what relationships between individual criteria might

possibly exist. The criteria catalog can be found in section 4[11]. Figure 9 shows an overview of relevant decision criteria for data infrastructures. The decision criteria for the implementation of a data infrastructure can be of an economic, organizational or technical nature. Economically motivated decision criteria result from the financial aspects of the business models to be supported. Organizational decision criteria describe decision criteria that are mainly realized by organizational measures. The technical decision criteria describe the way in which a service must be technically provided.

Figure 9: Decision factors for data infrastructures

For a specific data infrastructure, relevant decision criteria can be derived from the overview based on prior environmental and business model analysis. The first step is to assess whether these criteria are pertinent to the success of the data infrastructure initiative, reflecting on the preliminary work conducted. The format of "Joint Requirements Planning Sessions" is suitable for this purpose. These structured workshops bring together relevant stakeholders to identify and plan project requirements. In the collaborative setting, the decision factors listed in the overview are consolidated and amended as necessary. Decision factors should be traced back to specific environmental conditions or business model requirements, with the rationale documented to ensure long-term transparency. Engaging stakeholders cooperatively in identifying decision criteria prevents conflicting or redundant requirements. The table below illustrates an individualized list of criteria as an exemplary outcome of this step.

[11] In addition, an editable version of the criteria catalog for conducting a multi-criteria assessment to select the data infrastructure design option is available online at https://doi.org/10.5281/zenodo.15223378.

Category	Decision factor	Description
Economical	Cost amortization	The costs of a data infrastructure refer to the total financial expenditure required to build, set up, operate and maintain the data infrastructure. Amortization refers to when the costs are refinanced through profits and depreciation.
Organizational	Data sovereignty	Data Sovereignty describes the ability of a data provider to be self-determined about the usage of its data assets (including data and metadata) along their whole lifecycle, from data creation to data deletion and including actions conducted by third parties. Data Sovereignty requirements are usually formulated in contracts or policies.
Organizational	Trustworthiness	Trust refers to the willingness of one party to expose itself to the actions of another party, regardless of the ability to monitor or control that other party. In the context of data sharing, trust concerns the actions of the data sharing partners, the provisioning of data and the behaviour of the data infrastructure itself. It is governed by the technological design choices such as the implementation of zero-trust mechanisms.
Technical	Resource intensity	Resource intensity of a data infrastructure comprises storage and computation demands leading to energy expenditures.
...	...	...

Table 4: Exemplary criteria catalog

Not all requirements are equally relevant for the successful implementation of the data infrastructure. As such, requirement criteria must be prioritized to facilitate informed decision-making. Conventional prioritization methods from the field of multi-criteria decision-making are employed. Commonly used methods include the Analytical Hierarchy Process (AHP), the Hundred-Point Test, the Weighted Sum Model, and simple evaluation based on criteria such as High-Medium-Low. The first three methods use numerical values to mathematically assign a distribution (100 percent of the requirement points) to decision factors, while the High-Medium-Low scheme first requires defining the associated numerical values for individual factors. Additionally, each decision criterion must be assessed to determine whether it positively or negatively impacts the selection decision. For instance, high transparency of a data infrastructure may be desirable in certain situations but not in others. The table below visualizes an exemplary result of a prioritization exercise.

Category	Decision factor	Description	Interpretation	Priority
Economical	Cost amortization	...	Positive	3
Organizational	Data sovereignty	...	Positive	2
Organizational	Trustworthiness	...	Positive	2
Technical	Resource intensity	...	Negative	3
...	...	...	...	...

Table 5: Exemplary weighted criteria catalog

3.4 Phase 4 - Preliminary design

Based on the individual criteria catalog, the appropriate design option for the data infrastructure is identified in this phase. The selection of the design option serves as the crucial starting point for the subsequent detailed planning and implementation of the data infrastructure. Two steps are required to select the best fitting design option. First, the architectures that enable value creation within the scope of the planned applications are identified. Then, these architectures are compared and evaluated using the previously defined and weighted criteria.

A preselection of possible design options for data infrastructure services is conducted based on their potential for value creation. This assessment determines whether a design option's inherent characteristics allow for the implementation of the planned applications. For instance, legal requirements might prohibit data transfer to a centralized or fully distributed service, eliminating these designs from further consideration. This preselection process narrows the decision-making scope, thereby reducing the complexity of the final evaluation.

For the final assessment of the design options, they are compared and evaluated according to the requirements criteria. Using the same evaluation system as in the requirements prioritization helps minimize comprehension issues and stakeholder effort. The final ranking of design options is then established based on the prioritization results. The table below provides an example of a decision catalog including evaluation. An initial indication of how well potential design options meet the decision criteria for data infrastructure services can be found in the criteria catalog presented in the next section.

Category	Decision factor	Description	Interpretation	Priority	Centralized	Distributed
Economical	Cost amortization	...	Positive	3	3	1
Organizational	Data sovereignty	...	Positive	2	0	1
Organizational	Trustworthiness	...	Positive	2	1	3
Technical	Resource intensity	...	Negative	3	2	1
...	...	...	...	...	...	...
				Total	17	**22**

Table 6: Exemplary criteria catalog including evaluation

4 A Criteria Catalog for Data Infrastructure Services

This criteria catalog lists and evaluates relevant decision factors when selecting a design option for data infrastructures. The criteria catalog provides guidance in the phases of requirements definition and the preliminary design of data infrastructures. The relevant decision factors in the selection of a data infrastructure are described in more detail below. Each influencing factor is represented by an individual canvas. The canvas describes the following information:

- Category of the decision factor: Primary motivation of the decision factor (economic, technical, organizational)
- Designation of the decision factor
- Description of the decision factor
- Relevance of the decision factor: Description of the applications or circumstances necessitating a more comprehensive analysis of the decision factor
- Fulfillment of the decision factor through the data infrastructure design option: Description and justification of the extent to which the decision factor can be realized through the data infrastructure design option
- Important interdependencies (if any): Other decision factors that exert influence on the decision factor under consideration

An editable version of this criteria catalog, including a template for conducting a multi-criteria analysis to select the design option, is also available in the form of a table document[12]. The criteria catalog was developed through literature research and subsequently validated, elaborated, and expanded via expert interviews.

Economical	**Cost amortization**						
The costs of a data infrastructure refer to the total financial expenditure required to build, set up, operate and maintain the data infrastructure. The amortization refers to when the costs are refinanced through profits and depreciation.							
Relevance	The consideration of cost amortization for a data infrastructure is particularly pertinent when enabling commercial applications with low margins.						
Fulfillment through data infrastructure design option							
Centralized	+	Federated	o	Decentralized	o	Distributed	o
Centralized data infrastructures usually cause lower overall costs than decentralized, federated or distributed systems in which a network of several nodes has to be set up.							
Important interdependencies							
Operational efficiency	↗	Cost amortization					
Complexity	↘	Cost amortization					
Resource intensity	↗	Cost amortization					

[12] The digital version of the criteria catalog is available online at https://doi.org/10.5281/zenodo.15223378.

Economical	Operational efficiency						
Operational efficiency refers to the performance, speed and efficiency with which transactions can be executed by the data infrastructure. Operational efficiency is promoted by automation, for example.							
Relevance	Operational efficiency plays a particularly important role in data sharing applications in which data has to be exchanged between different parties at short intervals and is constantly changing.						
Fulfillment through data infrastructure design option							
Centralized	+	Federated	○	Decentralized	-	Distributed	-
Centralized infrastructures reduce the technical overhead of data replication and support the rapid automation of processes, leading to operational efficiency.							
Interdependencies							
Scalability		↗		Operational efficiency			
Complexity		↗		Operational efficiency			
Orchestrability		↗		Operational efficiency			

Economical	Sustainability						
The sustainability of a data infrastructure refers to the ability to use it in the long term without compromising economic or social aspects (Tan et al., 2024).							
Relevance	Economic and social sustainability is especially significant for applications with social and societal impact, as well as for data infrastructures intended for long-term use.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	○	Decentralized	○	Distributed	+
Independence from individual actors implies that a distributed infrastructure can function autonomously over the long term, thereby avoiding lock-in effects negating sustainability.							
Interdependencies							
Portability		↗		Sustainability			
Fairness		↗		Sustainability			
Generativity		↗		Sustainability			
Complexity		↘		Sustainability			

Organizational	Trustworthiness						
Trust refers to “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (Mayer et al., 1995). In the context of data sharing trust concerns the actions of the data sharing partners, the provisioning of data and the behaviour of the data infrastructure itself. It is governed by the technological design choices such as the implementation of zero-trust mechanisms.							
Relevance	Trustworthiness is a relevant factor for almost all applications. However, it plays a particularly important role when interacting with potentially unknown actors.						
Fulfillment through data infrastructure design option							
Centralized	○	Federated	+	Decentralized	+	Distributed	+
The capability to independently operate a part of the system and remain free from central authority enhances trust in the data infrastructure.							
Interdependencies							
Reliability		↗		Trustworthiness			
Autonomy		↗		Trustworthiness			
Auditability		↗		Trustworthiness			

Organizational	Data sovereignty						
Data Sovereignty describes the ability of a data provider to be self-determined about the usage of its data assets (including data and metadata (Bastiaansen et al., 2019)) along their whole lifecycle, from data creation to data deletion and including actions conducted by third parties (Scherenberg et al., 2024; Hellmeier und Scherenberg, 2023). Data Sovereignty requirements are usually formulated in contracts or policies (Scherenberg et al., 2024).							
Relevance	Data sovereignty is especially crucial when handling highly sensitive data for industrial applications and when directly integrating personal data of individuals.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	+	Decentralized	+	Distributed	○
Data infrastructure designs that avoid the physical or virtual centralization of data increase the data sovereignty of the actors involved.							
Interdependencies							
Legal certainty	↗	Data sovereignty					
Transparency	↘	Data sovereignty					
Autonomy	↗	Data sovereignty					

Organizational	Transparency						
Transparency describes the degree of visibility and accessibility of information about the data infrastructure, the players and the data transactions taking place. This includes elements such as the current status as well as the scope, location and unit of service provision.							
Relevance	Transparency is vital for applications where tracking the actions of the data infrastructure is necessary, such as enabling fair cooperation or facilitating financial compensation. However, in certain situations, high transparency regarding data or transactions may be undesirable.						
Fulfillment through data infrastructure design option							
Centralized	○	Federated	○	Decentralized	○	Distributed	+
In distributed (DLT-based) data infrastructures, the transactions carried out are visible to all actors operating a node, thus creating high transparency.							
Interdependencies							
Data integrity and data verifiability	↗	Transparency					

Organizational	Auditability						
Auditability refers to the ability of the data infrastructure to be audited (i.e. inspected and verified) by a third party.							
Relevance	The auditability of a data infrastructure is particularly relevant in areas with high regulatory requirements.						
Fulfillment through data infrastructure design option							
Centralized	+	Federated	○	Decentralized	-	Distributed	-
The effort required to audit a data infrastructure is lowest for centralized architectures. In multi-node architectures, each individual node potentially needs to be audited.							
Interdependencies							
Autonomy	↘	Auditability					
Complexity	↘	Auditability					

Organizational	Autonomy						
Autonomy refers to the extent to which an actor can use data infrastructure services without being dependent on another specific actor.							
Relevance	High demand for autonomy is particularly anticipated in industrial sectors characterized by "winner-takes-it-all" effects, where infrastructure providers also have a vested interest in the development of smart services.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	+	Decentralized	+	Distributed	+
In federated, decentralized and distributed architectures, data infrastructure services can be implemented without dependence on a central actor.							
Interdependencies							
Portability		↗		Autonomy			
Complexity		↗		Autonomy			

Organizational	Complexity						
Complexity is a characteristic that is perceived by natural persons. It refers to the comprehensibility of a technology and the simplicity with which it can be developed, implemented, maintained and used. The greater the complexity, the greater the know-how required to implement the solution and the time needed to achieve productive use for the first time.							
Relevance	Low complexity and the resulting technology acceptance are particularly important for applications involving interactions between small and medium-sized enterprises (SMEs) or individuals via the data infrastructure. Moreover, low complexity facilitates the achievement of quick wins.						
Fulfillment through data infrastructure design option							
Centralized	+	Federated	○	Decentralized	+	Distributed	○
In this sense, the "+" corresponds to good (low) complexity. Centralized data infrastructures are usually easy to implement and maintain. The same applies to single services in decentralized data infrastructures.							

Organizational	Fairness						
Fairness refers to the degree to which the data infrastructure establishes a "level playing field" by balancing the needs of the involved parties and ensuring openness to additional service offerings.							
Relevance	A fair opportunity to utilize the data infrastructure is crucial in scenarios where competing service offerings are supported by the same data infrastructure.						
Fulfillment through data infrastructure design option							
Centralized	○	Federated	+	Decentralized	+	Distributed	+
Fair use of the data infrastructure is ensured by providing dedicated data infrastructure nodes, which allow customization to meet the specific needs of each actor.							
Interdependencies							
Complexity		↗		Fairness			
External interoperability		↗		Fairness			
Portability		↗		Fairness			

Organizational	Confidentiality						
The confidentiality of a data infrastructure includes mechanisms that promote the protection of privacy in order to meet the needs of data providers and data subjects as well as legal requirements.							
Relevance	The confidentiality of a data infrastructure needs to be especially considered when personal data or data containing business secrets is transmitted via the data infrastructure.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	+	Decentralized	+	Distributed	-
Confidentiality is particularly ensured by architectures where data is shared only with selected parties and can be processed locally.							

Organizational	Legal certainty						
Legal certainty refers to compliance with applicable legislation. This includes general data legislation such as the General Data Protection Regulation, the Data Act or the AI Act as well as additional area-specific regulatory requirements, e.g. in relation to data security or data localization.							
Relevance	Compliance with relevant legal regulations is relevant for any data infrastructure.						
Fulfillment through data infrastructure design option							
Centralized	○	Federated	○	Decentralized	○	Distributed	○
In general, no architecture inherently offers greater legal security; it is always prudent to assess each individual case.							

Organizational	Strategic fit						
Strategic fit refers to the extent to which the data infrastructure harmonizes with the strategies relevant to the individual actor. These include the organizational strategy, the IT strategy or the data strategy.							
Relevance	The need to assess the strategic fit must be evaluated for the individual organization.						
Fulfillment through data infrastructure design option							
Centralized	○	Federated	○	Decentralized	○	Distributed	○
A general assessment of the strategic fit is not possible, as this depends on the individual strategies of each actor.							

Organizational	Internal interoperability						
Internal interoperability describes the ability of individual data infrastructure services to exchange information with each other and to use the exchanged information mutually. From the perspective of the data infrastructure, interoperability mainly refers to organizational, semantic and technical interoperability (Firdausy et al., 2022).							
Relevance	Internal interoperability between data infrastructure services is pertinent for applications requiring spatially distributed data infrastructure services.						
Fulfillment through data infrastructure design option							
Centralized		Federated	+	Decentralized	-	Distributed	○
Since a federated system is based on interoperability and information exchange, internal operability is generally greatest here. If there is only one central service, interoperability between services of the same type is not necessary.							

Organizational	External interoperability						
External interoperability describes the ability of a data infrastructure to integrate with existing and utilized enterprise data infrastructures, platforms and other existing data infrastructures (e.g. in related sectors) on a technical and process level to provide a seamless experience. External interoperability is required in terms of legal, organizational, semantic and technical aspects.							
Relevance	A high level of interoperability with external systems is particularly necessary if seamless processes are to be ensured between the systems of an actor and the data infrastructure services or if there are frequently used platforms whose replacement by the data infrastructure is associated with high costs.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	+	Decentralized	○	Distributed	-
Due to the flexibility and adaptability of the services, federated data infrastructures can potentially achieve the highest level of interoperability with external services.							

Technical	Portability						
Portability is the ability to easily move data or software services from one system to another.							
Relevance	The portability of a data infrastructure service is typically relevant when its benefits entail higher long-term costs or there is a risk that the system may become inoperable.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	+	Decentralized	+	Distributed	○
Federated and decentralized architectures generally provide the best support for moving services from one system to another.							

Technical	Scalability						
Scalability of a data infrastructure refers to the ability to support potentially increasing sizes of data, numbers of transactions, or numbers of supported participants as while ensuring constant performance.							
Relevance	The scalability of a data infrastructure is particularly relevant for use cases with a large number of participants or a high frequency of data transactions.						
Fulfillment through data infrastructure design option							
Centralized	○	Federated	+	Decentralized	+	Distributed	-
Centralized systems represent a bottleneck in the processing of requests. Distributed systems are often limited in the number of transactions due to their need to find consensus among all nodes.							
Interdependencies							
Cost amortization			↘			Scalability	

Technical	Resource intensity						
Resource intensity of a data infrastructure comprises storage and computation demands leading to energy expenditures.							
Relevance	The resource intensity of a data infrastructure is an important criterion, particularly for sustainability-oriented applications. This criterion should also be considered for highly scalable data infrastructures.						
Fulfillment through data infrastructure design option							
Centralized	+	Federated	○	Decentralized	○	Distributed	-
Since resource consumption rises with the increasing number of redundant services, a centralized data infrastructure design typically exhibits the lowest resource intensity.							

Technical	Generativity						
Generativity describes the characteristic of a data infrastructure to be adapted or extended to enable further value added services or use cases.							
Relevance	The generativity of a data infrastructure is particularly significant when the scope of the supported use cases cannot be anticipated at the time of design or is not yet fully defined.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	+	Decentralized	+	Distributed	-
In particular, federated and decentralized data infrastructures enable the adaptation of individual services as long as defined minimum standards are adhered to.							

Technical	Ability to orchestrate						
Orchestration describes the ability to globally optimize and coordinate the data infrastructure.							
Relevance	The global orchestration capability of a data infrastructure is especially beneficial when an application or system spans multiple geographical regions, necessitating coordinated and efficient data processing and management.						
Fulfillment through data infrastructure design option							
Centralized	○	Federated	○	Decentralized	-	Distributed	+
A distributed architecture offers the best conditions for global orchestration, as it combines virtually centralized control with distributed and redundant data storage.							

Technical	Data integrity and data verifiability						
Data integrity refers to the ability of a data infrastructure to ensure the accuracy and completeness of data throughout its lifecycle (Bastiaansen et al., 2019). Verifiability refers to the ability to check the accuracy and authenticity of the data itself.							
Relevance	The integrity and verifiability of data are crucial for applications that utilize data-based control mechanisms or allocate financial resources based on the provided data.						
Fulfillment through data infrastructure design option							
Centralized	+	Federated	○	Decentralized	○	Distributed	+
Ensuring data integrity is easiest to implement in centralized data infrastructures through centralized control instances. However, verification by external parties is difficult. A distributed infrastructure can be used for this purpose.							

Technical	Reliability						
Reliability characterizes the ability of the data infrastructure to work as intended without interruptions or failures at any time.							
Relevance	The reliable functioning of a data infrastructure is essential for almost all applications. It is most relevant for applications that require real-time data exchange and where delays in data transmission could cause the system to fail.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	○	Decentralized	○	Distributed	+
If individual nodes fail in federated or decentralized data infrastructures, the remaining nodes can continue to operate the system. In distributed systems, even if individual nodes fail, the overall image remains accessible through other nodes.							
Interdependencies							
Scalability			↗			Reliability	

Technical	Recoverability						
Recoverability refers to the ability to regain access to data and systems, e.g. in the event of a hardware failure or operating error.							
Relevance	The recoverability of a subsystem of the data infrastructure can be a particularly relevant criterion if there is a great fear of data loss.						
Fulfillment through data infrastructure design option							
Centralized	-	Federated	○	Decentralized	-	Distributed	+
Nodes of a distributed data infrastructure can restore data after errors by synchronizing with the other nodes.							
Interdependencies							
Reliability			↗			Restoration	

Technical	Information security and cyber security						
Security generally refers to the protection of assets against the various threats posed by certain inherent vulnerabilities. Information security refers to information as an asset, while cyber security also considers the protection of people and other assets accessible via cyberspace (e.g. machines) (Solms & van Niekerk, 2013).							
Relevance	Information and cyber security are particularly relevant for areas of critical infrastructure, but also for sectors such as industrial production, where cyber-attacks can potentially cause significant economic damage.						
Fulfillment through data infrastructure design option							
Centralized	○	Federated	○	Decentralized	○	Distributed	○
It is not possible to make a general statement on information or cyber security based on the architecture options.							

5 Outlook: the Next Steps

When designing data infrastructure services, an informed decision must be made based on the actual requirements of business models, considering essential social framework conditions and balancing the interests of various stakeholders involved in data sharing. This decision aid facilitates the informed design of the system architecture through a structured design process. After the preliminary design is chosen, the detailed design, implementation, and operation of the data infrastructure follow. Existing reference architectures from various initiatives can be utilized to support the detailed design of the data infrastructure.

Against the background of a centralized design of the data infrastructure, the reference architectures for big data from ISO (ISO, 2020) and NIST (NIST, 2019) explain roles, activities, functions and architectural components for implementing big data architectures. The architecture documents of the IDSA (International Data Spaces Association, 2022) and Gaia-X (Gaia-X European Association for Data and Cloud AISBL, 2024) can be used as a reference for the detailed design of federated data infrastructures. In addition, the SIMPL technical specification provides an overview of existing centralized and decentralized technical components for implementing the functional requirements of data infrastructures (Sovereign-X, 2024).

A reference architecture for DLT-based systems is defined in ISO 23257 (ISO, 2022). Due to the limited interoperability of individual DLTs and their varying characteristics, such as the consensus protocols employed, specific DLT frameworks must be evaluated for a particular implementation. In the domain of data sharing, the Hyperledger Fabric (Hyperledger, 2023) and Ocean Protocol (Ocean Protocol Foundation Ltd., 2024) frameworks are among the most utilized.

In addition to implementation by cloud services providers, dedicated companies that focus on the provision of data infrastructure services are increasingly being considered for the implementation and operation of data infrastructures. When operating federated or distributed infrastructures, it is possible to choose more than one operator or to give the individual actors a free choice regarding the operation of their data infrastructure services. This helps to reduce dependence on a singular provider.

The selection of system architecture for data infrastructures highlights the complexity of issues faced by actors in the data economy. Consequently, companies should develop expertise in data infrastructures and ecosystems to leverage the potential of applications based on shared data. Taking on one of the diverse roles such as application developer, system integrator, data provider, and or infrastructure operator offers vast value-creation opportunities. Expertise in data infrastructures and ecosystems also enables early participation in design phases, offering the advantage of shaping the infrastructure according to company needs. Data sharing skills are crucial for business continuity, especially considering increased regulatory obligations to provide data, such as those outlined in the Data Act (Straub & Bogenstahl, 2024).

References

Bader, S., Belyaev, A., Clauer, E., Diedrich, C., Fritz, R., Hasler, J., Karnouskos, S. & Schnicke, F. (2023). *Decentralized-Registries-Taxonomy-of-decentralized-registries-and-an-architectural-overview.* https://doi.org/10.13140/RG.2.2.34793.26723

Bastiaansen, H., Dalmolen, S., Kollenstart, M. & Punter, M. (2019). Infrastructural Sovereignty over Agreement and Transaction Data ('Metadata') in an Open Network-Model for Multilateral Sharing of Sensitive Data. *Fortieth International Conference on Information Systems.* https://aisel.aisnet.org/icis2019/economics_is/economics_is/23/

Demchenko, Y., Laat, C. de & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. In W. W. Smari (Ed.), *2014 International Conference on Collaboration Technologies and Systems (CTS 2014): Minneapolis, Minnesota, USA, May 19 - 23, 2014 ; [including symposia and workshops* (pp. 104-112). IEEE. https://doi.org/10.1109/CTS.2014.6867550

Dodds, L. & Wells, P. (2019). Data infrastructure. In T. Davies (Ed.), *The State of Open Data: Histories and Horizons.* African Minds. https://idrc-crdi.ca/en/book/state-open-data-histories-and-horizons

DS4SSCC. (2023). *The Data Cooperation Canvas.* https://www.datacooperationcanvas.eu/local/canvas/uploads/The%20Data%20Cooperation%20Canvas%20-%20Template.pdf

Firdausy, D. R., Alencar Silva, P. de, van Sinderen, M. & Iacob, M.-E. (2022). Towards a Reference Enterprise Architecture to enforce Digital Sovereignty in International Data Spaces. In *2022 IEEE 24th Conference on Business Informatics (CBI)* (pp. 117-125). IEEE. https://doi.org/10.1109/CBI54897.2022.00020

Gaia-X European Association for Data and Cloud AISBL (ed.). (2024). *Gaia-X Architecture Document - 24.04 Release.* https://docs.gaia-x.eu/technical-committee/architecture-document/24.04/

Gane, C. & Sarson, T. (1979). *Structured systems analysis: tools and techniques*. *Prentice-Hall software series*. Prentice-Hall.

Hellmeier, M. & Scherenberg, F. von (2023). A DELIMITATION OF DATA SOVEREIGNTY FROM DIGITAL AND TECHNOLOGICAL SOVEREIGNTY. *Thirty-first European Conference on Information Systems (ECIS 2023).* https://aisel.aisnet.org/ecis2023_rp/306/

Hyperledger (ed.). (2023). *A Blockchain Platform for the Enterprise.* https://hyperledger-fabric.readthedocs.io/en/release-2.5/

International Data Spaces Association (ed.). (2022). *IDS-RAM 4.* https://docs.internationaldataspaces.org/ids-knowledgebase/ids-ram-4

ISO (2019). *ISO 15704:2019: Enterprise modeling and architecture - Requirements for enterprise-referencing architectures and methodologies*.

ISO (2020). *ISO/IEC 20547-3:2020 : Information technology - Big data reference architecture - Part 3: Reference architecture*.

ISO (2022). *ISO 23257:2022: Blockchain and distributed ledger technologies - Reference architecture*.

Mayer, R. C., Davis, J. H. & Schoorman, D. F. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review.* https://doi.org/10.2307/258792

NIST. (2019). *NIST Big Data Interoperability Framework.* https://doi.org/10.6028/NIST.SP.1500-6r2

Object Management Group. (2015). *Business Motivation Model: Version 1.3.* https://www.omg.org/spec/BMM/1.3/

Ocean Protocol Foundation Ltd (ed.). (2024). *Architecture Overview.* https://docs.oceanprotocol.com/developers/architecture

Osterwalder, A., Pigneur, Y. & Clark, T. (2010). *Business model generation: A handbook for visionaries, game changers, and challengers*. Wiley.

Otto, B. & Burmann, A. (2021). European data infrastructures. *Informatics Spectrum.* Vorab-Onlinepublikation. https://doi.org/10.1007/s00287-021-01386-4

Otto, B., Hompel, M. ten & Wrobel, S. (Eds.). (2022). *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage* (1st ed. 2022). Springer International Publishing. https://doi.org/10.1007/978-3-030-93975-5

Scherenberg, F. von, Hellmeier, M. & Otto, B. (2024). Data Sovereignty in Information Systems. *Electronic Markets*, *34*(1). https://doi.org/10.1007/s12525-024-00693-4

Schleimer, A. M., Jahnke, N. & Otto, B. (2023). Architecture Design Options for Federated Data Spaces. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 3643-3652. https://scholarspace.manoa.hawaii.edu/bitstreams/941a67b1-d2e4-4394-8f3f-11b8b7d370e9/download

Solms, R. von & van Niekerk, J. (2013). From information security to cyber security. *Computers & Security*, *38*, 97-102. https://doi.org/10.1016/j.cose.2013.04.004

Sovereign-X. (2024). *Functional and Technical Architecture Specifications.* https://simpl-programme.

Straub, S. & Bogenstahl, C. (2024). *The current EU legislation in the field of digitization and data economy.* https://www.iit-berlin.de/wp-content/uploads/2024/12/KI_Leitfaden_EU_Recht-2024.pdf

Tan, K. L., Chi, C.-H. & Lam, K.-Y. (2024). Survey on Digital Sovereignty and Identity: From Digitization to Digitization. *ACM Computing Surveys*, *56*(3), 1-36. https://doi.org/10.1145/3616400

Weill, P., van den Berg, E., Birnbaum, J. & Planta, M. de. (2024). *What's Next: Top Performers Are Becoming Real-Time Businesses.* https://cisr.mit.edu/publication/2024_0801_RealTimeBusiness_WeillvanderBergBirnbaumdePlanta?utm_source=pressrelease&utm_medium=pr&utm_campaign=realtimebusiness#fn_1