GAP analysis

From data sharing proofs-of-concept towards operationalization of the system architecture

NL AI Coalition

December 2020

Contents

Management summary

Artificial Intelligence (AI) needs data to train and run algorithms. Therefore, an adequate data sharing infrastructure for AI is essential for improving the position of the Netherlands in the AI domain. However, as data is considered a valuable and potentially sensitive asset organizations want to be in control on who may use their data, for what purposes and under what conditions. This is referred to data sovereignty.

Data and data sharing are clearly on the radar of the European Commission. The NL AIC working group Data Sharing develops an overarching system architecture for controlled data sharing for AI as much as possible aligned with the EU Data Strategy and the architectural concepts and technology components that it promotes. These have previously been described in its report 'Responsible data sharing in AI', which is publicly available on the NL AIC website.

Based on an system architecture, the NL AIC working group Data Sharing has developed three illustrative and representative PoCs in 2020 on the main architectural concepts and technology components. The PoCs are based on use cases stemming from sectors represented in the NL AIC coalition: (1) a PoC on (flexible) permission management with a case from the government sector, (2) a PoC on distributed collaboration models with a case from the health sector and (3) a PoC on hybrid data sharing environments with a case from the energy sector. The system architecture and the three PoCs are described in this report.

The basic technology for realizing the system architecture is maturing. Therefore, the focus of the GAP analysis as described in this report is on the system operations gaps and the governance gaps to be bridged between the architectures and technology as demonstrated in the PoCs and the large-scale deployment and adoption thereof:

- The *system operations gaps* address the effective and efficient deployment and operations of the overarching system architecture for controlled data sharing for AI. It is a prerequisite for large-scale adoption. From each of the three PoCs, lessons learned for system operations have been derived. Furthermore, as a multitude of data sharing domains will emerge, e.g. to support AI in individual sectors or communities, interoperability becomes key: it enables seamless sharing of data over data sharing domains, extends the available data sets for AI-algorithms and prevents from a siloed approach. Therefore, the aspects of technical, semantic, organizational and legal interoperability have been addressed, together with gradual migration for data providers and interworking in hybrid data sharing environments.
- The *governance gaps* address the alignment of organizations within NL AIC to adopt a joint strategy for developing and deploying a system architecture for controlled data sharing for AI on aspects such as business viability, legal conditions, interoperability, standards and interfacing. Jointly, these aspects are referred to as a common 'Trust Framework for AI Data Sharing'. The development, introduction and adoption of this common trust framework will be a major goal of the NL AIC working group Data Sharing for the coming time period 2021 – 2024.

This 'GAP analysis' report is more technical in nature than the previous reports of the NL AIC working group Data Sharing. Nevertheless, the reader is encouraged to spend the effort in going through the report to grasp the overarching challenges and issues involved in further developing the system architecture towards operationalization and large-scale adoption. A further detailed technical elaboration of the system architecture is provided by the NL AIC working group Data Sharing in the report '*Blueprint NL AIC Data Sharing System Architecture*' which will be periodically updated with the latest insights.

1 Introduction

1.1 Background

"Data is a key strategic asset in our times. Data sharing is a means for valorization".

This certainly holds true for personal data, as data can tell so much about us. But it also holds for organizational data as it can optimize business ecosystems and supply chains, help the advancement of research, improve the functioning of government agencies and spur the economic and strategic position of countries and even regions.

Therefore, data and data sharing are key ingredients that are clearly on the radar of the European Commission, also in the context of artificial intelligence (AI). Together with an AI White Paper [1], the commission has released a paper on data governance and the role of data in AI [2]. Moreover, its release of the Data Governance Act [3] and the additional input sought on data spaces through OPEN DEI [4] point to the importance that the EU attribute to data and data sharing for our society and economy.

Similarly, also the Netherlands AI Coalition (NL AIC) has indicated that data and data sharing are essential for improving the strategic position of the Netherlands in the AI domain. That is a logical conclusion: AI-algorithms need data to allow algorithms to train, improve and be executed.

1.2 Scope of this report

Ideally, data is freely accessible. But reality is often different. Data has inherent value and is considered a strategic and valuable asset. In addition, there may be regulatory restrictions on sharing data, such as the General Data Protection Regulation (GDPR). Therefore, organizations want to be in control on who may use the data, for what purposes and under what conditions [5] [6]. This also applies to sharing data for AI.

In previous reports [7] [8] the NL AIC working group Data Sharing has identified the specific challenges for (responsible) data sharing for AI. These reports also provide an overview of architectures and technologies that can be used in addressing these challenges. Moreover, three Proofs-of-Concept (PoCs) have been developed in 2020 to demonstrate these architectural concepts and technical components, using illustrative and representative use cases in the sectors 'government', 'health' and 'energy'. These PoCs form the starting point in the process from first-time engineering towards operationalization of a data sharing infrastructure for AI in the Netherlands, as described in [9].

This report describes the overarching system architecture for controlled data sharing for AI. Its basic architectural concepts and technology components are maturing. Therefore, the focus of the GAP analysis in this report is on system operations, i.e. the gaps to be bridged between the architectures and technology as demonstrated in the PoCs and the large-scale deployment and adoption thereof.

1.3 Structure of this report

The following chapter 2 describes the ambition of the system for controlled data sharing for AI in terms of its architecture principles and system architecture. Chapter 3 addresses each of the three PoCs in 2020, including the lessons learned for further developing towards operationalization. Subsequently, chapter 4 focusses on the important aspects for large-scale adoption: interoperability, migration and interworking. Finally, chapter 5 provides the overarching conclusions on both system operations and governance.

2 A system architecture for controlled data sharing for AI

In the emerging data-driven economy, data markets and data sharing systems are currently attracting major attention. They provide functions for controlled sharing of data. However, they are to a large extend provided as part of a closed sector-specific domain, each with its own specific solutions. This poses major challenges for organizations that share data within multiple domains: they are faced with both a threat of vendor lock-in by their IT providers, and with major integration efforts across multiple data sharing relationships. Moreover, data sharing across domains and solutions is difficult. How advantageous would it be when data can be seamlessly shared in a controlled manner over domains such as smart industry, logistics and mobility…. Hence, a new system architecture for controlled data sharing for AI is needed, offering AI-algorithm providers low-barriers for access to data, whilst giving data providers the means to maintain control over their sensitive data.

The subsequent sections in this chapter addresses various aspects of the envisioned system: the architecture principles, the IDS-based reference architecture and the overarching system architecture. They are further elaborated in the Architectural Blueprint of the NL AIC working group Data Sharing [10].

2.1 Architecture principles

The architecture principles for the system architecture for controlled data sharing for AI are described in the following paragraphs.

2.1.1 Open network model approach with single entry point to a federation of building blocks

Data sovereignty is important for data providers to share their potentially sensitive data. Data sovereignty is currently mainly handled in a siloed 'hub-model' approach, i.e. as community-specific closed ecosystems, in which data providers are faced with both a threat of customer lock-in. Therefore, network-model approaches are currently attracting major attention in enabling a single entry point for data providers for sharing data over multiple data sharing relationships and within various data sharing domains [11], as depicted in Figure 1.

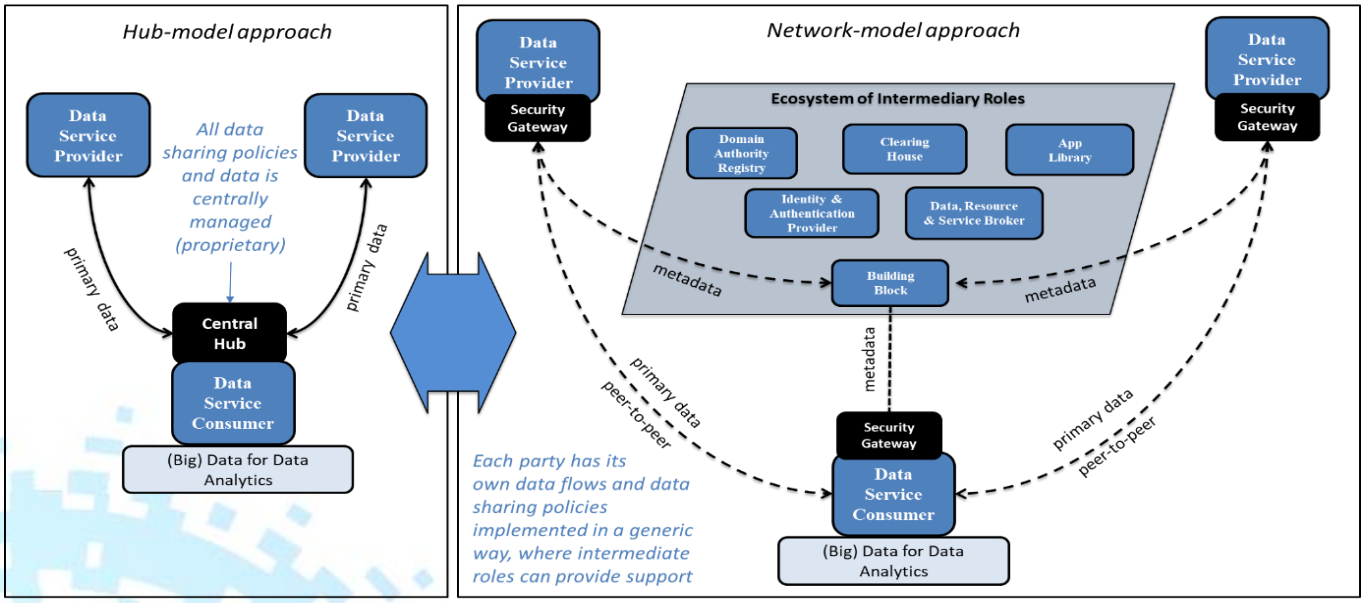

Figure 1. Hub-model (l) versus open network-model (r) approach for controlled data sharing.

The open network-model approach builds upon the principle of peer-to-peer data sharing between data providers and consumers, without the need for centrally storing or processing the data. This is enabled by a service-based system architecture in which a multitude of independent organizations federatively provide data sharing capabilities as generic and re-usable services, referred to as 'building blocks'. This allows building blocks to be developed as self-contained services so they can easily be functionally extended and adapt new technologies without major changes to the overarching system architecture. Moreover, they are not restricted in the technology used for their internal implementation. Jointly, the building blocks give data providers control and sovereignty when sharing potentially sensitive data. The service-based approach is a driver for development of an adequate portfolio of building blocks for data providers and consumers for maintaining sovereignty over their sensitive (meta)data [12].

By agreeing upon a basic set of building blocks and an agreed-upon information model in the open network model approach, a single entry point for the data provider can be created to simultaneously manage and control his data sharing relationships, yielding clear operational benefits over siloed hub-model approaches in user-friendliness, complexity, efficiency and costs [13]. Such a single entry point is also referred to as 'security gateway'. Moreover, it prevents data providers from both a threat of customer lock-in by specific data sharing solutions and from major integration efforts on managing data control and sovereignty capabilities over multiple data sharing relationships.

2.1.2 Enabled for locally executing data apps at the data provider: distributed data analytics

Data apps may be used to process data locally within the domain of the data provider or Data Service Consumer. This is referred to as 'app enabling'. Locally executing data apps may for instance be used for data enrichment, for semantic conversion, data quality management and de-identification (anonymization, pseudonymization).

A specific type of locally executing data apps are for distributed data analytics. Privacy preserving techniques are becoming available for which the data to be processed does not have to be gathered into a single database or location. As such, federated learning is able to learn by distributed data analytics algorithms. Furthermore, secure Multi-Party Computation offers possibilities to execute

algorithms on encrypted data without external parties having the opportunity to decrypt the source data itself [7] [8]. These technologies can be used in case the different data sources cannot be simply brought together and should remain at their source location, either because the amounts of data are too large or due to confidentiality. Think of privacy restrictions due to GDPR or company confidentiality.

As such, two basic 'collaboration models' can be considered in training a data analytics system involving sensitive data [8]. [14], as shown in Figure 2:

- *Data-to-Algorithm (D2A),* in which the data is sent to the data analytics algorithm and processed 'centrally', along with data from other sources. In this manner a central data set is created.
- *Algorithm-to-Data (A2D),* in which the data analytics algorithm is sent to the data source and executes 'locally' on data sources. This removes the need to transfer sensitive data.

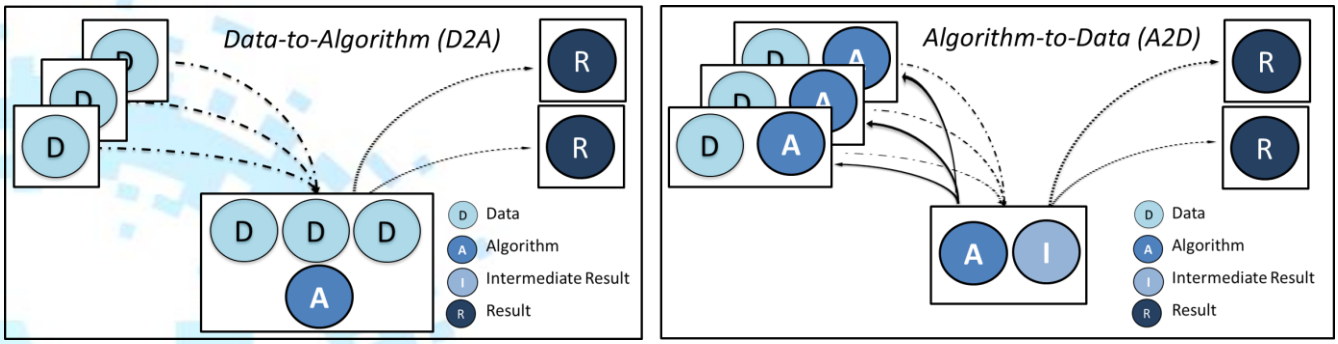

Figure 2. Collaboration models: 'Data-to-Algorithm (D2A)' (l) and 'Algorithm-to-Data (A2D)' (r).

A point of attention is that the A2D collaboration model does not fully shield the input data in all cases. For example, a trained model might contain information that can be traced back to the original data. Therefore, the algorithms should be checked or validated that they do not leak sensitive data.

Running data apps to process data locally within the domain of the data provider or Data Service Consumer is referred to as 'app enabling'. The deployment of data apps may either be instantiated and managed by the data provider or Data Service Consumer themselves or originating, instantiated and managed by external third parties. The latter is referred to as 'third part app enabling'.

Enabling third parties to deploy apps within the security domain of the data provider or Data Service Consumer requires a secure and controlled environment in a security gateway with app enabling capabilities providing the external third parties with an interface to deploy apps. Given the current state of the art in data processing it seems best to enable the data providers environment with an edge processing (computing) capability conform the (Application Container Management Layer of the) security gateway architecture as recently standardized [15]. GAIA-X seems to be a good match to provide some such functionality [16]. However, it is not clear yet what the adoption of GAIA-X will be and when implementations will become available. Alternative technologies such as OSGi [17] or cloud native container management technologies (e.g. Docker and Kubernetes) [18] could already provide a solution on the short term.

2.1.3 Data sovereignty and control based on standardized frameworks

Data sovereignty is a natural person's or corporate entity's capability of being entirely self-determined with regard to its data, i.e. it allows a legal person to exclusively decide about the usage of its data. It requires organizations to be in control over the conditions under which their data is shared and how it may be processed by other parties. Data sovereignty requires building blocks from the data sharing system to define, manage and support their data sharing policies, operational data sharing statements

and the enforcement thereof. These building blocks are required for controlling (access to and usage of) data flows. For reference, the capabilities as distinguished within the widely used XACML policy framework [19], as depicted in Figure 3, are used.

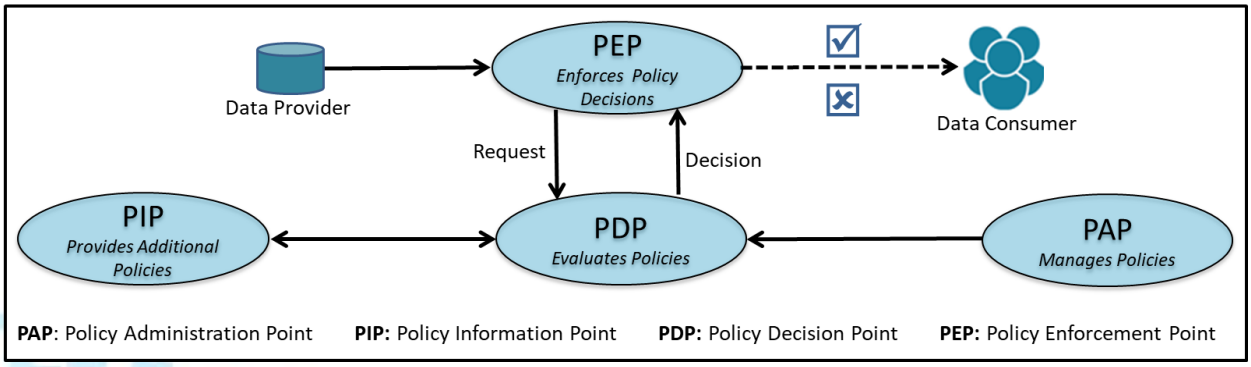

Figure 3. XACML Policy Framework.

The XACML policy framework as depicted in the figure has the following main capabilities: The *Policy Administration Points (PAP)* administer the applicable data sharing statements (referred to as 'usage contracts') and policies (referred to as 'usage rules'). The *Policy Enforcement Points (PEP)* are the triggering points at specific locations in the data sharing flow where data sharing statements should be applied and enforced, by allowing or restricting the requested sharing of data. The *Policy Decision Points (PDP)* validate the data sharing statements against the detected requests and a decision is made whether or not to allow the data sharing. The *Policy Information Points (PIP)* are used to collect additional (mostly dynamic) information on the data sharing to make accurate decisions.

2.2 Overarching system architecture

The system architecture for controlled data sharing for AI enables sharing of data between data providers and data consumers. The data providers provide the source data for the AI-algorithm. The data consumers provide the AI-algorithms.

From a data provider perspective, data control capabilities for data sovereignty, trust and security are necessary to prevent the misuse of shared data and are, as such, sine qua non conditions for organizations to share potentially sensitive information [5]. Therefore, various (types of) generic capabilities are required to enable a process for controlled data sharing between data providers and data consumers following an open network model approach as described in the previous sections.

The following paragraphs describe how the required generic capabilities jointly define the overarching system architecture for controlled data sharing for AI.

2.2.1 Reference architecture: IDS

Two main reference architectures are emerging that adopt the architecture principles for controlled data sharing for AI as addressed in the previous section. Their associated technologies are currently maturing:

- *International Data Spaces (IDS)*

  IDS is currently gaining major international traction for realizing an open network model approach for multi-lateral data sharing with infrastructural data sovereignty capabilities, following the architecture principles for controlled data sharing as addressed in the previous section. The IDS

reference architecture [20] is aimed at enabling the trusted sharing of sensitive data, whilst maintaining sovereignty. It can be considered an architectural elaboration of a zero trust architecture [21]. It is based on the network-model architecture principles as described in paragraph 2.1.1, with peer-to-peer data sharing with local data storage and processing in a federated and open infrastructure for support services.

An initial IDS implementation in the Netherlands is the Smart Connected Supplier Network [22], a field-lab initiative of Brainport Industry to enable improved cooperation in the supply chain of many companies behind large high-tech companies in the Eindhoven area.

- *SOLID (SOcial LInked Data)*

  SOLID, is a proposed set of conventions and tools for building decentralized social applications based on linked data principles [23]. SOLID is m**o**dular and extensible and it relies as much as possible on existing W3C standards and protocols.

  The AMsterdam data Exchange [24], a data space initiative of the Amsterdam Economic Board for enabling local and (inter-)national collaboration on an open data market uses the SOLID approach.

IDS is currently (aimed at) becoming part of the EU Data Strategy and is primarily addressing data sharing between organizations in business contexts (and not on individual consumers). The IDS architecture has recently been standardized [15]. Therefore, IDS will be used as reference architecture for the system for controlled data sharing for AI. This will be done while adhering to the standardized (XACML) framework for data sovereignty and control based as described in paragraph 2.1.3. The overarching system architecture will be further elaborated in the following paragraphs.

2.2.2 Roles, functional areas and building blocks

To allow data providers and data consumers to share data in a trusted and secure manner, while data providers maintain sovereignty over their sensitive data, a set of enabling generic capabilities is needed. Such generic capabilities can be provided as services according to the open network model approach. The generic services are referred to as 'building blocks', the providing organization as 'role'. As building blocks are developed as self-contained services, they can functionally extend the generic capability they provide and adapt to new technologies without major changes to the overarching system architecture. Moreover, they are not restricted in the technology used for the internal implementation of their building block.

Building blocks can provide different types capabilities in the system architecture. Hence, the building blocks can be grouped into functional areas. As depicted in Figure 4, a functional area for data sharing domain control and three functional areas for data sharing operations support are distinguished.

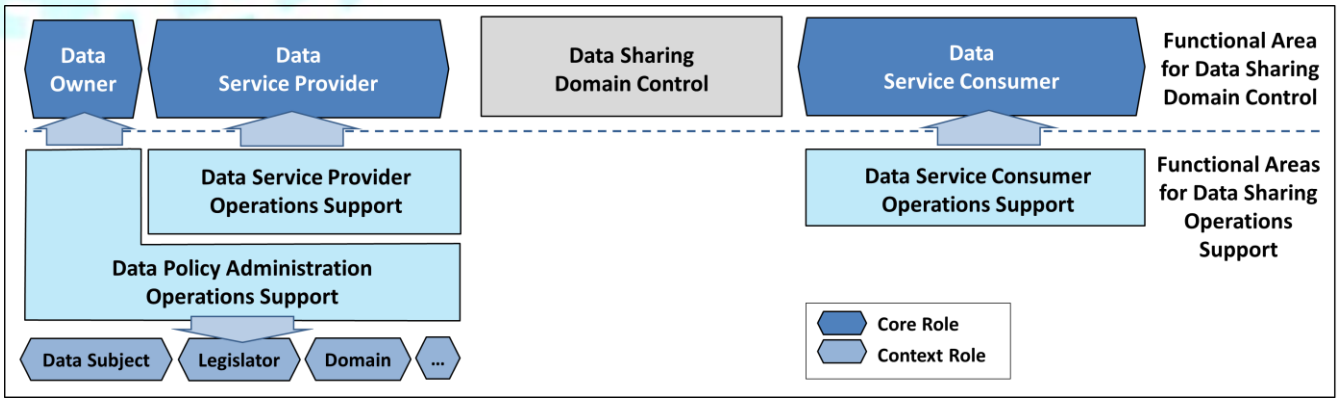

Figure 4. Functional areas for grouping the building blocks for controlled data sharing.

It is to be noted that in the figure the data provider and the data consumer are referred to as 'Data Service Providers' and 'Data Service Consumer', respectively. This indicates that they manage the availability and processing of data in a service oriented manner. Furthermore, roles and building blocks are referred to with capital letters. This notation will be used throughout the remainder of this report.

The Data Owner, Data Service Provider and Data Service Consumer are 'core roles' as they bear the primary responsibility when sharing data. The figure also shows context roles. Context roles represent stakeholders which are relevant for a Data Service Provider for obtaining the necessary permission information to be legally allowed to share data. They may be Data Subjects, Legislators and Domains.

Table 1 provides a description for both the core roles and the context roles.

Role	Description
Core roles	
Data Service Consumer	A core role in the data sharing system architecture that requests and uses data provided by a Data Service Provider. The Data Service Consumer may be a machine or a human (person). The Data Service Consumer can be an Entitled Party or perform this role on behalf of and authorised by another Entitled Party.
Data Service Provider	A core role in the data sharing system architecture that exposes data sources and provides data to a Data Service Consumer. The Data Service Provider must have explicit consent of the Data Owner to provide the data. The Data Service Provider may be an enterprise or other organization, a data marketplace, an individual, or a "smart thing".
Data Owner	A core role in the data sharing system architecture that owns the legal rights for, and has complete control over, the data it makes available. It defines the terms and conditions of use of its data and is responsible for the data, including being accountable for the quality of the data. The Data Owner is an identifiable natural person (within an organization).
Context roles	
Data Subject	A context role in the data sharing system architecture being an identifiable natural person to which data pertains.
Legislator	A context role in the data sharing system architecture being a governmental body with the mandate to make or change laws.
Domain	A context role in the data sharing system architecture representing a co-operating group of organizations with common interests and being allowed to define polices and rules on data sharing on their behalf.

Table 1: Core and context roles in the ecosystem for controlled data sharing

The following paragraphs elaborate the individual functional areas and the building blocks that they contain.

2.2.3 Functional area for data sharing domain control

The *functional area for data sharing domain control* contains building blocks for intermediation between data providers and data consumers within a (single) data sharing domain. The building are key enablers for (bilaterally) executing data sharing transactions. They provide essential intermediary functions for controlling the trustworthy and secure data sharing between a data provider and a data consumer within a data sharing domain. As such, they form the basis for the 'Trust Framework' between Data Service Providers and Data Service Consumers.

Figure 5 provides the building blocks for each the functional areas as derived from in IDS reference architecture [20], with Table 2 providing a description of the building blocks.

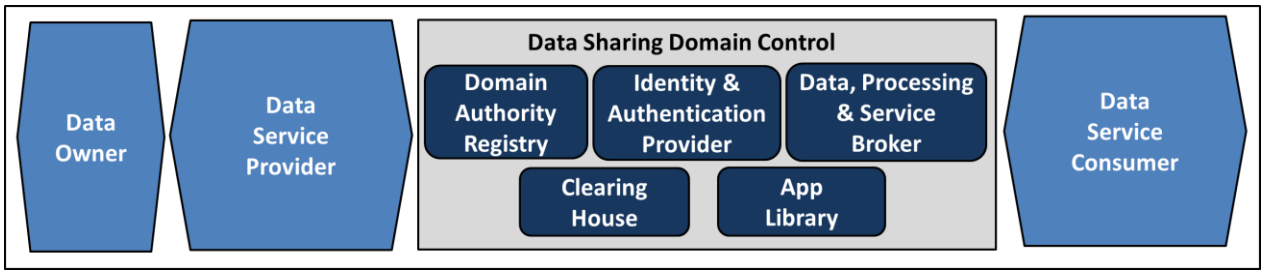

Figure 5. Building blocks in the functional area for data sharing domain control.

Building Block	Description
Functional Area: Data Sharing Domain Control	
Domain Authority Registry	A building block that provides central registration and information point for the (status of the) participating roles within a specific data sharing domain. It manages the master data and information on domain membership status, security profiles, certification status and applicable legal agreements.
Clearing House	A building block that provides clearing and settlement services for all financial and data sharing transactions, including conflict resolution and support for data sharing transactions requiring non-repudiation.
Data, Processing & Service Broker	A building block that manages a metadata repository that provides information about the data sources, the (processing) resources and services available in a data sharing domain.
Identity and Authentication Provider	A building block that enables identification of parties in a data sharing domain (which can be a machines, persons or companies) and authentication of the registered identities, e.g. using certificates.
App Library	A building block that provides a secure platform for registering and distributing data apps. It may feature different search options, e.g. by functional or non-functional properties, pricing model, certification status, community ratings, etc..

Table 2: Building blocks for the functional area for data sharing domain control.

2.2.4 Functional areas for data sharing operations support

The *functional areas for data sharing operations support* contain building blocks that help the core roles in fulfilling the functions as needed for controlled sharing of their data. These building blocks are attributable to a specific core or context role, i.e. they don't fulfill an intermediation function between core roles. As such, the building blocks in these functional areas are optionally. Moreover, the core roles could implement these activities by themselves. However, these building blocks may prevent the core and context roles from an extensive effort to make their own implementations and as such lower the barriers to participate.

As Figure 4 shows, three functional areas for data sharing operations support are currently distinguished:

- *Data Service Provider Operations Support:* This functional area contains the building blocks that provide data providers the capabilities in a data sharing system infrastructure for managing their data sharing policies and transactions, e.g. for providing the lawful ground for being allowed to share data, for providing machine-interpretable usage contracts (with access and usage control statements), for enabling data apps to be locally executed and for administering data transactions.

- *Data Service Consumer Operations Support:* This functional area contains the building blocks that provide data consumers (e.g. data analytics providers) the capabilities in a data sharing system infrastructure for managing their data sharing policies and transactions, e.g. for delegating rights, enforcing usage contracts, for accounting compliance to (agreed upon) contracts and for enabling data apps to be locally executed.
- *Data Policy Administration Support:* This functional area contains the building blocks in a data sharing system for managing the data sharing policies over the various stakeholders in the context of the data provider, e.g. for defining data sharing rules by data owners / subjects themselves (in case 'explicit' consent is required) or by entitled parties such as legal or community stakeholders (in case a form of 'implicit' consent is adequate).

Figure 6 provides the building blocks for each of the functional areas for data sharing operations support, with Table 3 providing a description for the building blocks.

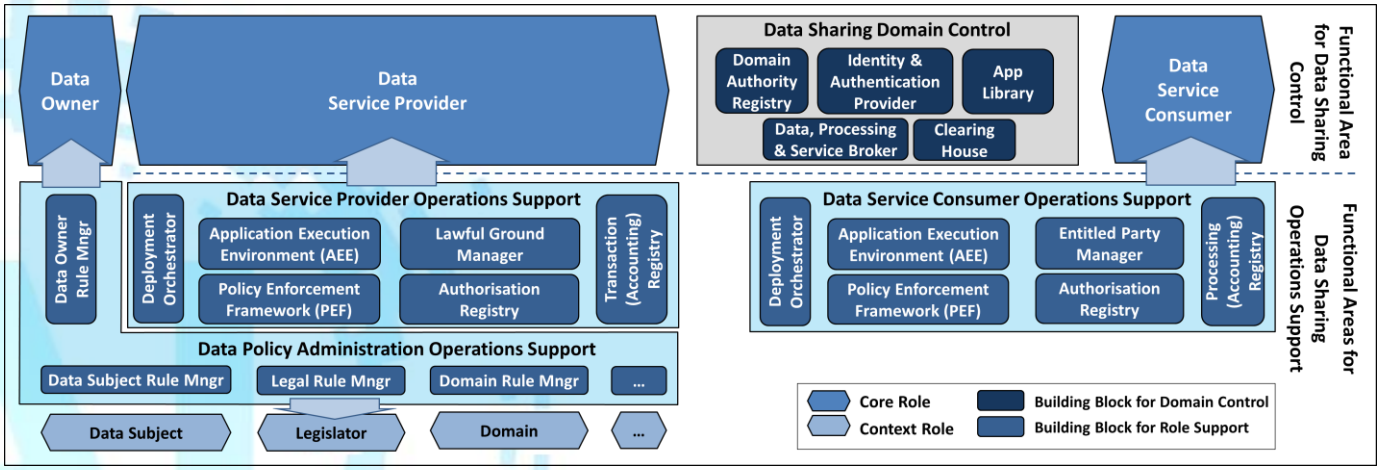

Figure 6. Building blocks in the functional areas for data sharing operations support.

Role	Description
Functional Area: Data Service Provider Support	
Application Execution Environment (AEE) Provider	A building block that provides an environment to execute data apps within the (security) domain of the Data Service Consumer. *The AEE Provider calls the PEP-function in the XACML policy framework as depicted in Figure 3.*
Policy Execution Framework (PEF) Provider	A building block that technically enforces the conditions under which data is shared (as expressed by a usage contract) within the (security) domain of the Data Service Provider. *The PEF Provider orchestrates the PEP and PDP-functions in the XACML policy framework as depicted in Figure 3.*
Authorisation Registry	A building block that registers formal technical usage contracts with specific access and usage statements, including the capability to resolve / merge multiple (potentially conflicting) data owner's or Data Subject's consent statements, using a machine-interpretable language. *The Authentication Registry fulfils the registry part of the PAP-function in the XACML policy framework as depicted in Figure 3.*
Lawful Ground Manager	A building block that manages and registers the applicable legal and policy conditions under which a Data Owner or Data Subject allows his data to be shared. It retrieves usage rules from external consent registries and supports in deriving the legal grounds to share data

	from, including resolving conflicting policies. The actual decision for releasing the data on legal grounds remains a responsibility of an authority within the Data Service Provider. *The Lawful Ground Manager fulfils the management part of the PAP-function in the XACML policy framework as depicted in Figure 3.*
Transaction (Accounting) Registry	A building block that registers logging transactions by a Data Service Provider, including the usage contracts and underlying applicable legal and policy conditions under which a provider has done the specific data transaction.
Data Service Provider Deployment Orchestrator	A building block that provides both service and deployment orchestration capabilities over an integral set of building blocks as offered by various third parties and does integration thereof into an integral service package towards a Data Service Provider.
Functional Area: Data Policy Administration Support	
Data Owner Rule Manager	A building block that manages and registers the usage rules for sharing data as defined by the Data Owner, using natural language. *The Data Owner Rule Manager fulfils (an extension of) the PAP-function in the XACML policy framework as depicted in Figure 3 by providing context role input.*
Data Subject Rule Manager	A building block that manages and registers the usage rules for sharing data as defined by the Data Subject, using natural language. *The Data Subject Rule Manager fulfils (an extension of) the PAP-function in the XACML policy framework as depicted in Figure 3 by providing context role input.*
Legal Rule Manager	A building block that manages and registers the usage rules for sharing data as defined by legislators, using natural language. *The Legal Rule Manager fulfils (an extension of) the PAP-function in the XACML policy framework as depicted in Figure 3 by providing context role input.*
Domain Rule Manager	A building block that manages and registers the usage rules for sharing data as defined by a specific domain (e.g. an group of organizations), using natural language. *The Community Rule Manager fulfils (an extension of) the PAP-function in the XACML policy framework as depicted in Figure 3 by providing context role input.*
Functional Area: Data Service Consumer Support	
Application Execution Environment (AEE) Consumer	A building block that provides an environment to execute data apps within the (security) domain of the Data Service Consumer. *The AEE Consumer calls the PEP-function in the XACML policy framework as depicted in Figure 3.*
Policy Execution Framework (PEF) Consumer	A building block that technically enforces the conditions under which data is shared (as expressed by a usage contract) within the (security) domain of the Data Service Consumer. *The PEF Consumer orchestrates the PEP and PDP-functions in the XACML policy framework as depicted in Figure 3*.
Processing (Accounting) Registry	A building block that provides the registry for logging transactions by a Data Service Consumer, including the usage contracts and underlying applicable legal and policy conditions under which a consumer has done the specific data transaction with the provider.
Entitled Party Manager	A building block that Party that registers and manages the Entitled Parties.
Data Service Consumer Orchestrator	A building block that provides both service and deployment orchestration capabilities over an integral set of building blocks as offered by the external intermediary roles and does integration thereof into an integral service package.

Table 3: Building blocks for the functional areas for data sharing operations support.

An elaboration in detail for the building blocks in each of the functional areas is provided in the architectural blueprint [10]. In addition, it is to be noted that various initiatives are currently (further) developing ecosystems of building blocks and roles for the controlled sharing of data, not specifically focusing on data sharing for AI. As such, the OPEN DEI white paper 'Design Principles for Data Spaces' [25] should be mentioned as it also addresses the aspects of building blocks to be developed for controlled data sharing, adhering to the principles of the EU data strategy. The system architecture as described in this paper is aligned with the approach as described in the OPEN DEI white paper.

3 PoCs on architectural concepts and technologies

New architectural concepts and technologies for controlled data sharing for AI are maturing. Therefore, three Proofs-of-Concept (PoCs) have been developed in 2020 to demonstrate and validate their potential and to identify lessons learned on effective and efficient system operations to enable large scale adoption.

The three PoCs have been defined such that they address complementary architectural concepts and technical components of the system architecture, as depicted in Figure 7.

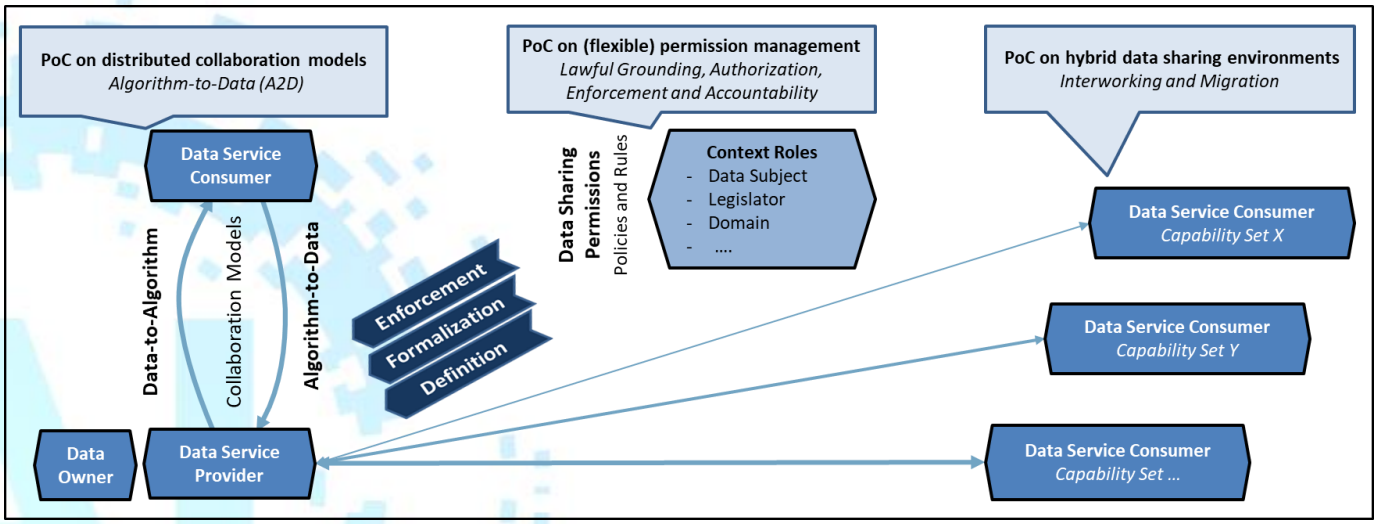

Figure 7. Three PoCs on complementary architectural concepts and technical components for controlled data sharing for AI.

The following sections subsequently address each of the three PoCs in 2020 as depicted in the figure: their background, goals, an illustrative representative use case and the lessons learned. The lessons learned are identified from the perspective for further developing the NL AIC system architecture towards operationalization. They provide input for the NL AIC working group Data Sharing in further developing the system architecture.

The (demonstration of the) PoCs is publicly available [26].

3.1 PoC on (flexible) permission management: lawful grounding and accountability

Concerns about trust, safety and lack of control over the use of available data currently hamper large-scale data sharing [5] [6]. All of these slow down the development and operationalization of new data analytics applications. As such, permission management (encompassing the aspects of lawful grounding and accountability) is of major importance when developing data sharing infrastructures to support AI.

3.1.1 Background

Data Service Providers need a lawful ground for being allowed to share potentially sensitive data. A permission management process takes care of obtaining the lawful ground. For data analytics a permission management process can have various complexities:

- The data to be provided might be (privacy) sensitive data from a multitude of Data Subjects For instance, hospitals have medical data related to patients, energy supplier have energy consumption data from their customers. For being allowed to share the sensitive data, these Data Service Providers need a lawful ground. Hence, a permission management process is need so that authorizations to share data may are delegated from the Data Subjects to the actual Data Service Providers, either through explicit consent or through other legal methods of delegation.
- Processing by a data analytics algorithm derives new information from source data, which can then be shared with other parties. The permission management architecture must be able to proliferate the required authorizations and accountability in the processing chain.

These complexities in permission management become even more complex taking into account that various collaboration models may be deployed, i.e. the algorithm-to-data (A2D) and data-to-algorithm (D2A) collaboration models as elaborated in paragraph 2.1.2 and depicted in Figure 2. Moreover, legal conditions and ethical opinions on sharing sensitive (e.g. personal) data are subject to societal debate and are expected to vary per sector (domain) and may change over time.

In addition, Data Service Providers and Data Service Consumers are accountable for the lawful ground for sharing data and for compliance with both legal and organizational conditions when sharing data. It is essential to be able to respond to questions or complaints on the processing of data. Accountability measures are needed on how the results came about and based on which data, requiring logging and traceability. Traceability can range from transparent information that provides high-level insight through a detailed description of all data, data analytics models and configurations, to the (automatic) reproduction of the delivered result. Legal accountability and compliance of data sharing for data analytics hold for instance on privacy sensitive data for which the GDPR requires absolute transparency.

Accountability also holds for processing data. Legal conditions apply when processing sensitive data, e.g. by AI-algorithms. For instance, the EU GDPR legislation prescribes the obligation to keep a record of processing operations and to take appropriate security measures. In addition, processing operations involving fully automated decision-making are not allowed for decisions that have legal consequences (for example, the granting of a benefit) or other significant effects (for example, whether or not to pass a recruitment procedure). However, the GDPR offers a number of exceptions, e.g. if there is explicit consent from the Data Subject or if the processing is necessary due to an important social interest.

3.1.2 Goals of the PoC

Taking into account the complexities as described in the previous paragraph, a generic permission management architecture is needed to support a broad variety of differing cases for managing lawful grounding and accountability. The permission management architecture must provide adequate flexibility to deal with situations of changing ethical, regulatory and organizational policies. To address the challenges of flexible permission management, the main goals of this PoC are:

- To demonstrate and asses the suitability of the set of building blocks in the functional areas for Data Service Provider Support and Data Policy Administration Support (as depicted in Figure 6) to jointly provide permission management processes with adequate flexibility for dealing with different and changing situations of ethical, regulatory and organizational policies. Therefore, the (interaction between these) building blocks must enable (1) defining data sharing rules / policies in natural language by the context roles as input for deriving the legal ground and (2) the automated

translation thereof into machine-interpretable data sharing statements (i.e. a 'usage contract' with access and usage control statements) to be used and enforced in individual data transactions. Asking for explicit consent to the Data Subject for sharing privacy-sensitive data for various purposes may be an initial option and may even be the only one under current legal conditions. However, this may not always be the desired approach, e.g. due to complexities for data subjects to grasp the essence consequence of the consent being requested and the multitude of consent requests. Therefore, alternative legal policies are expected to emerge as lawful ground for Data Service Providers to share data, e.g. provided by context roles such as the legislator or the domain.

- To demonstrate the technical enforcement of these machine-interpretable usage contracts when performing actual data sharing transactions.

3.1.3 Illustrative and representative case from the government sector

The PoC uses an illustrative and representative case from the government sector. Various Dutch governmental organizations collect data from citizens to fulfill their duties and provide services to society. For instance, the Dutch Tax Authority ('Belastingdienst') collects information to determine the amount of tax an individual or organization is due and to avoid tax evasion practices. The Dutch 'Centraal Justitieel Incassobureau' (CJIB) collects fines and as such gathers data in order to facilitate its duties. As described by the GDPR [27] (article 6, paragraph 1 and article 9, paragraph 2), these governmental organizations are allowed to collect the data without explicit consent from citizens as the data is necessary for their operational process. Consent is implicitly and irrevocably given, as long as the data is used exclusively for what is necessary according to law and regulations.

The GDPR forbids data usage or data sharing of this personal data by these governmental organizations outside the scope of the organization's primary duties. However, data sharing among different governmental organizations could lead to solutions to problems that are otherwise difficult to tackle. For instance, this applies for debt prevention and associated target group approaches ('doelgroepbenadering'). The government is obliged to offer support to citizens that are dealing with severe debt. However, it would be better and cheaper for all parties involved if such situations could be prevented by early detection of and acting upon potential debt situations. This can be achieved by sharing data between the various governmental organizations as mentioned above to feed AI-algorithms that detect individuals that are at risk of running into debt problems. Local governments can use the results to offer support to these individuals to solve their debt problems in an earlier stage.

Asking for explicit consent by the governmental organizations to the citizens for sharing its data for such purposes may be an initial option (and maybe the only one under current legal conditions / legislation). However, this may not always be the desired approach due to complexities of the essence of what consent is being asked for, the multitude of the associated consent requests and difficulties for citizens to grasp the consequence of the consent being requested. Moreover, ethical opinions and legal conditions on sharing such sensitive (personal) data are subject to societal debate and are expected to vary per sector / application area. Moreover, they may change over time. Therefore, this PoC demonstrates the generic permission management architecture based on the set of building blocks in the functional areas for Data Service Provider Support and Data Policy Administration Support (as depicted in Figure 6). It assesses its suitability in providing adequate flexibility for dealing with varying and changing situations of ethical, regulatory and organizational policies.

The PoC builds upon the 'data-to-analysis' collaboration model, in which an AI-algorithm provider (CBS) gathers and combines data from different sources (Belastingdienst, CJIB and CAK) and shares the analysis result with a third party, i.e. a local municipality ('Gemeente'). The PoC demonstrates the required processes for lawful grounding and accountability. The PoC includes two main permission management processes provided by the various building blocks and adhering to the standardized

XACML Policy Framework as described in the architectural principle in paragraph 2.1.3 and shown in Figure 3:

- *The fulfillment process (configuration)*

  The various building blocks for managing data usage rules in natural language (i.e. Subject Rule Manager, the Community Rule Manager and the Legal Rule Manager) make the formulation and provisioning of usage rules easier by enabling usage rules to be defined in natural language. Any changes in usage rules invalidates any formal usage contract derived from and referring to that specific usage rule. Any analysis done, prior to the changes in the usage rules and based on the invalidated usage contract, stay valid.

  The following process steps are taken in the fulfillment process: (1) The Rule Manager building blocks manage and register the data sharing rules (defined in natural language) for the various context roles, e.g. the Data Subject, Legislator and Domain. (2) The Lawful Ground Manager building block retrieves the data sharing rules of the various context roles. Among its task is to resolve conflicting usage rules. When necessary, the 'Lawful Ground Manager' may contact the data owner directly for asking explicit consent. The Lawful Ground Manager translates the usage rules into a proposed set of usage contracts. (3) The derived usage contracts are formally approved Data Owner acting as main authority for the Data Service Provider and serve as input for negotiation with Data Service Consumers. (4) In case the usage contract is formally agreed upon, it is administered in the Authorization Registry to be used in the individual data transactions.

- *The data transaction process (usage)*

  In the data transaction process, a Data Service Consumer requests data from a Data Service Provider.

  Subsequently, the following process steps are taken in the fulfillment process: (5) The Policy Enforcement Framework (PEF) building block orchestrates technical enforcement for the data request. It asks the Authorization Registry to resolve an applicable usage contract for the requested data. If no valid usage contract applies, access to the data is denied. (6) The usage contract, the usage rules where it is based upon and the individual data transactions are stored in the Transaction (Accounting) Registry for reporting and conflict resolution. To be seamlessly usable by data apps, it requires aligned building blocks and automated real-time processes to minimize deployment efforts.

Figure 8 gives an overview of the involved building blocks, technical components and the process steps (as numbered above).

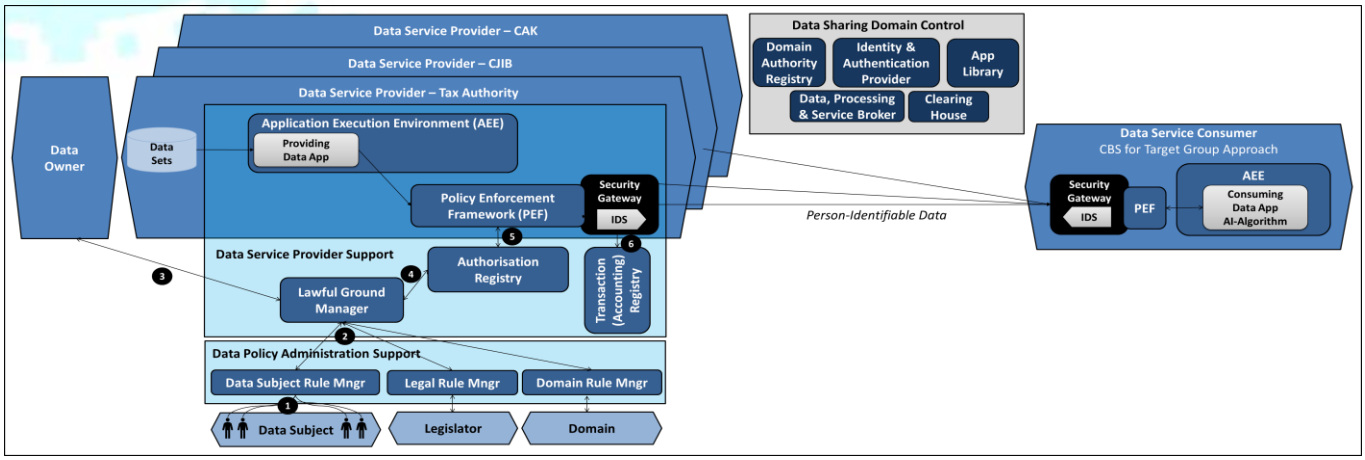

Figure 8. Building blocks, technical components and process steps for both the fulfillment process (configuration) and data transaction process (usage) for the government sector PoC.

3.1.4 Lessons learned for system operations

This PoC and its use case from the government sector have resulted in the following lessons learned for further developing the NL AIC system architecture towards operationalization:

- *Support flexible permission management structure through a set of consistent building blocks.*

  A large variety of permission management cases in different sectors requires an architectural approach that is on the one hand flexible enough to support the various cases and on the other hand is easy to use for the involved parties. The set of building blocks as elaborated in the PoC provides this flexibility, enabling a broad set of Data Service Providers to obtain lawful ground for sharing sensitive data.

- *Design for standardized API's.*

  The system architecture is to be used for a multitude of types of data sharing, use cases and sectors. Being able to simultaneously support various data sharing applications in an ecosystem of collaborating roles, the building blocks and their interfaces may need to be standardized. Initiatives like IDS already have specified some of the APIs and some reference implementations available. Thus far, GAIA-X only has defined a high level reference architecture [28] and is not in a phase yet to define the APIs.

- *Automate the translation of abstract usage rules into enforceable usage contracts.*

  To support the various context roles in stating applicable usage rules, the corresponding Subject, Legal and Community Rule Manager building blocks allow usage rules to be defined in natural language/ This makes the formulation and provisioning of usage rules easier. The Lawful Ground Manager building block retrieves these usage rules and translates them into a (proposed) set of technical usage contracts in machine readable format to be used in the policy enforcement process. Automation of the translation process will considerably improve process efficiency. However, it requires specific expertise to avoid compromising legal conditions in the automated translation process.

3.2 PoC on distributed collaboration models: 'Algorithm-to-Data (A2D)'

Due to their ability to derive complex relationships, AI-systems are ideally suited to analyze many different data sources simultaneously. However, the different data sources for AI-systems cannot always simply be brought together. Either because the amounts of data are too large, or other reasons like confidentiality. Think of privacy restrictions due to GDPR or company confidentiality, for example. Such reasons imply that data should remain at its source not to be transferred to other organizations, and it is necessary for the AI-algorithm to be transferred to the data (instead of vice versa).

3.2.1 Background

As described in paragraph 2.1.2 and depicted in Figure 2, privacy preserving techniques (such as Federated Learning (FL) and secure Multi-Party Computation (MPC)) are becoming available to design distributed AI-systems in which the data to be processed does not have to be gathered into a single database or location. They use an 'Algorithm-to-Data (A2D)' collaboration model: the AI system is sent to the data source and runs in the domain of the Data Service Provider. Hence, for running the AI-

algorithm the (potentially sensitive) source data doesn't have to be shared by a Data Service Provider with external parties: they are able to learn 'locally' from source data.

With the maturing of these privacy preserving two basic 'collaboration models' are to be considered in designing and training an AI-system involving sensitive data: 'Data-to-algorithm (D2A)' and 'Algorithm-to-data (A2D)'.

3.2.2 Goals of the PoC

As indicated, FL and MPC are maturing as 'Algorithm-to-Data (A2D)' collaboration model. However, it is thus far unclear how the promising FL and MPC technologies are positioned in relation to data sharing reference architectures that are currently emerging in the context of the EU Data Strategy, more specifically the International Data Spaces (IDS) [20] initiative and the FAIR principles [29] (Findable, Accessible, Interoperable and Reusable). Therefore, the main goals of this PoC are:

- To demonstrate that privacy preserving technologies based on de 'Algorithm-to-Data (A2D)' collaboration model can work in combination with IDS and its security gateway (e.g., IDS-connector), and that they are complementary and reinforce each other's functions in providing data sovereignty in an overarching system architecture.
- To demonstrate the FAIR data principles in combination with IDS approach by showing that data from FAIR data stations can be made Findable and Accessible by publishing their metadata through IDS data brokers, using the concept of Resources within the IDS information model.

To address the goals of this PoC, a federated learning application is implemented. Federated learning is a distributed machine learning approach that prevents the need of sharing sensitive data. In federated learning, there are multiple Data Service Providers in a 'network' that each have their own set of data. A 'centralized' federated learning server initiates and orchestrates the learning process. Roughly speaking, federated learning works as follows: (1) Data Service Providers train the same machine learning algorithm using its own machine learning model on its own data set, containing only their own (sensitive) data, (2) after some training epochs, the individually trained model is sent to the orchestrating federated learning server, (3) the orchestrator combines the models of all individual Data Service Providers into a single model, (4) the orchestrator sends the updated model back to the Data Service Providers, and (5) the previous steps are repeated until the training algorithm converges. In this manner, the sensitive source data that is being trained on is never shared with the federated learning server. Only the 'weights' in the federated learning model are interchanged.

3.2.3 Illustrative and representative case from the health sector

The PoC uses an illustrative and representative case from the health sector, including both the A2D and D2A collaboration model. Collaboration is of vital importance for the health sector to perform research. Especially in the field of machine learning that thrives on massive amounts of data to be trained on. For example, machine learning algorithms for image recognition can be trained to recognize tumors and get better recognition accuracy when trained on more data. Getting more data often involves data sharing between hospitals. However, in the context of data sharing in the health sector, data privacy and data ownership are two main concerns. Both concerns make collaboration between hospitals and/or third party organizations a challenging task. Preprocessing of data in the form of anonymization can be performed to solve the issue of data privacy, but it must be performed thoroughly to preserve the privacy of patients and the impact of data leakage might be large. Alternatively, privacy preserving data analytics algorithms based on an A2D collaboration model, such as FL and MPC, offer a compelling alternative.

Various complementary initiatives already exist in the health domain to improve data sharing of health data. An example with a lot of traction is around the FAIR principles [29] which provide guidelines to

improve the Findability, Accessibility, Interoperability, and Reuse of data assets. These guidelines have an emphasis on reducing the manual actions necessary to share and use data. A paradigm based on both the FAIR principles and an A2D collaboration model is the Personal Health Train (PHT) [30]. It introduced the metaphor of trains travelling from station to station. FAIR data-stations are locations where health data resides. They act as Data Service Providers in a 'network' of data stations, allowing the data to be used for data analytics. Trains consist of the algorithms that travel from station to station executing analytics based on the algorithm-to-data (A2D) collaboration model. Tracks are the technical infrastructure enabling the trains to move between the stations in a controlled manner.

In the PoC, the PHT federated learning application executes in an IDS-ecosystem. Users can upload data sets to a Data Service Provider IDS-connector. One IDS-connector is deployed as a Data Service Consumer, being the server for the federated learning algorithm (also referred to as the 'FL-researcher') that centrally coordinates the execution of the federated learning algorithm. The FL-researcher can start the federated learning process at each of the Data Service Providers using the locally available data within the Data Service Provider IDS-connector. Model weights are interchanged between the FL-researcher (Data Service Consumer) and each of the FL-workers (Data Service Providers). The sensitive source data does not leave the Data Service Provider IDS-connector.

The PoC entails both an A2D and D2A collaboration model in the health sector.

In the A2D federated learning case, hospitals are Data Service Provider. Datasets (e.g. collections of CT images) are available in the hospitals FAIR data station within the IDS-connector. Data apps within the App Execution Environment of the IDS-connector persist the data sets, forward the data to the federated learning worker data app and publish the metadata to the IDS Data, Processing & Service broker. A federated learning researcher has a Data Service Consumer role. He can search available data sets in the network by querying the broker and can initiate and orchestrate federated learning training by choosing available data sets and selecting federated learning workers, which are deployed in the hospitals IDS-connectors. The broker makes the Data Service Providers and data sets findable. The A2D case provides a concrete implementation of the Personal Health Train (PHT) over IDS. The following process steps are taken: (1) The FL-Researcher coordinates the execution of the FL-algorithm. (2) The FL-researcher can initiate the federated learning process by installing a FL worker data app in the Application Execution Environment (AEE) building block of each of the Data Service Providers. (3) The local workers data app declares its ingress and egress data flow requirements to the Lawful Ground Manager. Similar as for the previous PoC, (4) the derived usage contracts are formally approved by the Data Owner, (5) are administered in the Authorization Registry and (6) are enforced by the Policy Enforcement Framework (PEF) building block. (7) The FL worker data app can now access the locally available data. (8) Only the resulting FL model weights are shared between the FL-researcher and each of the local FL-workers, without the sensitive local data leaving the Data Service Provider's security domain.

Figure 9 gives an overview of the involved building blocks, technical components and the process steps (as numbered above).

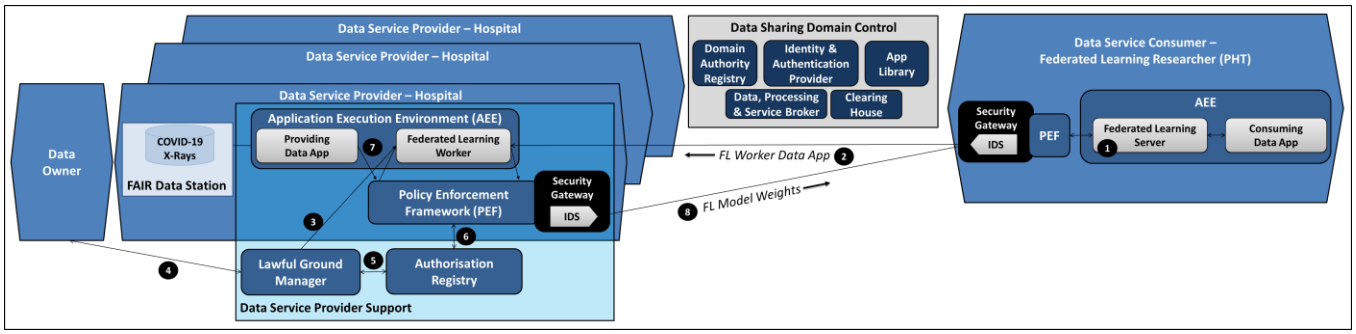

Figure 9. Building blocks, technical components and process steps for both the fulfillment process (configuration) and data transaction process (usage) for the health sector PoC.

This A2D collaboration model case has shown that various types of data can be supported. The PoC in the health sector deploys federated learning on data sets of x-ray images. In addition, it has also demonstrated the deployment on the MNIST data set [31], which is often used as a recognizable example data set for machine learning.

In addition to the A2D collaboration model case, the PoC also demonstrates a D2A collaboration model case. For the D2A collaboration model, source data is shared between a Data Service Provider hospital and a Data Service Consumer. Hospitals can create data sets that are made accessible via a Data Service Provider IDS-connector in the form of triple stores. A Data Service Consumer can send queries to a hospital's data providing IDS-connector to retrieve (sub)sets of data. This allows the Data Service Consumer to run analysis on the retrieved (sub)sets of data from the hospitals. The demo is focused on the case of knowledge sharing on COVID-19 registrations in the Virus Outbreak Data Network (VODAN) [32]. Again hospitals act as Data Service Providers. Its users can insert data in RDF-format in a triple store of the hospitals Data Service Provider IDS-connector through a user interface. All types of RDF-data is supported. For the PoC linked data is used in the eCRF semantic format for COVID-19 registrations [33]. Data Service Consumers can build a SPARQL queries on the COVID-19 registrations, providing basic tools to filter on symptoms like coughing and fever. Note that all types of data can be stored and queried as long as there is a valid ontology.

3.2.4 Lessons learned for system operations

This PoC and its use case from the health sector have resulted in the following lessons learned for further developing the NL AIC system architecture towards operationalization:

- *Support third party app enabling in a controlled manner.*

  In the PoC the federated learning workers were deployed statically. In a real world scenario this would be more dynamic, especially when Data Service Providers will allow multiple AI algorithms (simultaneous) access to sensitive data by means of FL and MPC privacy preserving techniques. For this, enabling third parties to deploy data apps within the Data Service Providers security environment has major potential. It gives external parties that want to use FL or MPC advanced options to access and pre-process data, whilst unburdening the Data Service Provider from extensive integration efforts. It therefore improves operations efficiency.

  Third party app enabling requires a secure and controlled environment to enable third parties with to deploy apps in the Data Service Provider's Application Execution Environment (AEE) building block. Given the current state of the art in data processing it seems best to view the AEE as edge processing (computing) capability. GAIA-X seems to be a good match to provide some such functionality [16]. However, it is not clear yet what the adoption of GAIA-X will be and when

implementations will become available. Alternative technologies such as OSGi [17] or cloud native container management technologies (e.g. Docker and Kubernetes) [18] could already provide a solution on the short term.

- *Integrate policy enforcement to control data sharing for apps in the execution environment.*

  Enforcement of data sharing policies and rules for specific data sharing transactions is considered a major driver for adoption of the system architecture. To be seamlessly usable by (third-party) data apps, it requires aligned processes and interfaces for matching and mapping data sharing polices and rules for the ingress and egress data flows of the data apps executing in the AEE onto the policy enforcement capabilities provided by the PEF. As far as possible this should be automated and real-time to minimize deployment efforts and costs.

- *Include processing capability in the broker building block.*

  Third party app enabling allows external organizations to deploy data apps within the Data Service Providers environment. To identify what the (processing, storage and network) resource availability of a Data Service Provider are and whether they suffice the needs to deploy a specific data app, each of these available resources should be exposed by the Data Service Provider. This can be done through a federated catalogue for brokering both data and available resources. GAIA-X interfaces may be used.

- *Support FL and MPC by means of an enabling data flow management app.*

  Third party app enabling can be facilitated with supporting data apps that allow external parties ease-of-use. Specifically, to support the most commonly used versions of FL and MPC, an enabling data flow management app can be developed that allows an FL or MPC algorithm provider to configure the required data pipeline such that the ingress and egress data flows of the data apps (e.g. the FL / MPC workers, data quality management, semantic conversions, locally executing in the AEE building block can be controlled by the Data Service Provider by means of its PEF building block. This unburdens both the Data Service Provider from extensive development and deployment efforts.

- *Ensure interoperability with the FAIR approach.*

  In the PoC, FAIR data stations can publish their metadata on available data resources in an IDS data broker. Data resources are expressed in the IDS ontology. By making use of a federated catalogue for data, processing & service brokering, the findability of resources is supported via the domain independent IDS Information Model. Some of the metadata elements specific for AI, e.g. the detailed FL data shapes, are not yet fully supported. However, the expectation is that such elements will be supported, either directly in the IDS Information Model or by extensions to the ontology. Standardized control mechanisms for access to the data provided improves Accessibility. For the new privacy preserving AI techniques, this point becomes more difficult since the data itself cannot be accessed directly but only actions on the data can be performed. For this an important role is reserved for the Interoperability and Reusability principles, which are more on the level of the data itself. As such, they are beyond the scope of the data sharing infrastructure. Nevertheless, third party app enabling will provide Data Service Consumers the capability to manage the data according to these principles.

3.3 PoC on hybrid data sharing environments: interworking

It is to be realized that the introduction of the system architecture for controlled data sharing for AI (as described in chapter 2 of this report and elaborated in the architectural blueprint [10]) will be gradual. Not all Data Service Providers and Data Service Consumers will have all (or even any) building blocks

in place at the same time. Or in short: a 'big bang' introduction of the controlled data sharing infrastructure for AI to all Data Service Providers and consumers is an utopia.

3.3.1 Background

The data sharing landscape will be characterized by hybrid data sharing environments with Data Service Providers and Data Service Consumers having implemented various sets of the building blocks. Nevertheless, this shouldn't prevent individual Data Service Providers or Data Service Consumers from being able to participate in the data sharing infrastructure and to share data for AI. Neither should they be forced to implement (all) building blocks. However, the extent to which sensitive data is shared may depend on the building blocks they do have in place.

3.3.2 Goals of the PoC

The hybrid data sharing environment poses challenges on interworking in situations in which Data Service Providers and Data Service Consumers have different capabilities. Therefore , the main goal of this PoC is to demonstrate the possibilities for data sharing for cases in which the Data Service Provider and Data Service Consumer have different data sharing building blocks with respect to identity and authentication, legal contracts, usage contracts and policy enforcement. To support such heterogenous data sharing relationships, a Data Service Provider can classify his data, distinguishing between open data and (various levels of) governed data. Data sharing decisions with specific Data Service Consumers are based upon the combination of the classification level of the data and the capabilities and building blocks they have implemented, including:

- *Being able to identify the Data Service Consumer:* In digital identification and authentication, the three main standards are SAML, (e.g. as used in e-Herkenning), OAUTH 2.0 (a Web Authorization Protocol) and OpenID Connect (OIDC, a simple identity layer on top of OAuth 2.0). In addition to having a claimed identity, the Data Service Provider may or may not have the capability to authenticate the Data Service Consumers claimed identity at its identity provider.
- *Having a legal and / or usage contract with the Data Service Consumer:* Various situations with respect to a (joint) legal contract may occur. The Data Service Provider and Data Service Consumer may or may not have a joint legal contract, e.g. when they are both member of the same data sharing domain. In addition, they may or may not have negotiated usage contract for specific data sharing transactions, which is not part of a bilateral or overarching legal contract.
- *Having policy enforcement capabilities with the Data Service Consumer:* This refers to the Policy Enforcement Framework (PEF) building block as depicted in Figure 4. Situations may occur in which both the Data Service Provider and Data Service Consumer may or may not have the PEF building blocks.

3.3.3 Illustrative and representative case from the energy sector

The PoC uses an illustrative and representative case from the energy sector. Energy related data is currently stored and managed by various organizations. As an outcome of the climate discussions in the Netherlands, it has been decided that data sharing of energy related data between organizations should be improved to stimulate new sustainability solutions. As such work has started on the casus of the 'Datastelsel Werkelijk Energieverbruik Utiliteit' (WEU) in which data from various sources is shared, including cadaster building data from the BAG (Basisregistratie Adressen en Gebouwen), usage data and energy label data. The initial approach taken for the WEU is based on a data sharing agreement framework. This is characterized by a joint (legal) data sharing agreement to be agreed upon between Data Service Providers and Data Service Consumers. Possibly, an authorization function is included for defining and enforcing usage contracts for individual data sharing transactions. This approach is

comparable to agreement framework approaches such as iSHARE for the logistics sector in the Netherlands [34].

The PoC extends upon the agreement framework approach taken for the WEU. It demonstrates how a migration / evolution path for Data Service Providers from a data sharing agreement framework approach to the system architecture as described chapter 2 can be followed, by activating the PEF building blocks for enforcing data sharing and the AEE Provider building block for executing data apps (i.e. the AEE Provider building block). Moreover, it shows the possibilities for supporting heterogenous data sharing relationships by means of a hybrid security gateway.

Similar to the PoC for the government sector as described in section 3.1, this PoC distinguishes two subprocesses as part of an overarching permission management process in providing various types of Data Service Consumer access to sensitive data:

- *The fulfillment process (configuration)*

  As stated in the previous paragraphs, to support heterogenous data sharing relationships a Data Service Provider must classify his data, distinguishing between open data and (various levels of) governed data. The following process steps are taken: (1) The Data Owner manages and registers the classification of the data under his responsibility by means by means of the Owner Rule Manager building blocks. It allow the data sharing rules to be defined in natural language. (2) The Lawful Ground Manager building block retrieves the data sharing rules from the registry in the Owner Rule Manager building block. The data sharing rules may be combined with the data sharing rules from various context roles, as described in paragraph 3.1.3. Among its task is to resolve conflicting usage rules. When necessary, the 'Lawful Ground Manager' may contact the data owner directly for asking explicit consent. The Lawful Ground Manager translates the usage rules into a proposed set of usage contracts. (3) The derived usage contracts are formally approved Data Owner acting as main authority for the Data Service Provider and serve as input for negotiation with Data Service Consumers. (4) In case the usage contract is formally agreed upon, it is administered in the Authorization Registry to be used in the individual data transactions.

- *The data transaction process (usage)*

  In the data transaction process, a Data Service Consumer requests data from a Data Service Provider. The Data Service Provider bases the sharing decisions with specific Data Service Consumers upon the combination of the classification level of the data being requested and the capabilities and building blocks the Data Service Consumer has implemented. The following process steps are taken: (5) The Data Service Consumer requests data from a Data Service Provider. (6) Based on the capabilities and building blocks that the Data Service Consumer has implemented, the security gateway does a validation of the identity provided (if any) and the protocols supported and makes a (7) routing decision for the incoming data request to the appropriated data providing app for further handling of the data request. (8) The Policy Enforcement Framework (PEF) building block orchestrates technical enforcement for the data request. (9) It asks the Authorization Registry to resolve an applicable usage contract for the requested data. Based on the usage contract the PEF decides whether or not the data request can fulfilled and whether (10) the requested data is shared with the Data Service Consumer to the of the If no valid usage contract applies, access to the data is denied. (11) The usage contract, the usage rules where it is based upon and the individual data transactions are stored in the Transaction (Accounting) Registry for reporting and conflict resolution.

Figure 10 gives an overview of the involved building blocks, technical components and the process steps (as numbered above).

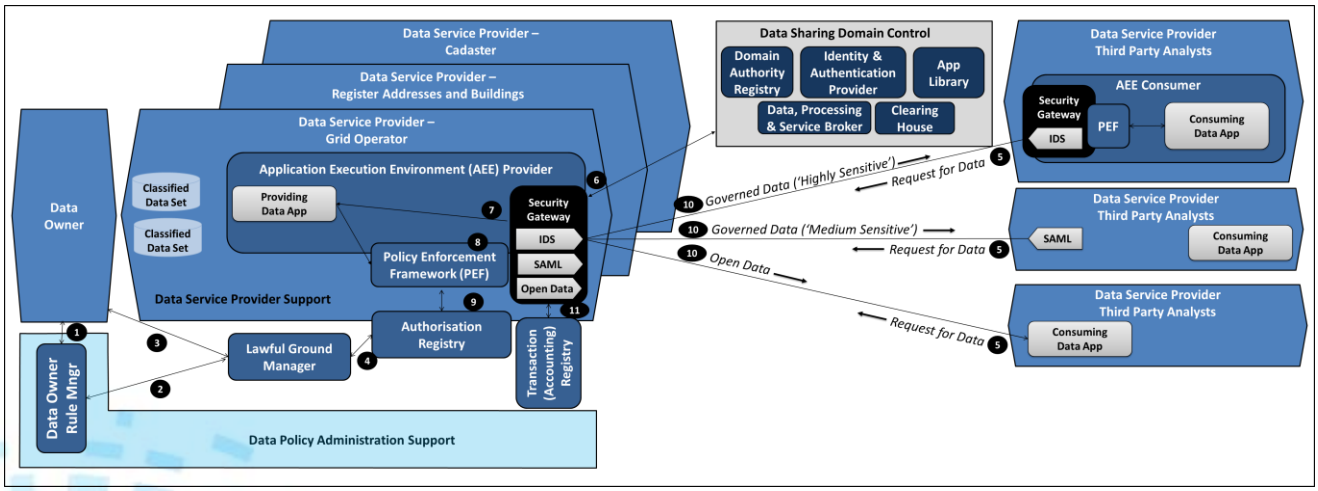

Figure 10. Building blocks, technical components and process steps for both the fulfillment process (configuration) and data transaction process (usage) for the energy sector PoC.

As the figure illustrates, the PoC includes the roles for various Data Service Providers and Data Service Consumers, The latter support varying protocols with respect to identification and authentication. Based on the differing identification and authentication capabilities of the Data Service Consumers, the PoC shows that Data Service Providers can deploy different rules and policies for actually sharing data.

3.3.4 Lessons learned for system operations

This PoC and its use case from the energy sector have resulted in the following lessons learned for further developing the NL AIC system architecture towards operationalization:

- *Develop a hybrid security gateway to support various data sharing architectures*

  To support hybrid data sharing environments as described in the previous paragraphs by means of a single entry point for a Data Service Provider a security gateway should be developed that allows multiple endpoints to accept messages from different data sharing schemes and routes these to a single data app. This is referred to as a hybrid security gateway. Different data sharing policies can be defined by a Data Service Provider based upon the combination of the classification level of data and the available capabilities of a Data Service Consumer, e.g. with respect to identification and authentication, authorization and / or policy enforcement. The various data sharing policies can be enforced on both ingress and egress data flows through the PEF building blocks.

  This approach also enables migration for both Data Service Providers and Consumers from more basic data sharing architectures towards the overarching system architecture as described in this report with more elaborate capabilities.

- *Strive for governance alignment to allow seamless interworking between data sharing initiatives*

  Apart from the technical embedding of different data sharing architectures, the governance alignment will become ever more important. Especially when a tighter integration is desired between different data sharing initiatives. This for instance applies to membership of various data sharing initiatives and the legal conditions associated with it. An approach in which Data Service Providers and Consumers have to register at each data sharing initiative separately would introduce (too) extensive management and integration efforts. A more optimal solution would be that different initiatives come to an agreement such that registering at a single scheme is sufficient for interworking with the other initiatives as well. This does require alignment on the governance structures.

4 Lowering the barriers for adoption: interoperability, migration and interworking

The required basic architectural concepts and technical components for realizing the system architecture for controlled data sharing for AI as presented in this paper are rapidly maturing. They are currently being introduced. This is exemplified by the Smart Connected Supplier Network (SCSN [22]) SCSN is an IDS-based data space initiative of Brainport Industry in the Eindhoven area in the Netherlands to enable improved cooperation in the supply chain fir high-tech companies.

Hence, the successful introduction is clearly within reach from the technical perspective. Nevertheless, its widescale realization will benefit greatly from having a gradual migration path for Data Service Providers and interworking and interoperability between data sharing initiatives. Therefore, these operational aspects are addressed in the following sections.

4.1 Interoperability: towards federation of data spaces

There will not be a single data sharing infrastructure for AI. Individual sectors or communities are expected to develop their own instance of a data sharing infrastructure for AI, preferably in accordance with the system architecture as described in chapter 2. This will result in multiple data sharing domains for AI.

Being able to seamlessly share data over these individual domains yields clear advantages. It extends the reach and scope of accessible data that may be used for AI-algorithms and allows AI solutions to be developed across sectors and regions. Therefore, interoperability between data sharing domains adds major value.

An approach to systematically address the interoperability challenges is provided by the new European Interoperability Framework as developed by the European Commission [35] and depicted in Figure 11.

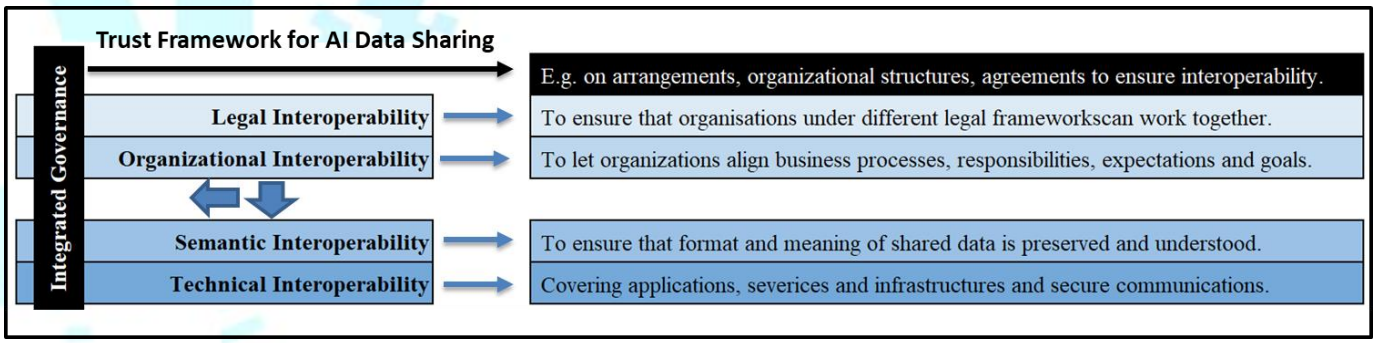

Figure 11. Interoperability model as defined in the New European Interoperability Framework [35].

The framework distinguishes four interoperability levels (legal, organizational, semantic and technical interoperability) under an overarching integrated governance approach. To enable interoperability between data sharing domains for AI, each of the interoperability levels as indicated in the figure are addressed in the following paragraphs.

4.1.1 Technical interoperability

Technical interoperability covers the applications, services and infrastructure for data sharing. Three main technical interoperability aspects need to be considered:

- *Handshake protocols: (hybrid) data sharing environments*

  The basis of secure data sharing is formed by the end-to-end secure handshake protocol. It handles aspects such as secure peer-to-peer connectivity and identification and authorization.

  From a Data Service Providers perspective, it is noted that not for all sharing of data with AI-algorithm providers a same and high level of security will be required. Supporting various handshake protocols can make his data available in an easy manner to a larger set of AI-algorithm providers. For instance, when open data is shared or when data has been anonymized it may already be used by a broad set of AI-algorithm providers without all 'heavy-weight' control and security measures. A 'light-weight' handshake protocol may be sufficient for being allowed to access the data., e.g. based on basic 'https'. Alternatively, if potentially sensitive data is shared, more heavy-weight secure handshake and identification protocols may be required with advanced end-to-end data control and security capabilities, e.g. based on the IDSCP secure handshake protocol which has recently been standardized for the IDS security gateways (IDS-connectors) [15].

  Therefore, a hybrid data sharing environment is to be supported with interoperability between various handshake and security gateway protocols. Its potential has been demonstrated in the PoC on hybrid data sharing environment as described in section 3.3.

- *Identity and authentication: legal and domain membership*

  Within a data sharing domain for AI, identification and authentication are done at two levels:

  - As *legal identities,* to uniquely identify and authenticate natural persons, organizations or software components as legal entities. For legal identification and authentication, a separate building block 'Identity and Authentication Provider' is included in the reference architecture as depicted in Figure 5.
  - As *domain members,* to administer and to continuously check that the identified and authenticated legal entities are actually registered as member of a data sharing domain, and as such adhere to the legal agreements as agreed upon within the domain. Identification of domain membership is part of the 'Domain Authority Registry' building block in the reference architecture as depicted in Figure 5.

  For identification and authentication across data sharing domains, it needs to be decided at which (or both) of the levels interoperability is required and the protocols and interfaces to do so.

- *Data, processing and service brokering: federated catalogue*

  Data, processing and service brokering entails registering and managing metadata on the data, processing and service resources available in individual data sharing domains. These capabilities are provided by the 'Data, Processing and Service Broker' building block in the reference architecture as depicted in Figure 5. Its activities focus on receiving and providing metadata on available resources. It provides an interface for Data Service Providers to send their metadata. The metadata is stored in an internal repository for being queried by Data Service Consumers.

  Multiple 'Data, Processing and Service Brokers' may co-exist at the same time, both within and across data sharing domains. To make their registered resources searchable and available across domains, a 'federated catalogue' approach is required. The federated catalogue consists of a federation of different catalogues of resources which are joined together in a standardized method, virtually acting as a single overarching data, processing and service broker over various data sharing domains.

  Recently the EU project EU Hubs for Data initiative [36] has started with the objective of Common European Data Spaces, based on a federation of regional data innovation hub / data space initiatives. It will address the topic of the federated catalogue.

4.1.2 Semantic interoperability

For semantic interoperability, it may be obvious that a shared and common semantic data model to be jointly used by Data Service Providers and Data Service Consumers has major advantages in minimizing complexity for interconnection and collaboration. However, such a jointly used common semantic data model will appear to be an utopia. Therefore, mechanisms for semantic conversion need to be supported in the system architecture for controlled data sharing. Enabled by the (Application Container Management Layer of the) security gateway architecture as recently standardized [15], this may be taken care of by means of semantic management data apps executing in the Application Execution Environment (AEE) in the security gateway. Semantic management data apps may be developed for specific semantic conversions or for enabling easy-to-use mapping between semantic models [37].

4.1.3 Organizational interoperability

Organizational interoperability refers to the way in which the processes, responsibilities and expectations are aligned to achieve the common and mutually beneficial goals for controlled data sharing for AI, whilst meeting the requirements of the user community by making services available, easily identifiable, accessible and user-focused.

The relationship between Data Service Providers and Data Service Consumers must be clearly defined. This may need the alignment of existing business processes or define and establish new ones. In addition, it may involve instruments to formalize mutual assistance, joint action and interconnected business processes as part of the data sharing relationship, e.g. Memoranda of Understanding (MoUs) and Service Level Agreements (SLAs).

4.1.4 Legal interoperability

The aspect of legal interoperability between data sharing domains presents a major challenge. Currently, legal aspects are mainly dealt with within a single data sharing domain by pre-defining the set of muti-lateral legal agreements to which individual Data Service Providers and Data Service Consumers are bound to adhere to when signing up for joining the domain. However, this provides interoperability challenges on the legal aspects in case a Data Service Provider and a Data Service Consumer are member of different (or even none) data sharing domains, with varying multilateral legal agreements. To address this challenge, various approaches may be thought of:

- *An overarching legal framework,* to which the individual data sharing domains (and their subscribers) agree to adhere. This provides a possibility to extend the scope of the shared legal agreements over multiple data sharing domains. However, although it extends the scope, it is still limited to adhering data sharing domains and their subscribers. For, sharing data with subscribers in data sharing domains beyond this extended scope, the same legal interoperability challenge remains.
- *A legal agreement negotiation approach*, in which Data Service Providers and Data Service Consumers bilaterally negotiate the legal conditions under which they share data. For this, a strong and formalized semantic fundament is essential to make sure that various organizations operating in different sectors and jurisdictions unambiguously understand each other. A machine-readable interpretation of the legal data sharing agreement and the usage contract is required, as this enables automatic reasoning to be executed on the complex system of rules and obligations.

For NL AIC the former is the most simple to realize in the short term and effective when initiating a joint effort for controlled data sharing for AI over various domains within the NL AIC context, and is therefore to be preferred as initial step.

4.1.5 Integrated governance

To overall handle and manage the aspects of technical, semantic, organizational and legal interoperability within the context of controlled data sharing for NL AIC, an overarching integrated governance umbrella is needed. In the context of NL AIC, these goals for integrated governance are met by a joint approach on defining and adopting a common *'Trust Framework for Data haring for AI'*, based on the system architecture for controlled data sharing for AI as described in this report. The development, introduction and adoption of this common trust framework will be a major goal and ambition of the NL AIC working group Data Sharing for the coming time period 2021 – 2024.

The roadmap for the common trust framework for AI data as currently adopted by the NL AIC working group Data Sharing is further addressed in the following, concluding, chapter.

4.2 Migration: gradual evolution for Data Service Providers

For a Data Service Provider, a flexible and gradual growth path is key for adopting the single entry point for controlled data sharing. Flexibility is provided through a gradual growth path by being able to subsequently implement the building block in the various functional areas as depicted Figure 5 and Figure 6. The building blocks in the 'Domain Data Sharing Control' functional area provide the basis. It contains building blocks for finding, sharing and managing specific data sets between Data Service Providers and Data Service Consumers, including a domain authority (encompassing domain membership registration with associated legal agreements on data sharing) and building blocks on identification and authentication (security) and authorization (data sovereignty). Subsequently, the building blocks as provided by the 'Data Sharing Operations Support' functional areas may be gradually adopted which help Data Service Provider in the management and administration of capabilities for the individual data sharing transactions.

To prevent (costly) migration and integration trajectories and to stimulate adoption, two aspects need to be taken care of in the gradual migration:

- *Technical migration* in which a Data Service Provider implements an additional building block with backward compatibility. For instance, a Data Service Provider can introduce the support for brokering or enforcement building blocks without impacting existing data sharing relationships in preparation for providing more advanced capabilities for sharing data.
- *Service migration,* in which a Data Service Provider offers the advanced data sharing capabilities as enabled by the new building blocks for both new and existing data sharing relationships. For supporting the new capabilities, this may require adjusting the integration with its internal systems.

To improve the ease-of-onboarding and migration, the Deployment Orchestration building blocks as depicted in Figure 6 are expected to emerge. They lower the barriers to participate, migrate and stimulate adoption by addressing the Data Service Provider's needs and challenges on minimizing integration efforts. They provide a single, user-friendly, entry-point for subscribing, configuring and managing their connectivity to a coherent and overarching set of building blocks, supporting various types of data sharing over multiple data sharing relationships. They unburden Data Service Providers from having to deal with complex and costly integration and management efforts due to a multitude of building blocks provided by separate organizations. Moreover, they allow the building blocks as 'wholesale'. The individual building blocks don't need to know end-users. This improves data sovereignty as customer-identifying information is only available at the deployment orchestrator.

4.3 Interworking: hybrid data sharing environments

The introduction of data sharing as described in the previous chapters will be gradual. The data sharing landscape will be characterized by hybrid data sharing environments with actors having implemented various sets of capabilities and building blocks. Nevertheless, this shouldn't prevent individual Data Service Providers or Data Service Consumers from being able to participate or force them to implement (all) building blocks. The extent to which sensitive data is shared with them may depend on the building blocks they do have in combination with the sensitivity classification of the data. Development of *a hybrid security gateway* to support situations in which the implemented building blocks of the Data Service Consumer may differ from the implemented building blocks of the Data Service Provider. It allows different endpoints to accept messages in different schemes and routing these messages to a single data app, provides an environment that is both flexible with respect to ingress and egress as well as simple with respect to the actual data processing itself. However, to stay in full control, it does imply that actual sharing of potentially sensitive data by the Data Service Provider with a Data Service Consumer may depend on the building blocks that the Data Service Provider has implemented, whether the Data Service Provider and Data Service Consumer are under the same domain authority or have a joint legal data sharing agreement.

5 Conclusions

This 'GAP analysis' report has described an overarching system architecture for controlled data sharing for AI, adhering to the principles and approach as currently being developed within the EU data strategy. Taking a system operations perspective, it has identified gaps to be bridged between the architectural concepts and technical components as demonstrated in three Proofs-of-Concept (PoCs) in 2020 and effective and efficient deployment and operations thereof conform the envisioned system architecture.

The required basic technology for realizing the system architecture is maturing. Therefore, the following sections focus on the system operations and governance gaps to bridge towards large-scale deployment and adoption of the system architecture.

5.1 System operations gaps

System operations addresses the effective and efficient deployment and operations of the overarching system architecture for controlled data sharing for AI. It is a prerequisite for large-scale deployment and adoption.

In this report, the system operations gaps have been identified as lessons learned from each of the three PoCs in 2020 as described in chapter 3: the PoC on (flexible) permission management, the PoC on distributed collaboration models and the PoC on hybrid data sharing environments. The lessons learned are summarized in Table 4.

PoC on (flexible) permission management	PoC on distributed collaboration models	PoC on hybrid data sharing environments
- Support flexible permission management structure through a set of consistent building blocks. - Design for standardized API's. - Automate the translation of abstract usage rules into enforceable usage contracts.	- Support third party app enabling in a controlled manner. - Integrate policy enforcement to control data sharing for apps in the execution environment. - Include processing capability in the broker building block. - Support FL and MPC by means of an enabling data flow management app. - Ensure interoperability with the FAIR approach	- Develop a hybrid security gateway to support various data sharing architectures - Strive for governance alignment to allow seamless interworking between data sharing initiatives

Table 4: Lessons learned on system operations from each of the three PoCs in 2020.

In addition, it is expected that a multitude of data sharing domains will emerge to support AI in individual sectors or communities. Interoperability between these data sharing domains adds major value as it enables seamless sharing of data over the domains, extends the available data sets for AI-algorithms and prevents from a siloed approach. Therefore, interoperability between data sharing domains is key for large-scale deployment and adoption of the overarching system architecture. Therefore, the four levels of interoperability distinguished in the new European Interoperability Framework as developed by the European Commission have been considered: technical, semantic, organizational and legal interoperability. Moreover, interoperability between data sharing domains has

been addressed together with adjacent topics for lowering the barriers to adoption: gradual migration for data providers and interworking in hybrid data sharing environments.

Both the lessons learned and the interoperability, migration and interworking aspects provide input for the NL AIC working group Data Sharing in further developing the system architecture in the time-period 2021-2024. A further detailed elaboration is provided in the report '*Blueprint NL AIC Data Sharing System Architecture'* [10]. It will be for review in 2021 by a group of data sharing experts in the Netherlands and be periodically updated with the latest insights.

5.2 Governance gaps

The three PoCs have demonstrated the architectural concepts and technical components that are part of the system architecture. It is important to further develop the system architecture by means of illustrative and representative AI use cases and gathering additional requirements from the current market. As basis, the three individual PoCs in 2020 will be integrated into an overarching reference architecture in 2021. The reference architecture reflects the state-of-the-art for the system architecture for controlled data sharing for AI across the sectors, as described in this report. It forms the foundation for 'system PoCs' to be developed in the time-period 2021-2022, which are based on concrete use cases provided and supported by individual sectors as represented in NL AIC.

To align the organizations within NL AIC to adopt a joint strategy for developing and deploying the system architecture for controlled data sharing for AI, various governance aspects need to be addressed in addition to the architecture and technology. These governance aspects encompass topics such as business viability, legal conditions, interoperability, standards and interfacing. Jointly these contribute to the development of a common 'Trust Framework for AI Data Sharing'. Figure 12 depicts its high-level development roadmap.

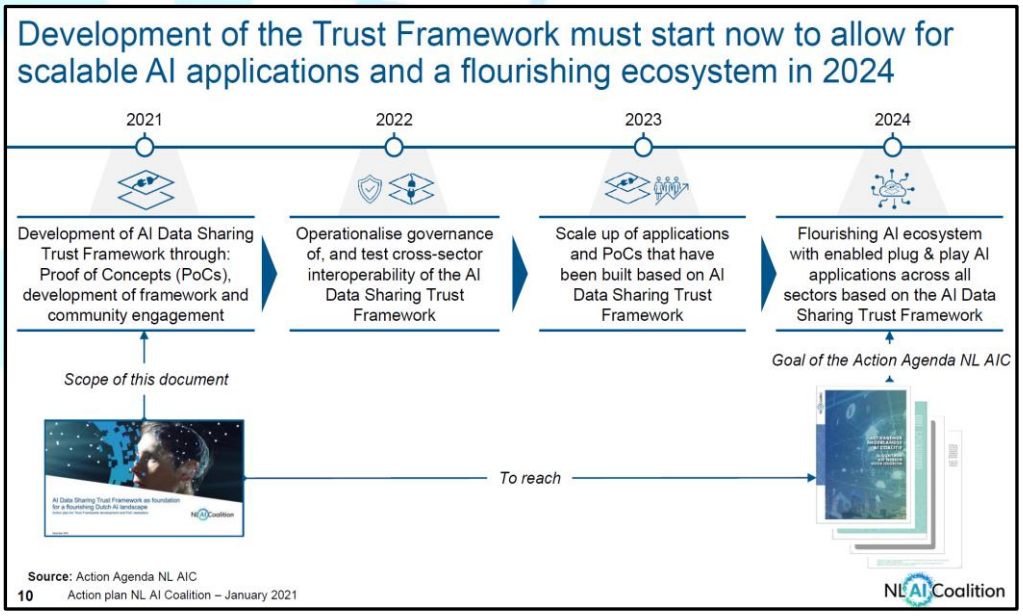

Figure 12. High-level roadmap for the development of the trust framework for AI data sharing [40].

The development, introduction and adoption of this common trust framework will be a major goal of the NL AIC working group Data Sharing for the coming time period 2021 – 2024.

References

[1] European Commission (2020). "On Artificial Intelligence - A European approach to excellence and trust". EC Communications 65. URL: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

[2] European Commission (2020). "A European strategy for data". EC Communications 66. URL: https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy.

[3] European Commission (2020). "Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act)". Communications 66. URL: https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-european-data-governance-data-governance-act.

[4] OPEN DEI. "Aligning Reference Architectures, Open Platforms and Large-Scale Pilots in Digitising European Industry". URL: https://www.opendei.eu/.

[5] PricewaterhouseCoopers (PWC), IDSA (2018). "Data exchange as a first step towards data economy". URL: https://www.pwc.de/en/digitale-transformation/data-exchange-as-a-first-step-towards-data-economy.pdf.

[6] Richter, H., Slowinski, P.R. 2019. "The Data Sharing Economy: On the Emergence of New Intermediaries". IIC 50, 4–29 (2019). https://doi.org/10.1007/s40319-018-00777-7.

[7] The Netherlands AI Coalition (NL AIC) working group Data Sharing (2020). "Verantwoord datadelen voor AI". URL: https://nlaic.com/wp-content/uploads/2020/03/Verantwoord-datadelen-voor-AI.pdf.

[8] The Netherlands AI Coalition (NL AIC), Data Sharing working group (2020). "Responsible data sharing in AI". URL: https://nlaic.com/wp-content/uploads/2020/10/Responsible-data-sharing-in-AI.pdf.

[9] The Netherlands AI Coalition (NL AIC), Data Sharing working group. 'Van First-time-Engineering naar Operationalisatie". URL: https://NL AIC.com/wp-content/uploads/2020/08/NL-AIC-Naar-First-time-Engineering-en-Operationalisatie.pdf.

[10] The Netherlands AI Coalition (NL AIC), Data Sharing working group (2020). "Blueprint NL AIC Data Sharing System Architecture". Version 0.1, January 2021. *Available on request from report editor, see Colophon.*

[11] Liezenberg, C., Lycklama, D., and Nijland, S. (2019). "Everything Transaction". Amsterdam, Lannoo Campus.

[12] Dalmolen, S., Bastiaansen, H., Kollenstart, M. and Punter, M. (2019). "Infrastructural Sovereignty over Agreement and Transaction Data ('Metadata') in an Open Network-model for Multilateral Sharing of Sensitive Data". ICIS2019 Conference, Munich, Germany, 15th – 18th December 2019. URL: https://aisel.aisnet.org/icis2019/economics_is/economics_is/23/.

[13] Bastiaansen, H., Kollenstart, M., Dalmolen, S. and van Engers, T. (2020). "User-Centric Network-Model for Data Control with Interoperable Legal Data Sharing Artefacts - Improved Data Sovereignty, Trust and Security for Enhanced Adoption in Interorganizational and Supply Chain IS Applications". Proceedings of the Twenty-Fourth Pacific Asia Conference on Information Systems, Dubai, UAE, June 2020. URL: https://aisel.aisnet.org/pacis2020/172/.

[14] Zhang, L. Cushing, R., Gommans, L., de Laat, C. and P. Grosso (2019). "Modeling of Collaboration Archetypes in Digital Market Places". IEEE Access, Volume 7, pp. 102689 - 102700, July 2019. doi: 10.1109/ACCESS.2019.2931762. URL: https://ieeexplore.ieee.org/document/8779607.

[15] Deutsches Institut für Normung (2019). "DIN SPEC 27070: 'Reference Architecture for a Security Gateway for Sharing Industry Data and Services". URL: https://www.beuth.de/de/technische-regel/din-spec-27070/319111044.

[16] GAIA-X 2020 "GAIA-X: A Federated Data Infrastructure for Europe". URL: https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html.

[17] Open Services Gateway Initiative (OSGi) Alliance. "OSGi - The Dynamic Module System for Java". URL: https://www.osgi.org/.

[18] Kubernetes. "Production-Grade Container Orchestration". URL: https://kubernetes.io/.

[19] OASIS (2013). "eXtensible Access Control Markup Language (XACML) Version 3.0". OASIS Standard. 2013. URL: http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html.

[20] Otto, B., Steinbuss, S., Teuscher, A., and Lohmann, S., IDSA (2019). "International Data Spaces: Reference Architecture Model Version 3," International Data Spaces Association – IDSA, URL: https://www.internationaldataspaces.org/wp-content/uploads/2019/03/IDS-Reference-Architecture-Model-3.0.pdf.

[21] NIST – National Institute for Standards and Technology 2020. "Zero Trust Architecture". NIST Special Publication 800-207, August 2020. URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf.

[22] Brainport Industries. "Smart Connected Supplier Network (SCSN)". URL: https://smart-connected.nl.

[23] SOLID 2020. "The SOLID project (SOcial LInked Data)". URL: https://solidproject.org/.

[24] AMsterdam data EXchange (AMdEX). "Towards an internationally trusted exchange of data". URL: https://www.towardsamdex.org/https://www.amsterdameconomicboard.com/initiatief/amdex.

[25] OPEN DEI. "Design Principles for Data Spaces – White Paper". *Currently under development. Publication expected in Q1 2021*.

[26] The Netherlands AI Coalition (NL AIC), Data Sharing working group. "NL AI Coalition". URL: https://gitlab.com/nlaic.

[27] European Parliament & Council of the European Union. (2016). "General Data Protection Regulation". URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[28] GAIA-X 2020. "GAIA-X: Technical Architecture," June 2020. URL: https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-technical-architecture.html.

[29] Go-Fair. "FAIR Principles". URL: https://www.go-fair.org/fair-principles/.

[30] Dutch Techcentre for Life Sciences. "Personal Health Train". URL: https://www.dtls.nl/fair-data/personal-health-train/.

[31] Wikipedia. "MNIST Database". URL: https://en.wikipedia.org/wiki/MNIST_database.

[32] GO FAIR. "Virus Outbreak Data Network". URL: https://www.go-fair.org/implementation-networks/overview/vodan/.

[33] FAIR Data Team. "WHO COVID-19 Semantic Data Model". URL: https://github.com/FAIRDataTeam/WHO-COVID-CRF/.

[34] Dutch Neutral Logistics Information Platform (NLIP). "iSHARE Data Sharing Initiative". URL: https://www.iSHAREworks.org/en/.

[35] European Union (2017). "New European Interoperability Framework (EIF) – Promoting seamless services and data flows for European public administrations". URL: https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf.

[36] EU Research Project EUHUBS4DATA. "European Federation of Data Driven Innovation Hubs". URL: https://euhubs4data.eu/.

[37] Dutch NWO Research Project "IDS-Connector Store and Interoperability Simulator for SMEs (CLiCKS)". URL: https://www.nwo.nl/en/research-and-results/research-projects/i/42/35042.html.

[38] World Wide Web Consortium (W3C) (2018). "W3C Recommendation - ODRL Information Model 2.2". URL: https://www.w3.org/TR/odrl-model/.

[39] IDS Association (IDSA). "Usage Control in the International Data Space". URL: https://www.internationaldataspaces.org/wp-content/uploads/2020/09/IDSA-Position-Paper-Usage-Control-in-IDS.pdf.

[40] The Netherlands AI Coalition (NL AIC), Data Sharing working group. 'Action Plan NL AI Coalition', January 2021.

Colophon

This 'GAP analysis' document is a result of the work being done in the NL AIC working group Data Sharing. It builds further upon the previous results of the working group, i.e. the report on identifying the specific challenges for data sharing for AI and overview of technologies and architectures that can be used in addressing these challenges [7] [8], and the description of the development process from first-time engineering towards operationalization [9].

The GAP analysis has been done in the fourth quarter of 2020.

Authors	With contribution of TNO: H.J.M. Bastiaansen, PhD (editor) M. Kollenstart, MSc S. Dalmolen, MSc F.T.A. van Ette, MSc T.R. Nijman, MSc B.T. Musters, MSc E.J.J. Somers, BSc J. Adriaanse, MSc
Contact	The Netherlands AI Coalition info@nlaic.com www.nlaic.com
Date	December 2020