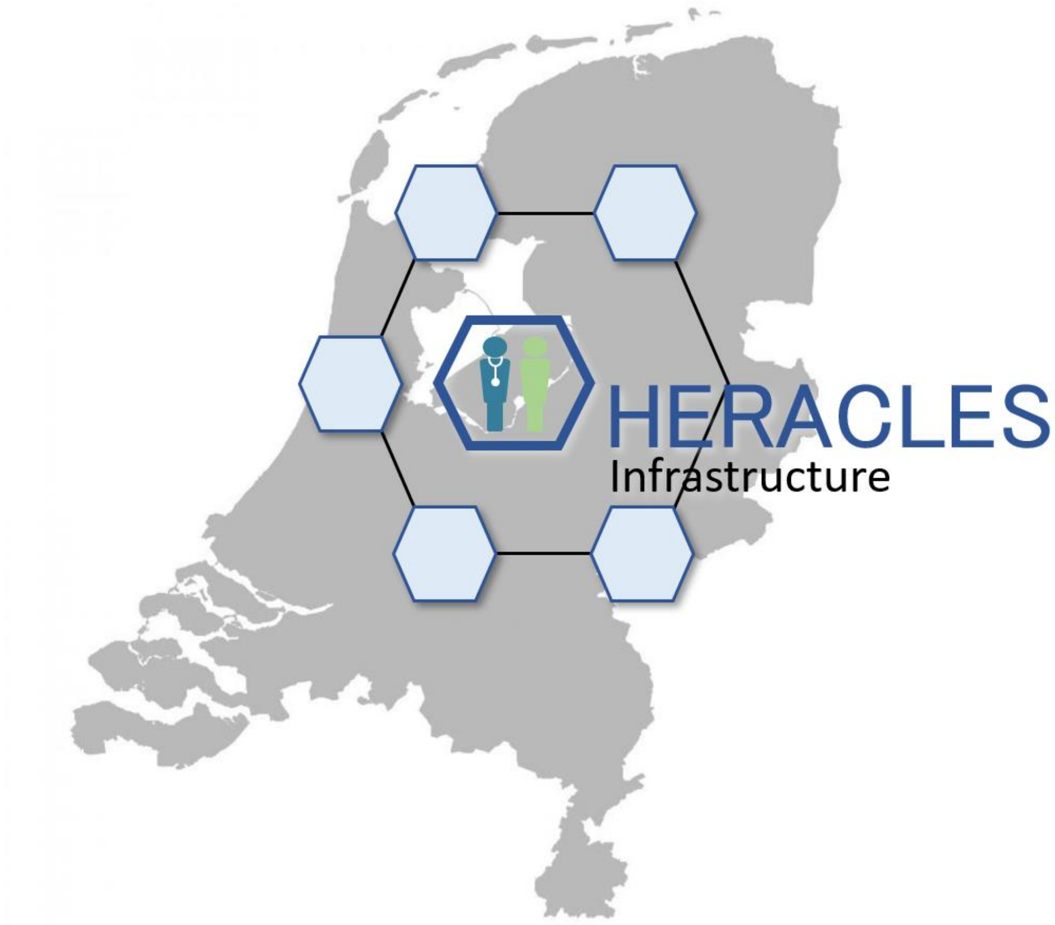
HERACLES

D2.2.1 : Architecture design of Health Dataspace (version 1)



Disclaimer and acknowledgements

Health~Holland
SHARED CHALLENGES, SMART SOLUTIONS

Granted by the Dutch Top sector Life Sciences and Health represented by Stichting LSH-TKI with brand name Health Holland, to promote and stimulate new public-private partnerships to undertake research and development projects in the life sciences

Disclaimer

The content of this document reflects only the author's view. The foundation Health Holland is not responsible for any use that may be made of the information it contains.

While the information contained in the documents is believed to be accurate, the authors or any other participant in the HERACLES consortium make no warranty of any kind with regard to this material including, but not limited to the implied warranties of merchantability and fitness for a particular purpose.

Neither the HERACLES consortium nor any of its members, their officers, employees or agents shall be responsible or liable in negligence or otherwise howsoever in respect of any inaccuracy or omission herein.

Without derogating from the generality of the foregoing neither the HERACLES Consortium nor any of its members, their officers, employees or agents shall be liable for any direct or indirect or consequential loss or damage caused by or arising from any information advice or inaccuracy or omission herein.

Copyright message

Acronym: HERACLES **LSH Match project number:** LSHM21060

Full Title	**HE**alth **R**ese**A**rch - **C**ancer **L**iving labs - setting up an **E**cosystem of trust (**S**ecure and Sovereign)		
Project coordinator	TNO		
Deliverable	D2.2.1 – Architecture design version 1		
Document Type	R		
Lead beneficiary	TNO	Lead beneficiary	TNO
Responsible author	Erik Cornelisse (TNO)		
Additional authors and contributors	Simon Dalmolen (TNO), Maarten Everts (LinkSight), Maarten Kollenstart (TNO)		

Document information

Issue	Date	Comment	Author
V0.1	2023-06-01	Initial document with table of content	EC
V0.2	2023-07-01	Introduction	EC
V0.3	2023-08-01	Describing the five layers + content business layer: roles & appendix A	EC
V0.4	2023-08-18	Chapter 3 - System layer	EC
V0.5	2023-08-29	Chapter 4 – Perspectives	EC
V0.6	2023-11-23	Update of chapter 3.4 – Process layer	EC
V0.7	2023-12-07	Chapter 3 - Update control & data planes	EC
v.08	2024-02-14	Chapter 3 – Update information layer description and protocols specification	EC
V.09	2024-03-12	Full concept for final review	EC
V1.0	2024-03-31	Full version	EC

Approved by:			
Issue	**Date**	**Name**	**Organisation**
V1.0	2024-03-31	Erik Cornelisse	TNO

Contents

List of figures

List of tables

Glossary of terms and acronyms used

Acronym / Term	Description
A2D	Algorithm to the Data
API	Application Programming Interface
Connector	A software service to enable and facilitate the exchange of information with participants of a dataspace in a secure and controlled way (see section 2.2)
Control plane	A system sub-layer that implements generic data space functionality such as identification, authorisation and verification of certificates.
D2A	Data to the Algorithm
Data Permit	A data contract, an agreement with all involved stakeholder which specifies the research objective, required data and involved data providers.
Data plane	A system sub-layer that is a generic implementation to facilitate the use data via HERACLES channels.
Data Sovereignty	The ability of a legal person to exclusively and sovereignly decide concerning the usage of data.
Dataspace	A decentralized infrastructure for trustworthy data sharing and exchange in data ecosystems based on commonly agreed principles.
DID	Decentralized Identifiers
DPM	Data Permit Management application
EHDS	European Health Data Space
EIF	European Interoperability Framework
FAIR	Findable, Accessible, Interoperable, and Reusable.
FAIR Principles	The FAIR principles are a set of guidelines for making data Findable, Accessible, Interoperable, and Reusable. These principles are designed to help organizations manage their data in a way that maximizes its value and minimizes the risk of data loss or misuse
FAIR applications	A FAIR application is a software or digital tool designed and developed with the primary goal of facilitating the Findability, Accessibility, Interoperability, and Reusability of data. In the HERACLES project a FAIR Data Point and FAIR Data Station are specialized FAIR applications required to implement the FAIR principles.
FDP	FAIR Data Point

Federated	Decentralized self-governing entities collaborating with each other to achieve a common objective based on common policies, controls and enforcement abilities governing the use of shared resources and data in particular among participants.
FL	Federated Learning
FDS	FAIR Data Station
GDPR	General Data Protection Regulation
GP	General Practitioner
HERACLES channel	A virtual private network where all resources and involved data are compliant to a specific Data Permit which is under control of the Control plane. Each HERACLES channel is related to one specific Data Permit.
HERACLES infrastructure	a federations of certified software services using the HERACLES Dataspace to exchange data in a secure and controlled way.
IDS	International Data Spaces
IDSA	International Data Spaces Association
IDS-RAM	International Data Spaces – Reference Architecture Model
Interoperability	Interoperability is the ability of organisations to interact towards mutually beneficial goals, involving the sharing of information and knowledge between these organisations, through the business processes they support, by means of the exchange of data between their ICT systems.
Local storage	Data persistency under strict and full control of the responsible service provider.
MPC	Multi-party computations
Parameter	Information element / variable
Participant	A legal entity with a unique verifiable identity, authenticated and accepted within a dataspace, who has accepted and is committed to the governance model of that dataspace.
PEF	Policy Enforcement Framework
PET	Privacy Enhanced Technologies
PHT	Personal Health Train
SSI	Self-sovereign Identities
TSG	TNO Security Gateway
TRL	Technology Readiness Level
TOC	Table of Content
TTP	Trusted Third Party

Table 1 – Glossary of terms and acronyms used

Executive summary

This document describes the architecture of the HERACLES infrastructure in terms of functional components and their interfaces in a *technology agnostic way*. The HERACLES infrastructure as a whole is based on certified services connected via a dataspace to enable *data sovereignty* and using real world sensitive medical data in a secure and privacy preserving way. Use of already existing solutions is one of the design principles and this document will refer to those specification and this document provides only a summary to provide a context. The use of the International Dataspace Association – Reference Architecture Model (IDS-RAM) is used as foundation for controlled and secure data exchange and is extensively documented. The HERACLES Architecture uses this technology as a starting point and focuses on the *HERACLES specific aspects*.

A five-layer structure is used to express various stakeholders' concerns and viewpoints at different levels of granularity at a Business, Functional, Information, Process and System layer. In addition three perspectives that need to be implemented across all five layers: Security, Certification and Governance are described as well.

The HERACLES Infrastructure takes in account the legal, organizational, semantical, and technical interoperability aspects. Permission to use data is specified in a *Data Contract* per research case from all four interoperability aspects. The technical implementation of such a Data contract is being supported with use of European standards for a dataspace and the creation of secure *HERACLES channels* which support the use of Privacy Enhanced technologies such Multi-party Computation and Federated Learning. FAIR Data Points, FAIR Data Stations and a Data Permit application are the key functional components for respectively data discovery, data usage and data contract negotiation.

Based on concrete specification of protocols which includes process and interface descriptions, multiple implementations can be realized based on already existing components, to test the architecture and to demonstrate the interoperability between those implementation and the feasibility of the HERACLES infrastructure.

The main recommendation is to realize multiple minimal viable products (MVPs) based on the same architecture (as described in this document) but as different implementations to test and demonstrate the interoperability aspects. This avoids a single solution / vendor lock-in and maximize the use of already working components.

1 Introduction

1.1 General

This document describes the architecture of the HERACLES infrastructure in terms of functional components and their interfaces in a technology agnostic way. This is not only common practise but also avoid a vendor lock-in situation. Furthermore, within the HERACLES project multiple different implementations will be realised to demonstrate the interoperability capabilities of the proposed architecture.

The purpose of this document is to define the boundaries primarily for the technical and semantical aspects of the HERACLES infrastructure and some organisational aspects regarding data governance.

The intended audience for this document is the stakeholders of the HERACLES project, which implies that knowledge of the Healthcare domain is required. Specific knowledge about IT architectures is required to provide contributions for this document and/or developing implementations based on this document.

This document is the result of WP2 - Health Dataspace infrastructure of the HERACLES project.

1.2 Design objective and starting points

One of the main objectives of the HERACLES project is: "*to design and develop a generic health data space infrastructure, coupling International Data Spaces (IDS) standards with privacy-enhancing technologies (PETs) supporting the Personal Health Train set of agreements*" [1].

In particular, FAIR-principles for the description of available data are supported by IDS, in order for 'Machine2Machine' communication to be possible. Datasets and services should be Findable, Accessible, Interoperable and Reusable. Next to the FAIR principles the reference architecture of IDS is applied to specify per organization, user, algorithm which data can be used and for which reason. In addition, a selection of generic secure analysis components will be developed, to enable specific privacy-enhancing algorithms and analyses using MPC/FL, and ensuring their integration with the data sharing infrastructure, in collaboration with TNOs open source MPC Lab. The dataspace shall also be able to support data to the algorithm (D2A) so that the value chain can benefit from data integration e.g. clinical, radiological, lab, genetic pathological data.

WP2 adopts the definition of a data space (IDS) according to the OPEN DEI Position paper [ref: Nagel et al, 2021] and the need for a data sharing infrastructure [ref: Bastiaansen et al, 2020]. Multiple types of data sharing methodologies to train AI models will be made possible in the data space infrastructure. The approach A2D (Algorithm to Data) in which the data stays at the source, and the algorithm 'travels' to the data will be made in two distinctions:

1. Federated learning (FL). Machine learning models are trained locally, and only aggregated model parameters are shared. This approach is primarily suitable for a 'horizontally partitioned' data setting;
2. Multi-Party Computation (MPC) uses cryptographic techniques to jointly train AI-algorithms on encrypted data without sharing any underlying sensitive data, suitable for 'vertically partitioned' data. MPC is more secure than FL, but requires more computation time.

In all cases it is crucial that the algorithms are trusted by everyone in the ecosystem and the infrastructure will facilitate the strengthening of trust between the various stakeholders by making it explicit what data is used when, for what, by whom.

1.3 System context

The HERACLES architecture describes the *HERACLES infrastructure* which is a federation of *certified software services* using the *HERACLES Dataspace* to use data in a secure and controlled way. See section 3.2 for a more elaborated description about the constellation of the HERACLES infrastructure and a decomposition into functional building blocks. The figure below describes the context of the HERACLES infrastructure and the types of *accountable actors*: responsible for their actions and answerable for the outcomes. Accountability is relevant in the HERACLES Infrastructure in regards to legislation, terms and conditions agreed within the HERACLES consortium and the organisational aspects. Section 3.1 elaborates on the types of actors (roles) and their interactions.

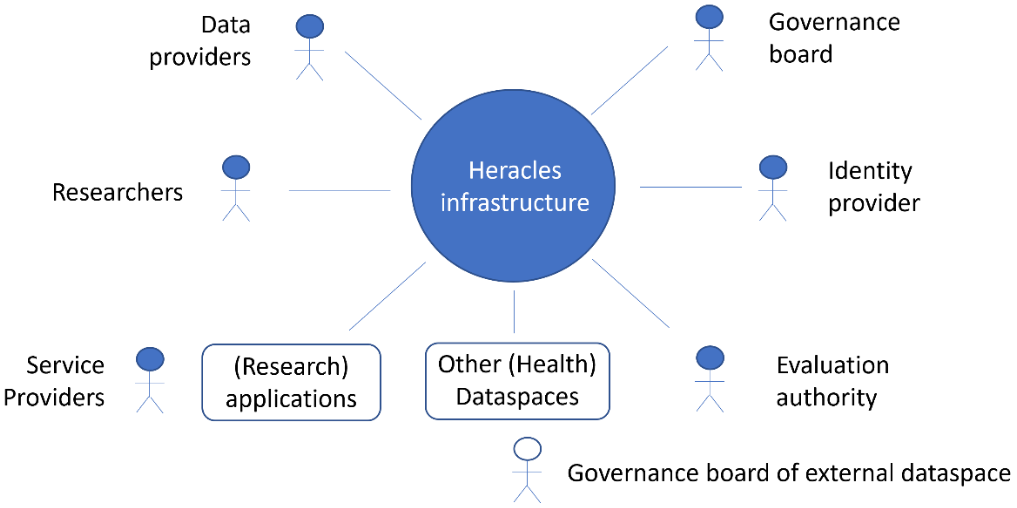


Figure 1 HERACLES infrastructure context diagram

The role of Data consumer is limited to *Researcher* to address explicitly the purpose of the Heracles infrastructure and that access to data is required for research purposes only. Also, the role of *Service provider* as accountable actor is mentioned in relation to providing software components (applications and software services). Service providers can also participate on behalf of other actors as a delegate responsibility.

A core functionality is the *HERACLES Dataspace* which facilitates secure access to data between trusted participants. This requires the identity and authentication of users such as researchers by a trusted *Identity provider*. For the HERACLES infrastructure verifiable credentials will be used based on YIVI. See section 3.5 for more details. To provide objectively and verifiable transparency of the workings of the HERACLES infrastructure, an *Evaluation authority* (auditor) is required. Not only to verify the correct inner working of the system as a whole based on a certification process, but also to monitor that the HERACLES infrastructure works according to the commonly agreed terms and conditions of the HERACLES consortium. The HERACLES *governance board* has the responsibility to manage and monitor the operations of the HERACLES infrastructure which will be checked periodically with independent audits. See section 4 for more details. Connection with other *(Health) dataspaces* are included in the design to become compatible with other European initiatives.

HERACLES infrastructure nodes

The HERACLES Infrastructure consists of an organisational and a supporting technical part. The organisational part will be described in a sperate document as the HERACLES Governance model. The technical part describes how the HERACLES infrastructure is built of HERACLES nodes which contain the following main functional components:

- A FAIR Data Point (FDP) enables discovery of available data at FAIR Data Stations to implement the "F" of findable in FAIR and to provide information about the data including the terms and conditions for using the data;
- A FAIR Data Station (FDS) enables controlled access to data to implement the "A" of accessible in FAIR where a copy of the data is being stored which is described in a FAIR Data Point. How the data is being processed can vary per research analysis and should be agreed and specified in advance as part of a Data Permit;
- Data Permit Management (DPM) supports a community driven approach to reach agreement about data usage. When all involved stakeholders agree on the research goal, approach and required means of resources and data in particular, the result will be a Data Permit. A Data Permit is an agreement for a specific research goal, and describes the terms and conditions for using data at the FAIR Data Stations. A digital subset of the Data Permit will be used by the HERACLES infrastructure to enforce the agreement and to control access accordingly.

Research applications are required for starting and handling the aggregated (anonymized) results collected at FAIR Data Stations. Research applications must implement the HERACLES API's and must be known as part of the agreements to grant data usage but are not part of the HERACLES infrastructure. Although the HERACLES API's are based on European standards to connect with other Health dataspaces, connecting to them is not within the scope of the HERALCES project.

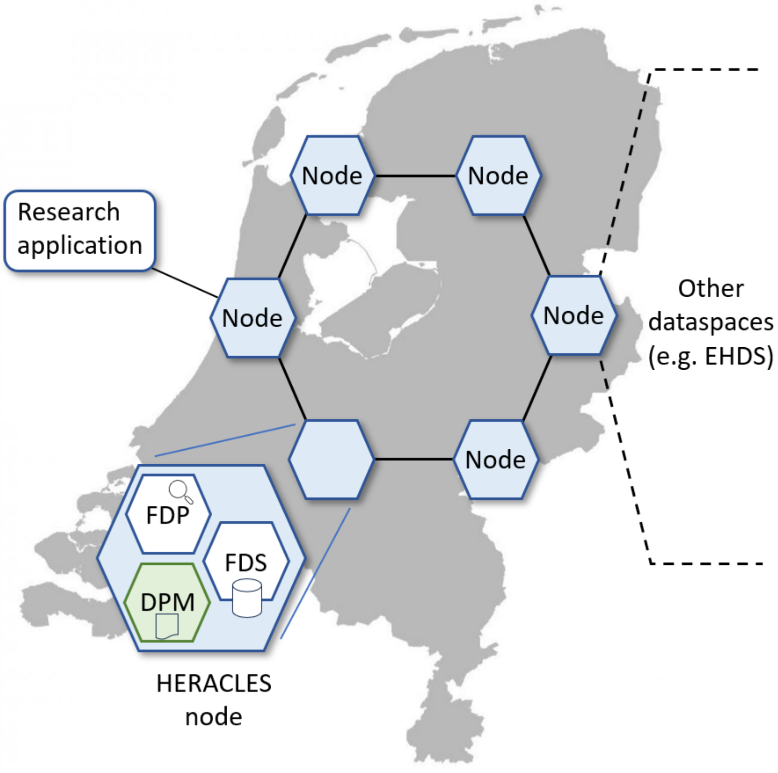


Figure 2 - Functional architecture of the HERACLES infrastructure

This document describes these HERACLES nodes and how they interact with each other.

1.4 Approach

To provide a starting point for the design of a secure and privacy preserving open infrastructure for federated analysis of Real World Data (RWD), a functional architecture and data architecture will be designed following International Data Space (IDS) standards, to establish semantic interoperability through a managed Health Data Space for the digital support needed for health AI specific applications [1]. Infrastructure components will be developed with IDS-RAM[4] compliant data sharing components to realize connections, authentication, authorization and data usage.

A data space[1] will be developed, enabling the integration of different data sources and services in a decoupled and modular way, ensuring data sovereignty[2]. Data sharing in a federated data sharing environment can be characterized as a system-of-systems, in which a multitude of dedicated systems pool their resources and capabilities together to create a new, overarching, system with value adding functionality and performance. The International Data Space (IDS) reference model (IDS-RAM) [ref: Nagel et al, 2020] is a starting point, since it is open, agnostic with respect to technologies and domains, based on zero-trust architectures and therefore an important basis for a health data space that should be open to different implementations and modular components (*open to all, no vendor-lock-in*), and providing method and tools for its implementation and promoting the adoption among consortium members. This includes the deployment of IDS-based connectors and essential components, and the participation of a certification Authority and a Dynamic Attribute Provisioning Service to ensure secure exchanges among certified participants. The IDSA reference architecture is at the basis of this Task and particularly its DIN SPEC 27070 pre-standard.

For the use cases defined in work-package 1(WP1) the data will be sensitive and vertically partitioned, which is particularly challenging. Therefore, the main focus is to enable A2D (MPC/FL) in the vertical setting in the Health Data Space. Different technology partners in this project bring in technology that can be used as basis for such a Health Data Space, including IDS-implementations, the Vantage6 Personal Health Train (PHT) infrastructure, the federated data infrastructure Feder8 and the technology of Linksight.

With use of these existing technologies the aim is to define a common denominator and to develop a scalable and vendor-independent solution based on these stacks such that (a) the PHT principles are served, (b) it is connected to an international community of developers (e.g. IDS a vendor neutral design based on a zero-trust architecture) and satisfies IDS-standards, (c) and it meets standards of the European Health Data Space. Working with different technologies and implementations, is a direct validation of the interoperability of the HERACLES infrastructure.

One of the main challenges is to enable and facilitate the application of secure federated AI solutions like MPC and FL on vertically partitioned data: a) Design and implementation of interoperable IDS components ("tracks") and b) Design and implementation of well-defined interfaces between IDS components and analysis components (MPC/FL).

Based on the first version of the architectural design c.q. this document, a minimal viable infrastructure will be implemented as a proof of concept, to validate the design decisions and to refine the specifications and the design in general.

[1] A dataspace is a decentralized infrastructure for trustworthy data sharing and exchange in data ecosystems based on commonly agreed principles. (source OpenDEI). This definition implies that the scope of a dataspace is based on common ground of the participants of that dataspace.

[2] Data sovereignty refers to the concept that individuals and organizations stay in control about the use of their data.

Compliant with the common system architecture models and standards (e.g., ISO 42010, 4+1 view model), the HERACLES architecture uses a five-layer structure expressing various stakeholders' concerns and viewpoints at different levels of granularity.

The general structure of the HERACLES Architecture is illustrated in the figure below. The model is made up of five layers:

- The *Business Layer* specifies and categorizes the different roles which the participants of the HERACLES infrastructure can assume, and it specifies the main activities and interactions connected with each of these roles.
- The *Functional Layer* defines the functional requirements of the HERACLES Infrastructure, plus the concrete features to be derived from these.
- The *Process Layer* specifies the interactions taking place between the different components of the HERACLES infrastructure; using the BPMN notation, it provides a dynamic view of the architecture.
- The *Information Layer* defines a conceptual model which makes use of linked-data principles for describing both the static and the dynamic aspects of the HERACLES infrastructure constituents.
- The *System Layer* is concerned with the decomposition of the logical software components, considering aspects such as integration, configuration, deployment, and extensibility of these components.

In addition, the HERACLES Architecture comprises three perspectives that need to be implemented across all five layers:

- Security,
- Certification, and
- Governance

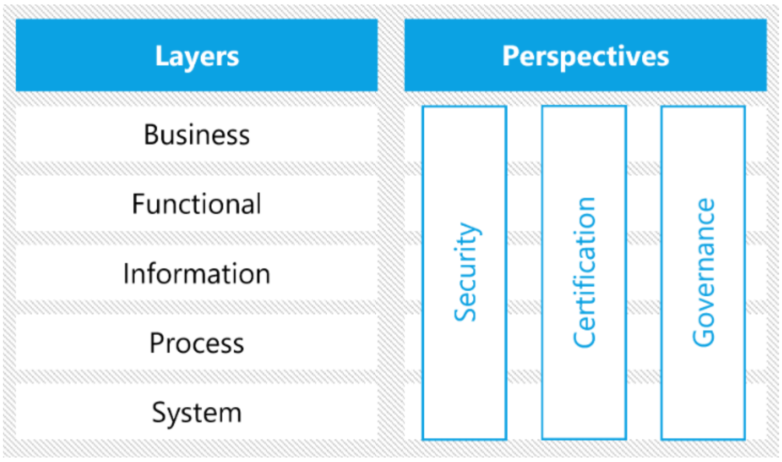


Figure 3 General structure of the HERACLES Architecture

The infrastructure will be tested, first on synthetic data, then on individual medical data in a pilot. The implementation of the HERACLES infrastructure is a success when:

- The consortium partners indicate that this infrastructure is modular with well-defined interfaces, ensuring no vendor lock-in [O1];
- At least three different analyses are performed on distributed real data sets from different organizations using this infrastructure, with approval from the organizations data protection officers [O2].

1.5 Document structure

To provide a quick overview of the content and structure of this document:

Chapter 1 - Introduction, describing the scope and objectives of this document;

Chapter 2 - Context of the HERACLES infrastructure described in terms of a data driven eco-system, where data sovereignty and data quality control are key capabilities. This second also addresses the frameworks and standards to be used as starting point;

Chapter 3 - The five layers of the HERACLES architecture are described here. Starting with the business layer where types of actors are identified. The second layer of the architecture describes the identified functional components, followed by the information layer where semantics have an important role. The fourth layer describes the processes and finally the system layer is about the technical components;

Chapter 4 - Three perspectives of the HERACLES architecture: security, privacy and data governance;

Chapter 5 - Recommendations and conclusion for the follow up tasks and activities;

Chapter 6 - References to sources;

Appendices - Contain detailed information.

The following figure visualizes the position of this document in relation to other deliverables.

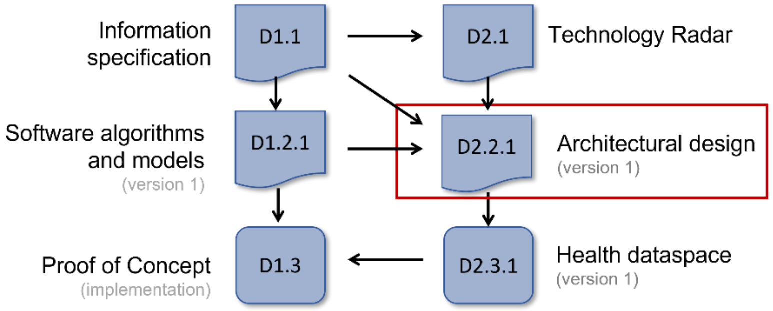


Figure 4 Document relations of D2.2.1 with other documents

The use-cases and information as described in the *Information specification*[D1.1] are the main starting point for this document *Architectural design (version 1)*[D2.2.1]. Applicable technical options and solutions as described in the *Technology Radar* [D2.1]. There is also a technical dependency with document *Software algorithms and models (version 1)* [D1.2.1] because the scope of the algorithms have requirements for the HERACLES infrastructure and therefore also the architecture.

Figure 3 visualizes also the importance of this document for the technical development of the HERACLES infrastructure also indicated as a *Health dataspace*[D2.3.1]. Implicitly it is therefore also important for the implementation of a *Proof of Concept* [D1.3] where the software algorithms and models[D1.2.1] are put to the test.

2 Context of the HERACLES infrastructure

2.1 Data-driven Business ecosystem

A data-driven business ecosystem is a collaborative network of organizations and stakeholders that leverage data and analytics to make informed decisions, drive innovation, and create value. In this ecosystem, data flows seamlessly among participants to optimize processes, enhance functional experiences, and achieve strategic goals.

In the proposal [1], the following two high level objectives are defined to describe when the implementation of the HERACLES infrastructure is a success:

O1 The consortium partners indicate that this infrastructure is modular with well-defined interfaces, ensuring no vendor lock-in;

O2 At least three different analyses are performed on distributed real data sets from different organizations using this infrastructure, with approval from the organizations data protection officers.

These objectives are used to derive the use-case in section 3.1 for each of the six categories of actors as identified and visualised in the context diagram in figure 1.

2.2 Data sovereignty as a key capability

Data sovereignty is a key capability for two main reasons:

1. Regulatory Compliance: Data sovereignty ensures that organizations can comply with local and international data protection laws and regulations. By maintaining control over where data is stored and processed, businesses can adhere to legal requirements, avoid fines, and protect their reputation.
2. Data Security and Privacy: Data sovereignty enhances data security and privacy by reducing the risk of unauthorized access or data breaches. When data is stored and managed within a specific geographic region, it becomes easier to implement robust security measures and safeguards to protect sensitive information from external threats.

Data sovereignty is about finding a balance between the need for protecting one's data and the need for sharing one's data with others. It can be considered a key capability of the HERACLES infrastructure for data providers to stay in control about which data is made accessible and with whom. To find that balance, it is important to take a close look at the data itself, as not all data requires the same level of protection, and as the value contribution of data varies, depending on what class or category it can be subsumed under.

Definition of Data Sovereignty:

> *Data Sovereignty is the ability of a legal person to exclusively and sovereignly decide concerning the usage of data.*

Access to data can be controlled with proper identification, authentication and authorisation (IAA) management based on mutually agreed contracts between the data provider and the data consumer (researchers) and technical enforcement. Additional agreements can be made about the usage of the data by the researchers based on an agreement as well but enforced by audits only.

2.3 Secure use of sensitive data

One of the main objectives of the HERACLES project is: "*to design and develop a generic health data space infrastructure, coupling International Data Spaces (IDS) standards with privacy-enhancing technologies (PETs) supporting the Personal Health Train set of agreements*" [1]

Data Spaces

A common starting point for the design of dataspaces is based on three European initiatives: the new European Interoperability Framework[2], OpenDEI[3] and IDSA[4]. A dataspace is an architectural concept for sharing data and defined as:

> *A dataspace is a decentralized infrastructure for trustworthy data sharing and exchange in data ecosystems based on commonly agreed principles*.

Source: OpenDEI [3]

The initial objective of a dataspace is to remain in control of the data and the infrastructure required for the exchange of information, also known as *data - and digital sovereignty*. This explains the "decentralized" part in the definition above.

Furthermore, it is a collaboration or cooperation of organisations to reduce the costs for implementing a secure mechanism for trustworthy data sharing by (re)using IT services under commonly agreed rules and principles. Developing and deploying IT services once and share them with many, not only reduces the costs but also allows new participants to join faster and without having specialised knowledge of deploying such services of their own. A dataspace is in the first place an organisation of companies who join forces to achieve a mutual objective and in the second place it consists of enabling IT services for data-exchange and is not intended for data storage in contrast to a data lake or database.

Privacy Enhance Technology (PET)

Privacy-enhancing technologies (PETs) are digital solutions that allow information to be collected, processed, and shared while minimizing the risk of privacy violations. PETs are a family of technologies aimed to allow the collaboration between different parties of the sharing of information while protecting the privacy and securing personal or confidential data. Based on the PET guide published by the United Nations in 2023, the following categories will be investigated and implemented as part of the HERACLES infrastructure:

- *Secure Multi-Party Computation*: an umbrella term consisting of different cryptographic protocols allowing several parties to jointly compute a function while preserving the privacy of the input data;
- *Homomorphic Encryption*: a cryptographic technology allows for computations on encrypted data;
- *Synthetic Data*: a family of statistical of ML-based techniques aimed to generate artificial data that preserve the relevant statistical properties of the original data, without exposing any private information;
- *Distributed/Federated Learning*: a family of protocols aiming to jointly train a ML model on data distributed among different parties, without the need of sharing or collecting the data;

These technologies will be used on top of Dataspace technology.

3 Layers of the HERACLES reference architecture

3.1 Business layer

The Business Layer categorizes roles and interaction patterns in the HERACLES infrastructure, fostering innovative business models and data-driven services. It acts as a blueprint for the technical implementation of the HERACLES infrastructure, ensuring alignment and specifying requirements for the Functional Layer.

Type of actors

The terms role and actor are synonyms and the term actor will be used in this document. There are six categories of actors identified within the HERACLES infrastructure:

- Category 1: Data providers (core participant)
- Category 2: Researchers (core participant)
- Category 3: Governance board (or Governance body)
- Category 4: Identity provider (intermediary actor)
- Category 5: Auditors (intermediary actor)
- Category 6: Service Providers (intermediary actor)
- Category 7: Other Health Dataspaces

See also figure 1 – HERACLES infrastructure context diagram. The equivalent type of actor[3] as identified for a IDS compliant dataspace are inserted between the brackets.

Appendix A provides background information about the identified HERACLES dataspace actors for a generic dataspace including their mutual relations. From a legal point of view, it is important to make the distinction between accountable and responsible actors and appendix A elaborates on the concept of delegated actions. This section focusses on the HERACLES specific aspects of the six categories of actors.

REQ *All actors* have a verifiable identity which has been authenticated and are known in the authorisation management process which will be described in the section about digital identities further on in this chapter.

REQ *All actors* also have to comply to the governance model of the HERACLES infrastructure which also state the accountability and responsibilities of the actors in detail.

REQ *All actors* can only interact with the HERACLES infrastructure if the have a verified identity and have the proper autorisation. *By design, nothing is allowed unless* an actor meets <u>all</u> the terms and conditions as stated in the governance model <u>and</u> has a proper data permit.

In this chapter only the relevant aspects of an actor for the HERACLES Architecture are mentioned in case there are related functional requirements involved.

<u>Data providers</u>

This category enables <u>access</u> to data to be used for research purposes via a so called FAIR Data Station within the terms and conditions agreed within the HERACLES consortium and the HERACLES Governance board in particular. The data provider ensures that the quality of the provided data meets the conditions as specified within the governance model. Pre-processing of the data will be necessary in most cases to meet the agreed quality and data exchange format.

[3] For completeness reasons, the technical architecture IDS-RAM[3] also identifies Software Developer also as category of actors which is omitted in the architecture of the HERACLES infrastructure.

The data exchange format will be specified as part of the HERACLES Architecture based on the information described in the HERACLES Information Specification [D1.1].

It is the responsibility of the *data provider* to select a proper subset of its source data in terms of parameters and records and to perform the require pre-processing. The pre-processing can only be done by the *data provider* to preserve the privacy of the records before they are made accessible for a *FAIR Data Station*. Although required, pre-processing and required processes, tools and software services are beyond the scope of the HERACLES infrastructure and remain within the scope of the *data provider* because of the previously mentioned reason.

In the proposed architecture, the data source will never be made accessible via the HERACLES infrastructure, but only a copy of the data will be imported to a *FAIR Data Station* where the data will be stored under strict, full and direct control of the *data provider*. This will be addressed for readability of this document as "*local storage"* as it would be within the walls of the company.

REQ *Data providers* are accountable for the selection of data to provide and to specify the corresponding meta data.

Researchers

This type of actor is a *data consumer* and shall only use the data for research purposes.

REQ *Researchers* shall not have access to individual patient records by design within the HERACLES infrastructure. They can have access to metadata via FAIR Data Point and anonymous statistics and aggregation of data via FAIR Data Stations.

Note There are different research areas which will not result in specific requirements for the design of the HERACLES infrastructure but might require specific or different governance aspects:

- Lung cancer (medical) research
- Ovarium cancer (medical) research
- (Healthcare process research ?)

REQ *Researchers* can have access to anonymous statistics and meta data made available via FAIR Data Points for discovery purposes and to test up front of the amount of data to be analysed meets the terms and conditions as specified in the governance model.

REQ *Researchers* can have access to anonymous statistics and aggregation of data collected via FAIR Data Stations.

REQ *Researchers* can have access via query functionality or indirectly with use of Federated Learning or MPC technology. A description of these three types of data usage can be found in section 3.5 which is about the System layer.

Governance board

This responsibilities and activities of this type of actor will be described in detail with document Governance Assessment [D3.3]. In the context of the HERACLES architecture it is import to mention the following aspects:

REQ The *Governance board* is accountable for preserving the privacy and security of the data that is being used within the HERACLES infrastructure.

It is the responsibility of all participants of the HERACLES infrastructure to comply with the terms and conditions as agreed in the governance model.

REQ It is a crucial requirement that the governance model of the HERACLES infrastructure is not only agreed within the HERACLES consortium but also be verified against compliancy with the National and European regulations.

REQ *The Governance board* represents the participants of the HERACLES dataspace c.q. the HERACLES consortium. Bound by the commonly agreed principles as stated in a Governance model, the Governance board is accountable for the enforcement of the Governance model.

Identity provider

This type of actor creates and authenticate the identity information of all participants of the HERACLES infrastructure. An *Identity provider* is a trusted intermediary service provider who will not use the HERACLES infrastructure. A proper choice for a "trusted" identity provider shall be stated in the governance model.

The identity management process as part of the dataspace will request an *identity provider* to provide verifiable credentials for all participants of the HERALCES infrastructure.

See for more information the section about digital identities.

Auditors (Evaluating authority)

The role is added to address the need to verify objectively if the security and privacy is being preserved. Auditors are as an Evaluation authority described in appendix A, to verify in advance and during operation if the HERACLES infrastructure, its implementation, operation and organisation comply to the terms and conditions as agreed in the governance model.

REQ Auditors must have access to the governance model, the logging of the systems and the HERACLES infrastructure itself to verify if its operation is compliant to the governance model.

Service Providers of components

Components are software services, applications or hardware which can be interact with the HERACLES Infrastructure if they are certified and expected by the Governance board and the HERALCES consortium in general. The

REQ Service providers are accountable for the functionality of the provided components.

Other Health Dataspaces

Interoperability with other dataspaces is taken in account by design, by using existing and upcoming European standards. The development of the European Health Dataspace (EHDS) is taken in account in particular while designing the HERACLES Dataspace but requires further research.

A connection with other Health dataspaces such as the EHDS is not in scope of the HERACLES project but considered as a must have feature for a sustainable solution. This topic will be extended in version 2 of the HERACLES architecture and be part of the work of WP4 as forward looking for an ecosystem.

Other Health Dataspace are expected to enable the use of FAIR Data Points and FAIR Data Stations outside the HERACLES infrastructure. In the first place it is a organisational aspect to decide which FAIR Data Points and Stations are allowed. The governance model for the HERACLES infrastructure should define specifically the terms and conditions for external connections.

3.2 Functional layer

The Functional Layer defines -- irrespective of existing technologies and applications -- the functional requirements of the HERACLES infrastructure, and the features to be implemented resulting thereof. An important requirement is the implementation of the FAIR principles with use of FAIR applications in particular. In the HERACLES Infrastructure FAIR applications are connected with each other via the HERACLES Dataspace based on open standards.

The HERACLES Infrastructure consists of five functional components:

- A Data Contract or Data Permit which describes the agreed terms and conditions for access to the data. The HERACLES architecture foresees in a supporting application to create and negotiate about a Data contract for each research request with all involved stakeholders;
- A FAIR Data Point is for discovery of available data at FAIR Data Stations to implement the "F" of findable in FAIR and to provide information about the data including the terms and conditions for using the data;
- A FAIR Data Station for controlled access to data to implement the "A" of accessible in FAIR where a copy of the data is being stored which is described in a FAIR Data Point;
- A Dataspace to facilitate a controlled, secure, trusted and privacy preserving exchange of data between all other functional components to implement the "I" of interoperability in FAIR. This includes also an interface with tools and applications to perform analysis and even future connections with other dataspaces. The HERACLES dataspace facilitates also the discovery of available software services such as FAIR Data Points & Stations and applications which are using the HERACLES Infrastructure;
- A HERACLES Channel is linked to a data contract and facilitates the access to data under strict control of the Heracles Dataspace.

Note The "R" of repeatable in FAIR is implemented by the organisational aspects in a Governance model of the HERACLES Infrastructure. See section 4.3 for more details.

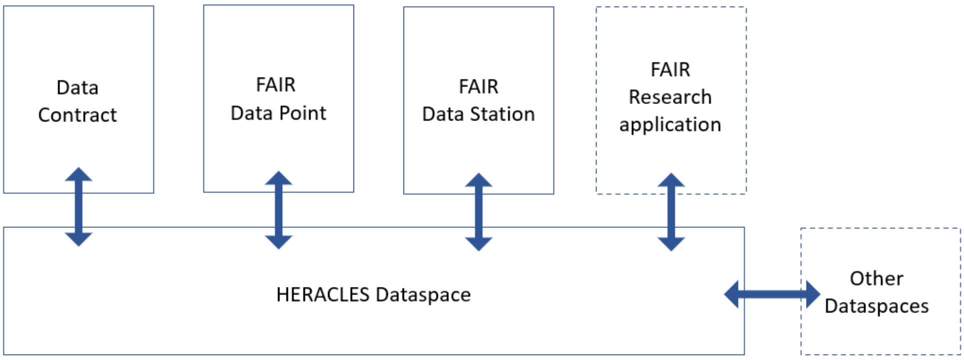


Figure 5 Functional architecture of the HERACLES infrastructure

Although not part of the HERACLES Infrastructure, there are two other functional components to take in account: FAIR Research applications[4] and other dataspaces because of their interactions with the HERACLES Infrastructure and the required functionality to enable them to participate.

REQ All valid/certified components shall ONLY allow data usage in case of a valid *Data Contract*.

REQ All valid/certified components shall enable data usage ONLY under strict control of the HERACLES Dataspace via so called *HERACLES channels* (the blue arrows in the figure above.

Note Components can only be valid/certified if they have properly implemented a *HERACLES channel*.

[4] A FAIR research application is a software application which uses FAIR Data Points and Stations.

HERACLES Dataspace

This component provides generic functionality which is not specific for the Healthcare domain or HERACLES in particular but essential for a controlled, secure, trusted and privacy preserving exchange of data between all other functional components.

REQ The Governance Board is accountable for the HERACLES Dataspace and can delegate the operational responsibility to a service provider.

The functional architecture of the HERACLES Dataspace divides the requirements into six groups of software functionality compliant to the IDS-RAM standard.

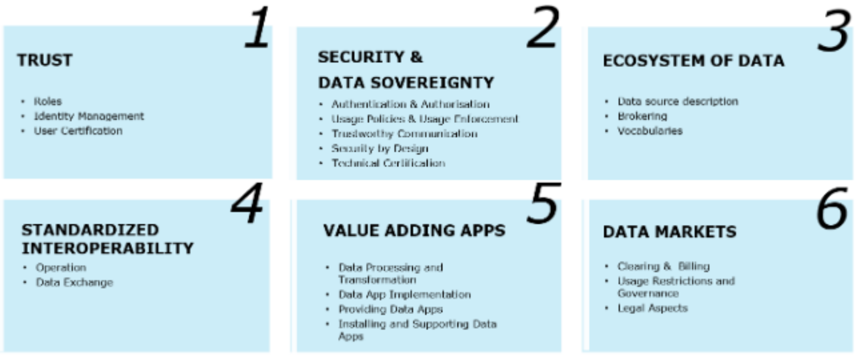

1
TRUST
• Roles
• Identity Management
• User Certification
2
SECURITY &
DATA SOVEREIGNTY
• Authentication & Authorisation
• Usage Policies & Usage Enforcement
• Trustworthy Communication
• Security by Design
• Technical Certification
3
ECOSYSTEM OF DATA
• Data source description
• Brokering
• Vocabularies
4
STANDARDIZED INTEROPERABILITY
• Operation
• Data Exchange
5
VALUE ADDING APPS
• Data Processing and Transformation
• Data App Implementation
• Providing Data Apps
• Installing and Supporting Data Apps
6
DATA MARKETS
• Clearing & Billing
• Usage Restrictions and Governance
• Legal Aspects

Figure 6 Functional architecture of the HERACLES Dataspace

REQ The HERACLES Dataspace shall be compliant with the IDS-RAM. A detailed description of all six building blocks can be found in the IDS-RAM[4].

Functional Dataspace building blocks five and six have a low priority for the implementation of the HERACLES Infrastructure but are included for completeness reasons and can be added in the future. Trust and secure data exchange are identified as the most important dataspace building block for the HERACLES Infrastructure and are therefore elaborated in more detail below.

Functional dataspace building block - Trust

Although requirements related to trust are usually non-functional, they are addressed by the Functional Layer, since they represent fundamental features of the HERACLES infrastructure. The Trust group comprises three main aspects (*roles, identity management, and user certification*), which are complemented by governance aspects (see Section on Data Governance).

- Roles: each role in the HERACLES infrastructure has certain rights and duties. More information about the roles is given in the *Business Layer*.
- Identity Management: every component (application or software service) participating in the HERACLES infrastructure must have a unique identifier and a valid certificate. In addition, each component must be able to verify the identity of other components (with special conditions being applied here; e.g., security profiles).
- User Certification: each participant in the HERACLES infrastructure must undergo certification in order to establish trust among all participants. More information about the certification process is given in section 4.3: the Certification Perspective.

Functional dataspace building blocks 2 & 4 – Security & Data Exchange

To implement controlled and secure use of data, the concept of HERACLES channels has been defined as a HERACLES specific feature on-top of the IDS – Reference Architecture. Before data can be accessed and used, the HERACLES Dataspace verifies if all conditions are met as described in a Data Permit: an electronic implementation of a data contract.

REQ The HERACLES Dataspace data usage of certified FAIR applications and components in general via HERACLES channels. This means that the HERACLES Dataspace verifies according to a *Data Contract* if the:

- involved user(s) and service providers are properly identified;
- involved components are certified;
- validity date is applicable (not overdue).

This implies that the configuration of the HERACLES infrastructure to enable the data usage for a specific research question is part of the *Data Contract*.

HERACLES Channel

A HERACLES Channel is a collection of secure digital links between two or more applications or software services, creating point-to-point tunnels that encrypt the data. Conceptually it is a group of network connections related to a specific *Data Contract* (data permit) under strict control of the Heracles Dataspace.

REQ A HERACLES Channel shall only enable data usage when all conditions in the corresponding Data Permit are met. For example, within a HERACLES Channel only software components stated in the Data Permit are allowed. In advance, the HERACLES Dataspaces verifies if the involved actors as stated in a Data Permit have valid credentials and certificates.

REQ Involved Data Providers can suspend a HERACLES Channel in the event of suspected misuse or unforeseen situations that violate security or endanger privacy.

The following two categories of data usage are identified:

- *Data to the algorithm* (D2A), query requests for data descriptions and
- *Algorithm to the Data* (A2A) with two variations:
    - Federated learning and
    - Multi-Part Computation (MPC).

Data contract

A Permit application is foreseen to create a *Data Contract* and to support the request and grant services between all involved participants and the researcher and the data providers in particular.

REQ A permit request service enables to specify at least the following aspects:

- The identity of the researcher;
- The research purpose and approach;
- The required data and related data sources discovered via FAIR Data Points;
- The required software and configuration for processing the data;

REQ A permit request service derives from the request the following aspects:

- The involved participants and in particular those who decided whether a request can be granted or not;
- The terms and conditions for data usage in general and per data source;

REQ A permit grant service coordintates the approval management among all involved participants for a specific permit request. This includes the ability to adjust/refine the terms and conditions for a specific request / use-case. If a request is granted, the result is a *Data Contract*.

REQ A permit grant service stores all approved Data Contracts to be accessible by all involved stakeholders. A subset of a Data Contract will be used to create digital contracts for automated enforcement by the *HERACLES channels*.

HERACLES FAIR Data Point

A FAIR Data Point provides access to data descriptions following the FAIR principles. It allows digital object owners/publishers to expose the description of their digital objects in a FAIR manner and enables consumers/users to discover information about the digital objects they are interested in. A Fair Data Point stores information about data sets including the terms and conditions for data usage. It aims to give anyone the power of putting their own data on the web, addressing issues related to the metadata needed for findability and reusability, and a uniform open way of accessing the data. FAIR Data Points can be used to describe data sets in a FAIR way, using standard metadata, and make them available through simple WWW protocols. The FAIR Data Point is a metadata repository that follows the DCAT schema and utilizes the Linked Data Platform to manage the hierarchical structure of metadata.

REQ The Data provider of the HERACLES Infrastructure is accountable for a HERACLES FAIR Data Point. The operational responsibility can be delegated to a service provider.

For the HERACLES Infrastructure a specialization of the generic description for a FAIR Data Point is defined to meet the main objective [O1].

REQ A HERACLES FAIR Data Point shall provide data descriptions available at a corresponding FAIR Data Station. The data provider is therefore also accountable for the decision which data is being made available to the HERACLES Infrastructure and which part of that data is accessible for a specific research task. (See section 3.4 about the Data Permit phase and the process of defining the scope of a data permit as part of defining the scope of data accessibility)

REQ A HERACLES FAIR Data Point shall not disclose information that can be linked directly or indirectly to a specific patient to preserve privacy. This means that in case of aggregated data in terms of amounts per selected group of parameters can ONLY be shared if these amounts are above a threshold to be defined in the Governance model.

REQ A HERACLES FAIR Data Point shall indicate as part of the data descriptions if the data is either: test/mock-up data, synthetic data or real-live data.

REQ A HERACLES FAIR Data Point shall provide the terms and conditions for data usage as specified by the data provider. There terms and conditions can be made more specific for a specific research request during the contract negotiation phase of the Data Permit process.

REQ The data provider is accountable for the correctness, quality and usage of the published meta data about the data that is potentially available in the HERACLES Infrastructure. (A *Data Permit* specifies the scope of data accessibility per research case)

HERACLES FAIR Data Station

A FAIR Data Station is a functional extension of a FAIR Data Point that supports interaction with the data itself, whereas the FAIR Data Point is focused on data descriptions only. A FAIR Data Station is designed to enable digital object owners/publishers to expose the data of their digital objects in a FAIR manner. In essence, a FAIR Data Station provides two things: access to data and an execution environment for certified components to use that data.

Note It can a design decision to combine the functionality of a FAIR Data Point and a Station. It is also a design decision to implement the FAIR Data Point and Station functionality in an existing application or to create a dedicated implementation.

REQ The Data provider of the HERACLES Infrastructure is accountable for a HERACLES FAIR Data Station and can delegate the operational responsibility to a service provider.

For the HERACLES Infrastructure a specialization of the generic description for a FAIR Data Station is defined to meet the main objective [O1].

REQ A HERACLES FAIR Data Station shall only allow data usage in case no information can be linked directly or indirectly to a specific patient to preserve privacy. This can be implemented by with use of valid and granted Data Contract and enforcing with use of the HERACLES Dataspace.

REQ A HERACLES FAIR Data Station shall implement data usage only with use of software components which are certified according to the terms and conditions as specified in the Governance model. In other words, a HERACLES FAIR Data Station provides an execution environment for certified components to use data.

REQ The available components to use data at a HERACLES FAIR Data Station shall be published in a service catalogue which is part of the HERACLES Dataspace.

REQ The execution environment of a HERACLES FAIR Data Station shall have the ability to suspend the execution of data usage even if there is a valid Data Contract.

REQ The data provider is accountable for the correctness, quality and usage of the provided data.

External applications – FAIR (research) application

External applications such as FAIR (research) applications can only be connected to the HERACLES Infrastructure if they are compliant with the following high level requirements:

REQ All applications must be compliant with the terms and conditions in general as stated in the HERACLES Governance model and research case specific in a Data Contract. This is foremost applicable for the legal and organisational constrains but also includes the technical aspects such as the use of certified components for the connection with the HERACLES Infrastructure.

REQ All applications must be technically and semantically interoperable with the five HERACLES functional components as described in this section 3.2 and the use of HERACLES Channels in particular.

Other dataspaces

The two high level requirements for external applications are also applicable for connections with other (Health) dataspaces.

Although the interfaces for an European Health Dataspace are not defined yet and therefore not applicable (yet), should be taken in account for further development.

3.3 Information layer

The Information Layer specifies the Information Model, the domain-agnostic, common language, i.e., Vocabulary of the HERACLES infrastructure. The Information Model is an essential agreement shared by the participants and components of the IDS, facilitating compatibility and interoperability. The primary purpose of this formal model is to enable (semi-)automated exchange of digital resources within a trusted ecosystem of distributed parties, while preserving data sovereignty of Data Owners. The Information Model therefore supports the description, publication and identification of data products and reusable data processing software (both referred to hereinafter as Digital Resources, or simply Resources). Once the relevant Resources are identified, they can be exchanged and consumed via easily discoverable services. Apart from those core commodities, the Information Model describes essential constituents of the HERACLES infrastructure, its participants, its infrastructure components, and its processes.

Scope. The Information Model has been specified at two levels of formalization: at a generic conceptual level and at a healthcare specific semantical level.

Core Model

The following core conceptual information model is used within the HERACLES architecture to describe the core concepts and their relations. These core concepts are visualised in yellow rectangles in the figure below where Identity, Data, Data Permit and Certificate are considered the be the most significant ones. These four concepts in particular are important to establish trust and contribute to required trust framework (See chapter 4).

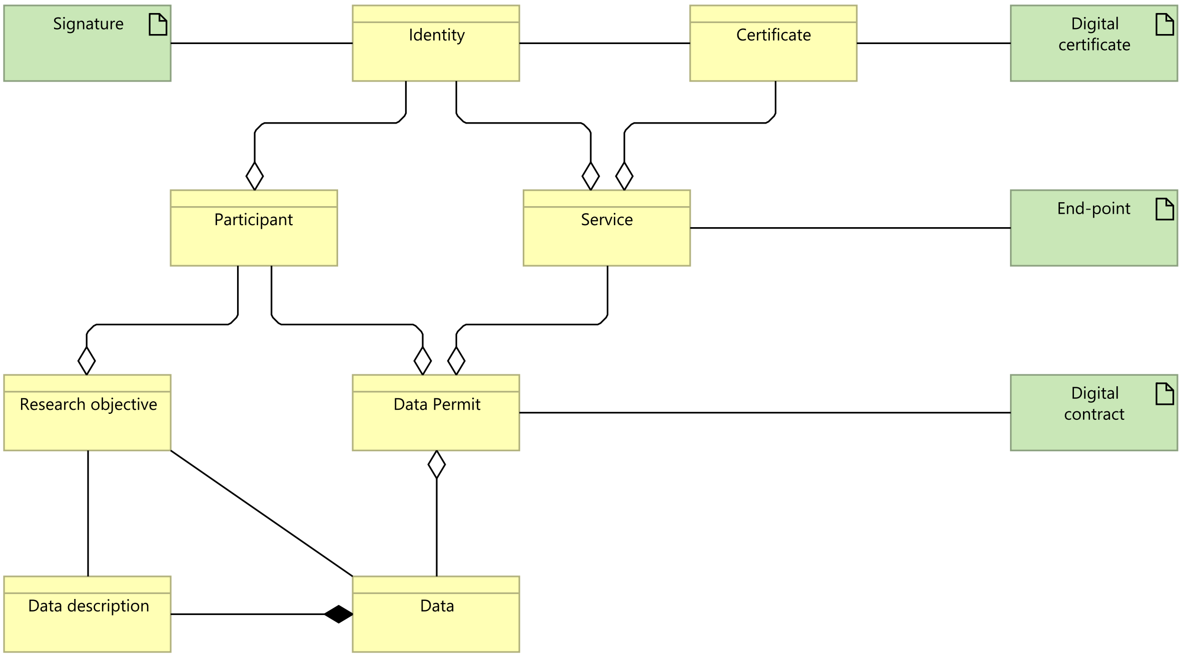


Figure 7 HERACLES core information model.

The green rectangles are digital representations of the related concepts which are used for automatic processing of requests at the system level.

A short description for each of these core concepts is:

- Research objective: a description of the research objective required to assess if there is enough (legal) ground to gain access to the required data;
- Participant: an actor such researcher or data provider as described in the business layer
- Identity: a unique verified (authenticated) identification of a participant or services. The digital representation is a *Self-Sovereign Identity (SSI)* in case of an individual or organisation and a Signature in case of a software service;
- Service: a software services with a verified identity, a valid certification which only performs tasks within the boundaries of a *Data Permit*. The location of a service or its digital presence is an *End-Point*;
- Certificate: proof that the subject is compliant with the terms and conditions stated in the HERACLES Governance model. The *Digital certificate* is the digital representation that can be used to automated the validation of software services;
- Data: the data that is available at a FAIR Data Station to be used for research purposes;
- Data Description: a description at a FAIR Data Point of the available data at a specific FAIR Data Station, required for discovery purposes;
- Data Permit: all the required agreements to fulfil the legal and terms of conditions of all involved stakeholders (participants) which must be met before data can be used for research purposes. The *Digital contract* is the digital representation of a Data Permit which can be used to distribute the agreements among the involved participants and to enable partly automated enforcement of these agreements.

Conceptual representation

For the generic dataspace functionality, the following concepts and relationships from the IDS-Reference architecture are used. See for a detailed description the IDS-RAM v4.0 specification[4] which will not be duplicated in this document. The actor Dataspace Authority is equivalent with the HERACLES Governance board.

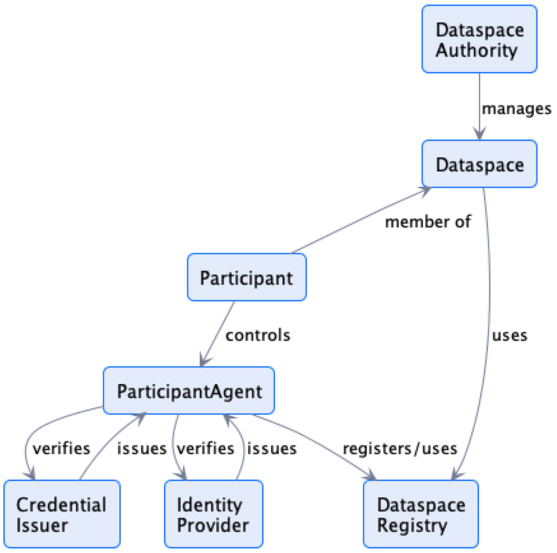


Figure 8 Dataspace entity Relationships (source IDS-RAM 4.0)

The Dataspace registry is a domain independent software services to facilitate the discovery and verification of valid participants, including software services.

The FAIR principles give special attention to metadata which can be defined as data that provides information about other data and includes descriptions about the origin, structure, provenance, rights and obligations or other characteristics also. FAIR Data Point's provides metadata about themselves and, from that point on, the client can navigate its metadata content to discover the other metadata records.

The Catalogue Service in the figure is a functional component to provide discovery functionality for data and services. The Connector represents a functional component to handle the connection with the HERACLES Dataspace and must be implemented as a software service with a known end-point and a valid self-description, signature and certificate. Both components are part of the HERACLES FAIR Data Point to enable discovery functionality via a dataspace.

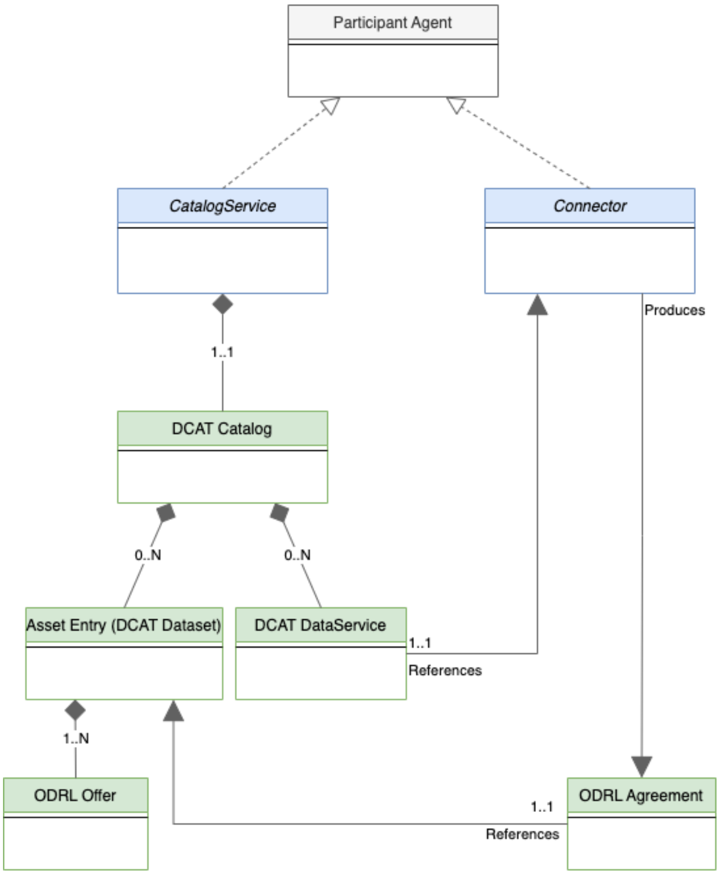


Figure 9 Participant Agent types (source IDS-RAM v4.0)

REQ The FAIR Data Point uses the W3C's Data Catalog Vocabulaire (DCAT) version 2 [13] model as the basis for its metadata content.

The figure below depicts the required extensions in green to the DCAT model according to the FAIR Data Point specifications[14]

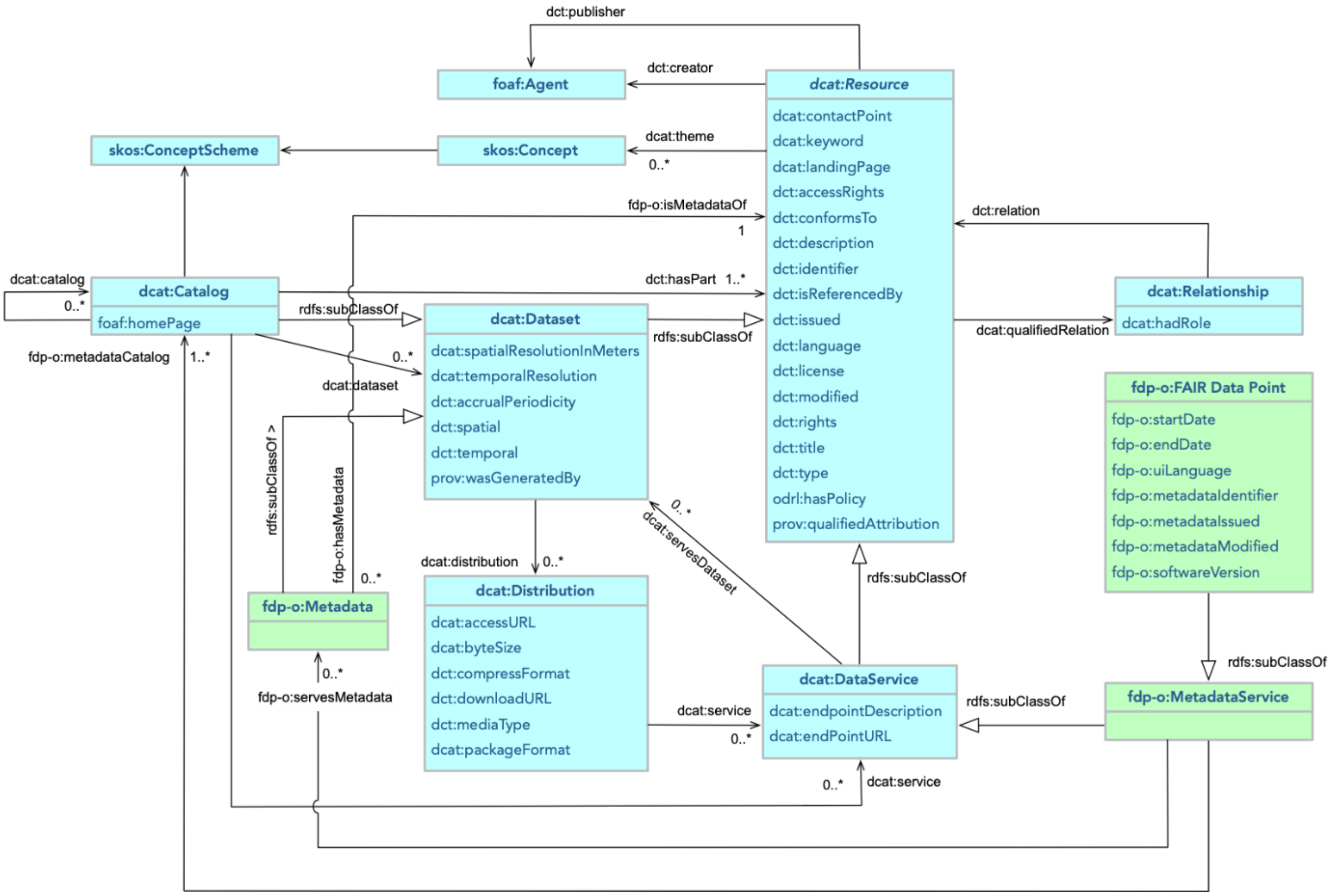


Figure 10 FDP extensions to the DCAT model (source: FAIRDataPoint.org)

The DCAT v2 data specification model is generic and can be used as upper ontology but it needs to be extended with a structured approach of describing the data. This enables a uniform way of exchanging information about the data across technical different implementations of FAIR Data Stations.

REQ JSON-LD (JavaScript Object Notation for Linked Data) is proposed as additional technical standard which is commonly used for describing data, especially in contexts where detailed, domain-specific attributes are crucial.

The next section describes a Health specific data description proposal which has to be proven practically feasible during the implementation of the HERACLES infrastructure.

Relation to Health-specific vocabularies

To become compliant with the **European Health Dataspace (EHDS)** the following data description standards are taken in account derived from the work of the TEHDAS initiative[5] selected on their generic purpose and the application for HERACLES cq. for cancer research in particular[15]

- Observational Medical Outcomes Partnership (OMOP) standard is a standardized data model and methodology developed to enable the aggregation, analysis, and sharing of healthcare information across different healthcare data systems. It aims to facilitate observational research and improve the quality of evidence generated from real-world data by harmonizing diverse datasets into a common format. This standardization allows for more efficient and reliable analysis of healthcare data for research, including the study of drug safety, effectiveness, and health outcomes.
- Fast Healthcare Interoperability Resources (FHIR) standard, developed by Health Level Seven International (HL7), is a framework designed to enable healthcare information to be exchanged more easily between different healthcare systems. FHIR leverages modern web technologies and standards to provide a comprehensive, robust, and flexible model for data representation and exchange. It is based on modular components called "resources," which can be used in a standalone or aggregated manner to address a wide variety of healthcare data exchange scenarios. These resources cover clinical, administrative, and infrastructural data domains, facilitating not just the exchange of electronic health records (EHRs), but also supporting a broad range of applications including those in mobile apps, cloud services, data analysis, and more. FHIR aims to simplify implementation without sacrificing information integrity and supports RESTful architectures, making it highly adaptable to the evolving technologies in healthcare informatics.

The Dutch **Health Research Infrastructure (Health-RI)** initiative aims to facilitate and promote the use of standardized data to enhance health research and innovation. While specific details about the standards promoted can evolve over time, Health-RI typically emphasizes the adoption of international and national data standards that ensure interoperability, data quality, and the efficient exchange of health information. Common standards and frameworks that initiatives like Health-RI promote the use of FL7 FHIR and OMOP Common Data Model (OMOP CMD) as well. Besides several ISO Standards relevant to health informatics and data management, Health RI also promotes the use of:

- Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED CT): a multilingual, comprehensive, healthcare terminology standard. It provides a systematic language that precisely represents medical terms used in healthcare documentation. Its intent is to represent clinical concepts across many domains, which includes conditions, diagnoses, symptoms, and signs, all of which are a type of finding and facilitates the accurate recording, retrieval, and analysis of health information.

There are mappings between the FHIR and OMOP standards but a practical obstacle to overcome is the fact that not all data sources are compliant with either one of these two standards. This is not a direct problem as long as the data is a known description available.

This results in the following practical requirement but very important generic requirement.

[5] THEDAS - Identification of relevant standards and data models for semantic harmonization

REQ In order to be findable and semantically interoperable, the data used within the HERACLES Infrastructure must be specified in terms of:

- Syntax and Semantics: to define a clear syntax for how data is structured and semantics for what the data means. This ensures that data is not only correctly formatted but also universally understood in terms of its content and context;
- Data Elements and Types: to specify the data elements and types, including how numerical values, text, dates, and other types of data are represented. This ensures consistency in how different kinds of information are encoded;
- Encoding and Serialization: how data is encoded or serialized for storage and transmission. This might include formats like XML, JSON, or others that dictate how data is converted into a format that can be easily shared or processed by computers;
- Terminology and Ontologies: To promote shared understanding and interoperability, a reference to a standard or controlled vocabularies, terminologies, and ontologies that define and relate concepts within a particular domain;
- Data Integrity and Security: to address aspects of data integrity and security, ensuring that data remains accurate, consistent, and protected from unauthorized access or alterations;
- Origin and purpose: to describe the data's origin, purpose, characteristics, and structure. This is crucial for data discovery, understanding, and reuse;
- Compliance and Governance: to align with relevant regulations, laws, and governance frameworks to ensure legal compliance and ethical use of data.

FAIR Data Points shall provide these aspects of data descriptions as categorised above.

Note It is strongly recommended (but not mandatory) to transform the data to a European or better a global semantic standard to benefit from:

- Interoperability: Standards facilitate data exchange between different systems and applications, enabling seamless sharing and integration of data across diverse healthcare environments.
- Scalability: Standardized data can be more easily expanded or adapted to new requirements or technologies without being tied to the limitations of a specific application.
- Data Quality and Consistency: Standards ensure that data conforms to a recognized format, improving its accuracy, consistency, and reliability across different platforms and use cases.
- Efficiency and Cost-effectiveness: Using standards reduces the need for custom integration solutions, lowering the time and cost associated with data sharing and processing.
- Innovation and Collaboration: Standardized data fosters innovation by making it easier for developers to create new applications and for researchers to collaborate and share findings, leveraging a common data language.

Overall, data standardization enhances the ability to aggregate, analyse, and apply data effectively, supporting better decision-making, research, and patient care outcomes in the healthcare sector.

With DCAT as upper ontology, the following additional description approach is proposed to describe datasets in terms of a common vocabulary, predefined tables and data formats.

Common Vocabulary:

Controlled Terminology: One of the key features of OMOP CDM is its use of a standardized, controlled vocabulary to describe medical terms, procedures, drugs, and more. This vocabulary standardizes the way healthcare concepts are represented across different data sources, turning disparate terminologies into a unified language.

REQ All concepts used for the description of dataset should be described in a vocabulary known to a FAIR Data Point and preferrably refer to a standardized vocabulary such as provided by OMOP.

Each entry in the controlled vocabulary is assigned to a unique Concept ID. These Concept IDs are used across the (predefined and use-case specific) tables to ensure that data from different sources can be used seamlessly.

REQ The common vocabulary also defines hierarchical relationships between concepts, allowing for the aggregation and analysis of data at different levels of specificity. For example, a specific brand of a medication can be rolled up to its generic form, or a particular type of cancer can be considered under the broader category of neoplasms.

Predefined tables

Tables are a generic type of datasets which can be described in terms of columns with fieldnames corresponding descriptions per column. The Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) uses a set of predefined tables to organize data. Each table serves a specific purpose, such as storing information about patients (Person table), visits (Visit Occurrence table), drug prescriptions (Drug Exposure table), conditions (Condition Occurrence table), and measurements (Measurement table), among others. This structured approach ensures consistency in how data is stored across different databases. Therefore the following predefined tables are proposed to be used within a FAIR Data Point:

- Person: contains demographic information about individuals in the database, such as gender and birth date. See section about Common patient identifier!
- Visit Occurrence: records each instance of a patient interacting with a healthcare provider, detailing the type of visit (inpatient, outpatient, emergency, etc.), and the start and end dates.
- Condition Occurrence: captures instances of medical conditions diagnosed or reported, linked to specific Concept IDs in the standardized vocabulary.
- Drug Exposure: documents the provision of a drug to a patient, with details on the drug type, dosage, route of administration, and duration.
- Procedure Occurrence: records medical procedures performed on patients, identified by standardized Concept IDs.
- Observation: contains observations or measurements about a patient, which could be clinical findings, patient-reported outcomes, or social determinants of health.
- Measurement: stores quantitative measurements or lab test results, such as blood pressure readings or blood glucose levels.
- Device Exposure: details about the use or exposure to a medical device during care.
- Death: Captures information on the death of individuals, including the date and cause of death when available.
- Cost: records the cost associated with healthcare events, such as procedures, drugs, or visits, providing insights into the financial aspects of care.

- Care Site: describes the locations where care is provided, such as hospitals, clinics, or community care centres.
- Provider: contains information about healthcare providers, including their specialty and the care site where they practice.
- Cohort: used to define groups of individuals for analysis based on specified criteria, supporting cohort studies within the database.
- Cohort Definition: stores the logic or criteria used to define cohorts, facilitating reproducibility of research studies.
- Condition Era and Drug Era: aggregate information from the Condition Occurrence and Drug Exposure tables to represent periods of time when the patient was known to have a particular condition or was exposed to a specific drug, respectively.

REQ If applicable the predefined tables as defined within OMOP CDM shall be used to describe the content of datasets in the form of tables. For example, if there is no data available about a specific topic such as death, that the related concepts and table to that topic are not required nor mandatory.

Note In case the data is not OMOP compliant, the description of the dataset can still be used but the "concept ID's" shall refer to another preferably standardised vocabulary being part of the *common vocabulary*.

Common Format scheme:

A common format scheme is required to standardize the format and data types of each data field. This schema dictates, for example, the format of dates, the representation of numeric values, and the use of specific codes for various entries (like condition codes, drug codes, etc.). This uniformity is crucial for enabling effective data integration and comparison across studies and sources.

Note A specification of such common format scheme will be included in an updated version 2 of this architectural design document.

Patient journeys

In the realm of data management and analysis, two types of data portioning —horizontal and vertical—play a crucial role in combining datasets from different sources:

- *Horizontally-partitioned data*, where two or more organizations record similar data items but for different data subjects.. An example is the combination of data from multiple cancer registries that cover different geographies and patients. Combining cancer registry data allows for inter-geographical comparisons and creates a large patient volume. The latter is particularly relevant for the research on rare cancers. As databases contain data from different patients, matching identifiers between databases is not a concern for horizontally partitioned data.
- *Vertically-partitioned data*, where data items for a group of individuals are distributed across several databases. For example, the data items on cancer patients in a cancer registry and the data items recorded by an insurance provider. For vertically partitioned data, the identifiers of the patient records should be matched across databases.

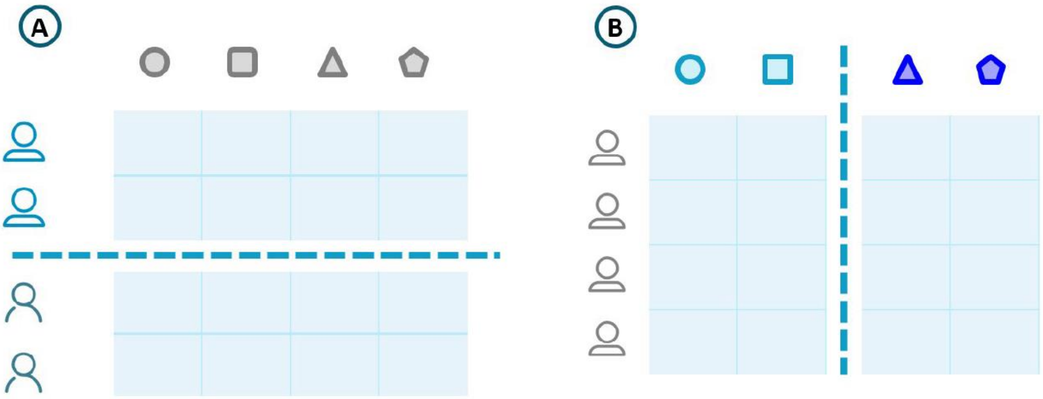


Figure 11 Horizontal and vertical partitioned data

(A) Horizontally-partitioned data: *same features from different patients,*
(B) Vertically-partitioned data: *different features with the same patients*.

To create a patient journey based on information from multiple data sources (situation B) requires a common to related the information about the same patient. To preserve the patients privacy, a common unique identifier such as the national social civil identification number (BSN) is not allowed from a legal point of view. This raises two challenges: finding the same person to enable a join and avoiding identification of that person.

The HERACES architecture distinguishes two approaches to cope with these challenges: probabilistic matching and the use of pseudonymized Identifiers.

Probabilistic matching

Probabilistic matching is a sophisticated technique employed to align multiple data sources containing varied information about the same group of patients, especially useful for executing horizontal joins. This method calculates the likelihood that different records across these datasets refer to the same individual, despite discrepancies or missing data. By evaluating the similarity of multiple attributes through statistical probabilities, probabilistic matching enables the integration of diverse patient data into a single, enriched dataset. This approach is indispensable for researchers and healthcare professionals seeking to consolidate fragmented healthcare information, thereby enhancing the completeness and accuracy of patient profiles for comprehensive analysis.

For example: the join of the datasets can based on a probabilistic (Bayesians) method for the following attributes:

- First character of the first name
- First two characters of the first name
- First letter last name(IJ & Y)
- Soundex last name (IJ & Y)
- Postal Code numbers
- Postal Code characters (in case postcode numbers are identical)

This might help to overcome the problem of different ways of registering a person across different datasets.

Pseudonymized Identifier

A method of creating pseudonymized identifiers is part of a broader approach to protect patient privacy while enabling the linkage of patient data across different healthcare datasets for research and analysis purposes.

This pseudonymization technique is designed to:

- Reduce the risk of re-identification: By not using direct identifiers such as the full name or BSN, the risk that someone could re-identify the individual from the data is minimized.
- Enable data linkage: The pseudonymized identifier allows for the aggregation of data from different sources about the same individual without revealing their identity. This is crucial for longitudinal studies, population health research, and other forms of health data analysis.

For example, a pseudonymized identifier can be constructed using a set of attributes known by all parties involved:

- the first characters of the surname and given name (XX)
- concatenated with date of birth (YYYYMMDD)
- and the postal code (9999)
- the gender (Z)

This results in the following format: XXYYYYMMDD9999Z which will be further scrambled by using a common hash function.

Note A pseudonymized identifier can be used as “person_id” in the proposed predefined table descriptions. As mentioned in the section about the predefined tables, the predefined table “Person” must be used in a modified form as measure to preserve the patient privacy.

The Person table in the OMOP Common Data Model (CDM) is designed to store demographic information about the individuals whose data is being captured in the database. For HERACLES only a subset is proposed, limited to the fields required for linking and the analysis.

Note The proposed predefined table Person is used to describe the available data sets and to determine what the possibilities are to construct a common probabilistic / pseudonymized identifier, because that is directly depending on the available attributes.

The content of the predefined table Person itself **will NOT be shared** across domains.

Note Execution of practical use-cases should determine which assisting functionality is required for each use-case and whether there is more research is required to facilitate a standardised approach by the HERACLES infrastructure.

3.4 Process layer

The Process Layer specifies the interactions taking place between the different components of the HERACLES infrastructure. It thereby provides a dynamic view of the architecture. The process layer distinguishes the generic and the healthcare and HERACLES specific aspects.

To be compatible with EHDS, the HERACLES user process flow should be compatible with the following revised user journey of EHDS[15].

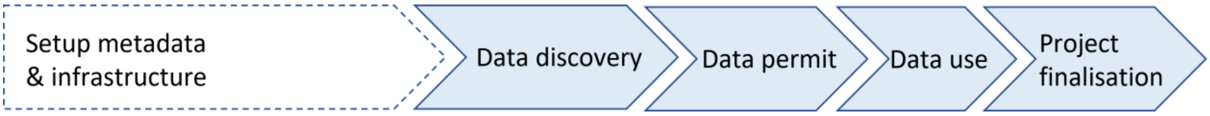


Figure 12 EHDS user journey phases

A complete description of these user phases can be found here [15]. The Setup metadata & infrastructure phase, is considered by EDHS not part of the user journey but is identified as a required and enabling step. In HERACLES there is a strong focus on establishing trust, identification management and certification which are part of this first step.

To explain more clearly the content of these processes and to reuse the dataspace processes defined in the IDSA Reference Architecture (IDS-RAM), the process names in the HERACLES infrastructure deviates slightly from the EHDS phase names, but their functionality is compatible with EHDS.

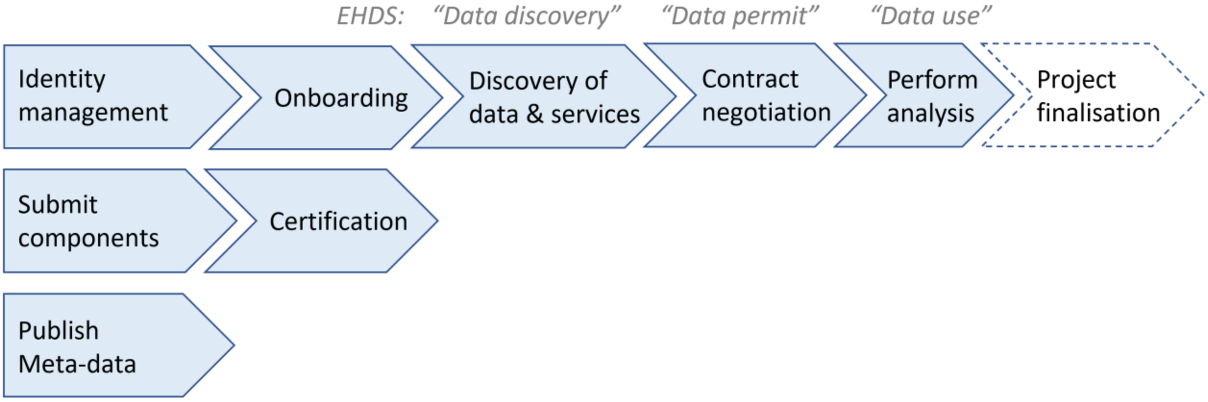


Figure 13 HERACLES user/research journey processes

The first five processes on left in the HERACLES user/research journey are considered to be generic processes and overlapping with the initial setup metadata & infrastructure step of EHDS. The HERACLES infrastructure focuses first on creating the proper conditions to enable the use of data. Supporting the project finalisation process has a lower priority.

Note The processes Submit components, Certification, Publish meta-data are enabling the main user/research journey line therefore visualised beside the centre process flow. Identity management is in general the first step and applicable for all other processes because only participants who are successfully onboarded can be valid actors. Therefore the identity management and onboarding processes are described first.

Identity management

Identity management is a critical aspect of information security and online services, dealing with the identification, authentication, and management of individuals' digital identities. The following key-aspects are taken in account:

- Identification, the process of recognizing an actor or entity within a system. This involves a unique identifier to be stored in a register as a core functionality of the HERACLES Dataspace;
- Authentication, the process of verifying that an individual or entity is who they claim to be. This is achieved through *verifiable credentials* based on Self Sovereign Identities which is a core functionality of the HERACLES Dataspace;
- Verifiable Credentials are digital documents that are used to establish trust in the identity and attributes (claims) of an individual or entity. They are issued by a trusted authority and can be securely stored and presented by the individual when required.

Overall, identity management is essential for controlling access to resources in within the HERACLES infrastructure, ensuring that users are accurately identified, authenticated, and granted appropriate access based on verifiable credentials.

The IDS-RAM[4] describes in more detail the Identity management process.

Onboarding

After acquiring verifiable credentials, a user applies for registration via the onboarding process. The objective of this process is to admit only users with verified credentials and are compliant with the Governance model. The following steps are identified:

- User Registration, the user initiates the process by providing verifiable credentials as result of the previously described Identification management process. The self-sovereign identities contain the meta-information that the user wants to share;
- Verification of Verifiable Credentials, the provided verifiable credentials are verified by checking the digital signatures or cryptographic proofs against the issuing authority's public key to ensure authenticity and validity. Valid issuing authorities should be described in the Governance model. Verification of verifiable credentials is a core function of the HERACLES Dataspace;
- Compliance with Terms and Conditions, the user is presented with the terms and conditions of the governance model of the system. This may include privacy policies, user responsibilities, and other regulatory compliances. The user must agree to these terms, often by checking a box or providing digital consent
- Assessment of Compliance, the system assesses if the user meets all the specified criteria and requirements as per the governance model. This may involve automated checks or manual review processes;
- Final Registration, if the user successfully verifies their credentials and agrees to the terms and conditions, the registration process will be completed and the user information will be stored securely, making them a recognized and authenticated entity within the system.

The result of a successful onboarding process is that the user's information is made findable within the system, subject to privacy settings and access controls. This means they can be identified and authenticated in future interactions, and they may be granted access to specific resources or services within the system. This process ensures that only verified and compliant users are onboarded, maintaining the integrity and security of the system.

Submit components

The objective of this process is to identify software components and to make them findable. It starts with assigning a unique identification to a software components which will be registered in a so called service catalogue, including meta data to describe the software component or service. This identification shall be linked to a verifiable signature of the software/service component.

This process is equivalent to the “Publishing and using Data Apps” process as described by IDS-RAM[4] with one exception, an App store for Data Apps or software components is not foreseen within the scope of the HERACLES project. Instead, meta-data of software components including their identification and signature are stored in the service catalogue which is a core functionality of the HERACLES Dataspace. The following three step approach is proposed:

- Collection and Preparation: This step starts with the software or service provider and involves gathering all the necessary meta-data including a description, categories, keywords for search optimization, a unique verifiable signature and privacy policies. The preparation phase involves organizing it in a structured format and ensuring it aligns with the schema as described in the previous section 3.3 – Information layer.
- Submission and Review: Once the software component or service and its meta-data are ready, the information is submitted to the service catalogue of the HERALCES infrastructure. After submission, the Governance board is accountable for a review process to checks if the provided information is compliant with the terms and conditions as stated in the HERACLES governance model regarding compliance with the guidelines, standards, technical requirements, and legal issues. After successful review the status of a software component or service changes to “candidate” and the certification process can start. Otherwise the submission will be rejected and the provider has to optimize the previous step.
- Post-submission Activities: After a successfully certification process, the post-submission activities include gathering user feedback, releasing updates to fix bugs, improve functionality, or add new features. All changes require a re-certification which practically means a submission of a new version and performing the preceding step of submission and review.

Overall, the submitted components (software or services) will only be visible to the HERACLES Infrastructure users after a successful certification process.

Note A Future optimisation can be the automation of distribution and deployment of certified software components from a HERACLES App store, to lower the technical barrier but this is not in scope of the HERACLES project.

Certification

The objective of this process is to verify that a software component or service is compliant with the terms and conditions stated in the Governance model. The role of auditor and the procedure how to perform the verification has to be described in the Governance model as well.

The certification process is successfully applied if all participants of the HERACLES infrastructure can trust a certified software component: that it preserve the privacy and security aspects and in general respect the *terms and conditions as defined in the Governance model*.

The result of a successful certification is a *verifiable certificate* linked to the identification and a verifiable signature of the software component or service. Practically it means that the certificate will be added to the service catalogue.

The IDS-RAM[4] describes how the certification processes for organisations and connectors are organised to verify IDS compliancy in more detail. Guidelines for certification are included in the IDS Rulebook [6]. For the HERACLES Infrastructure

The certification of software components or services by an evaluation authority involves a detailed and methodical process. This process ensures that the software meets specific standards of quality, security, and compliance. The following steps are identified:

- Detailed Audit and Testing: The auditing party conducts a thorough examination of the software. This includes code reviews, security audits, performance testing, and compliance checks with relevant standards and regulations. The auditing process is comprehensive and aims to identify any vulnerabilities, defects, or non-compliance issues. The auditor may use a variety of tools and methodologies to assess the software. If issues are identified, the developer is typically given an opportunity to address them and resubmit the software for reassessment.
- Certification and Issuance of Certificate: Once the software passes the audit and meets all the required standards, the auditing party issues a certificate. This certificate is linked to a unique identity and signature of the software, including details like the version number and release date. The certificate also features a verifiable signature from the auditing party, ensuring the authenticity and integrity of the certification. The software provider can then use this certification to demonstrate the reliability and compliance of their software to clients, users, or regulatory bodies.
- Publication Activities: After a successfully certification process the Governance board will publish the software component/service with a verifiable certificate in the service catalogue to become visible and findable within the HERACLES Infrastructure.
- Post-Certification: After receiving certification, the software provider must maintain the standards for which the certification was awarded. This may include regular updates, patches, and re-audits, especially if significant changes are made to the software or new regulations come into effect. The software provider should also monitor the software for any emerging threats or vulnerabilities and take prompt action to address them, ensuring continued compliance and security.

Overall, for the HERACLES Infrastructure, the certification process is an essential process to contribute to “trust” and mainly an organisational rather than a technical aspect. See also section 4.2 which describes certification as one of the three cross-layer perspectives.

Publish meta-data

After successful onboarding, data providers can publish information (meta-data) via a FAIR Data Point about what kind of data they provide including the terms and conditions for usage for the meta-data itself as well for the data to be provided.

Publishing metadata involves several key steps to ensure that the information is accurate, accessible, and useful. The following steps are identified:

- Collection and Preparation: This step starts with the data provider and involves gathering all the necessary meta-data. The preparation phase involves cleaning the meta-data, organizing it in a structured format, and ensuring it aligns with the standards or schema as described in the previous section 3.3 – Information layer;
- Metadata Creation: In this step, the metadata is being uploaded to the HERACLES Infrastructure in Data Fair Points. This includes adding the identity of the provider references to already defined / existing standards if applicable;
- Quality Assurance and Validation: Before publishing, it's crucial to ensure that the metadata is accurate, consistent, and adheres to the defined standards. This quality

assurance might involve manual checks or automated validation processes. It's important to ensure that the metadata accurately represents the data it describes and is free from errors;

- Publication and Maintenance: The final step is to publish (release) the metadata so that it can be accessed by users or systems as part of the discovery process. After publication, regular maintenance is essential to keep the metadata up-to-date and relevant. This might involve updating records to reflect changes in the underlying data, adding new metadata as new data becomes available or correcting any errors that are discovered after publication.

Overall, the availability and the quality of meta-data are crucial for collaboration within the HERALCES Infrastructure and facilitates this via the concept of *FAIR Data Points*.

Discovery of data & services

This process is compliant with the EHDS Data Discovery phase of the EHDS users' journey to request access to health data for secondary use. In this phase/process, researchers should be able to search the data available and needed to perform their work. For the HERACLES infrastructure this process is extended to finding supporting components (software/services) as well for specialized operations such as the use of Privacy Enhanced Technologies (PET).

This process depends fully on the previously described processes about Submitting Components & Services and Publishing meta-data. The following simplified steps are identified:

- Discovery of data: identifying and cataloguing various data sources, discovering the quality and availability of data resources. The outcome is a specification of data the researchers plan to use;
- Refine research objective and define approach: based on the available data (sources), the quality of data and the ability to join them, researchers can refine their plans. The research plan is required as justification to acquire permission to use the data but also specify how to use the data;
- Discovery of supporting components: helps researchers to identify tools, services, and resources that can aid in data analysis, visualization, and management. This can include tools for data cleaning, data mining, and data visualization. Within the HERACLES project the main focus will on components to use data while preserving the privacy of the patient records.

The outcome is a plan that specifies which data from which data sources are needed, how this data will be used and which components will be used to obtain permission in the next phase.

Contract Negotiation (request data permit),

This process is compliant the EHDS Data Permit phase, where researcher request data usage permission, and the governance board and all involved data providers verify compliancy with their usage policies. The following steps are identified:

- Request data permit application: a request contains a verified identity of the requester and a research plan which includes the research objective, a specification of the data that is needed and a description how the data is being processed;
- Request assessment: the Governance Board is accountable to verify compliancy of the request (c.q. the research plan) with legislation and the Governance model. This implies that the Governance model shall contain terms and conditions for data usage including applicable laws. The Governance Board is responsible to organise an operational process to perform the assessment in collaboration with the involved data providers. This operational process produces a compliancy list and a Data Protection Impact Assessment (DPIA);

- Contract negotiation: based on the compliancy list and the DPIA, omission, deviations and recommendations are return to the requester. In case the request is not compliant, a negotiation will be started to find out if the request can be altered or completed to become compliant with the law and the commonly agreed principles as stated in the Governance model;
- Grant permit or reject: in case a request is granted, the outcome is an agreed research plan, a DPIA and a technical permit in the form of an electronic contract. In case of a rejection, an explanation will be returned.

Overall, the Governance board is accountable for the contract negotiation process and therefore also for the permission. Involvement of the data providers in the assessment step is essential to enable data sovereignty and to ensure proper use of sensitive data.

The Data Permit phase c.q. this process of contract negotiation is currently a cumbersome and time-consuming procedure. Lead times of two years are not an exception but a huge obstacle to overcome for efficient and effect medical research. One of the HERACLES project objectives is to propose a procedure with a significant reduction in lead time. The last part of this section 3.4 describes how the HERACLES Infrastructure can facilitate and speed-up this process of contract negotiation.

Perform analysis (data use)

This process is similar to the EHDS Data Use phase and the Exchanging Data process of the IDS-RAM[4]. A different and more specific name has been chosen for this process, to avoid confusing about the concept of data sharing and data exchange. In the HERACLES Infrastructure, only anonymous data will be exchanged as results of the analysis to preserve the privacy and sensitive data of patients via so called "*HERACLES channels*". A HERACLES channel is similar to a VPN connection and provides secure and authorized use of data between multiple end-points such as FAIR Data Points & Stations and is always bounded to a related Data Permit (contract).

To enable a patient journey based on multiple data sources, Privacy Enhanced Technologies (PET) can be used to implement HERACLES channels. The following steps are identified:

- Verify and Control: after permission has been granted, the HERACLES Infrastructure verifies the identities of all involved actors, software components software services based on verifiable: credentials, certificates and permits. After successful verification, the use of the requested HERACLES channel, including the predefined and approved configuration and required processing, will be authorized;
- Authorise channel: this is an intermediate step to forward the authorization of a HERACLES channel to the involved software components and software services;
- Use channel: after receiving the authorisation information, a HERACLES channel can be used for a limited amount of time and only for the assigned researchers and specified data as dictated by the permit/contract;
- Close channel: when processing is completed or after a premature termination, the channel will be closed.

Overall, to preserve the privacy of patients the HERACLES Infrastructure facilitates the use of data via the concept of *FAIR Data Stations* based on secure and controlled *HERACLES channels* bounded by a Data permit (contract).

Note The proposed system architecture is processing agnostics which means that any kind of processing and data collaboration protocol can be used as long as the components are certified (trusted by the involved stakeholders and data providers/custodians in particular), processing is compliant to terms and conditions of the Data Permit and the Dataspace protocol is used to control the processing.

HERACLES specific process implementations

Although the previous process description refers multiple times to the HERACLES Infrastructure, FAIR Data Points/Stations and HERACLES channels, these processes are generic and can also be applied in other sectors and domains. However, the proposed process flow has not been implemented yet and is therefore a subject of research within the HERALCES project, especially the aspects of governance. This part describes additional process descriptions which are HERACLES specific in order to achieve an optimization of the current way of working: to reduce time and effort before data from multiple sources can be used.

This is work in progress will be completed in an updated version (2) of this document. The following processes are expected the have a HERACLES specific extension:

- Publish meta-data / Discovery of data & services: FAIR + usage policies to know in advance the terms and conditions for getting access to data;
- Contract negotiation: use of templates can speed up the process (to set a precedent) and even automate the assessment. This process results in a *Data Permit* and an electronic version for automated support within the HERACLES Infrastructure;
- HERACLES channels: use of automated verification of certificates and *Data Permits* to ensure controlled access.

3.5 System layer

The system layer specifies the technical components and their interactions. At this level interfaces and protocols are defined independent from the execution environment or programming language. A common practise is to distinguish four system sub-layers to separate concerns and to reduce complexity.

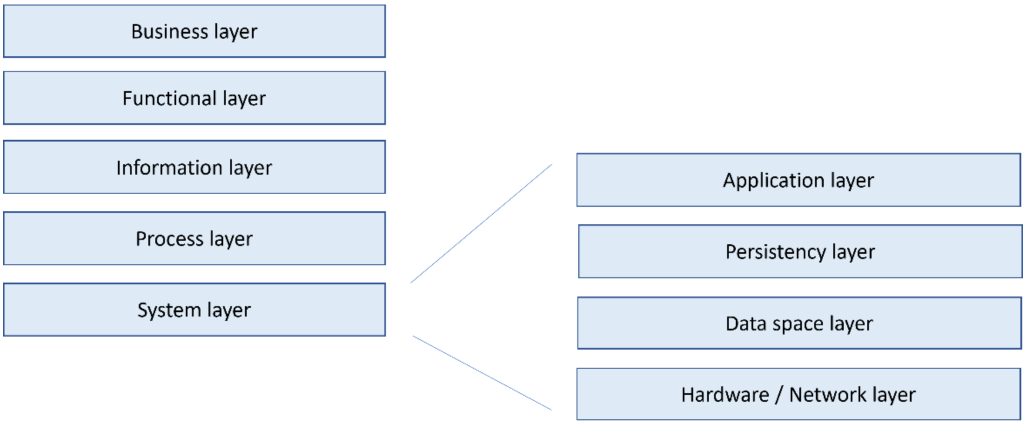


Figure 14 HERACLES System layers

The following three system sub-layers are described in this section:

- Application or software service layer where data is in use.
- Persistency or storage layer where data is at rest;
- Data space layer where data is in transit.

Note There are no specific requirements and design decisions for the hardware/network system sub-layer and therefore that layer is not described in this document. This sub-layer might become of interest in case of delegation to a service provider. This might require additional security and governance aspects to preserve privacy and security of the data.

Application layer

The figure below visualise that the HERACLES infrastructure is used by four types of applications:

1. *FAIR Data Points*: to provide meta data for discovery purposes;
2. *FAIR Data Stations*: to provide access to medical data via FAIR Data Stations;
3. *Healthcare research*: to analyse the collected data with tooling and;
4. *Dataspace control*: to manage and monitor the HERACLES Dataspace as part of the governance tasks.

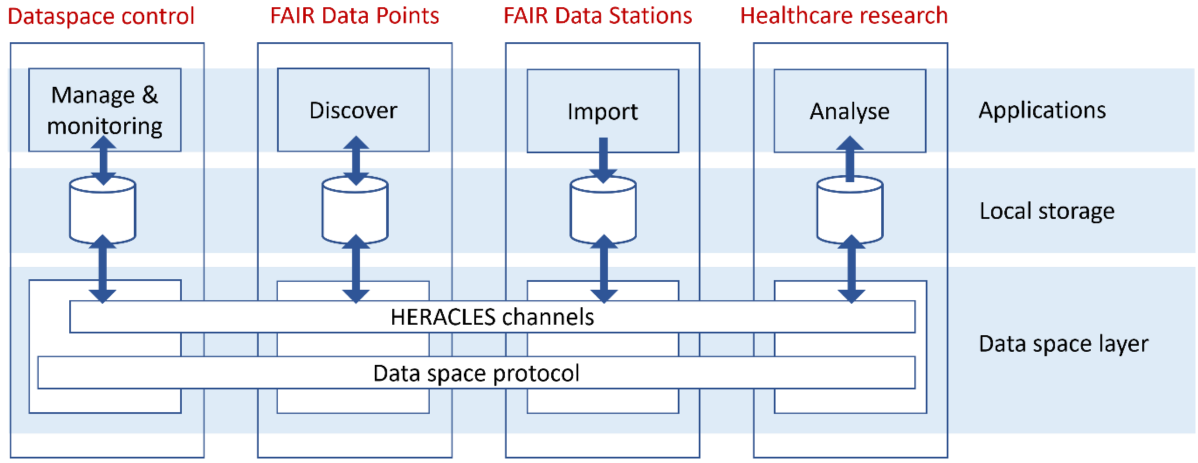


Figure 15 HERACLES four types of applications

The functional requirements for FAIR Data Points and Stations are described in section 3.2. At System level the following technical requirements are identified to enable the communication via the Data space layer:

REQ All four types of applications shall implement the *dataspace protocol* as specified in the Data space layer to enable the use of HERACLES channels, which is required to use data. This includes the implementation of:

- the *dataspace catalogue protocol* to publish meta data and discovery of data;
- the *dataspace data usage protocol* to control processing data locally.

REQ All four types of applications shall implement HERACLES channels for using data.

Note The design idea behind these system application requirements is that existing applications can be used by extending them with standardized API's, to reduce the development effort instead of building complete new applications. It also leaves open the option to implement use case specific API's. Therefore, 100% standardized API's is an objective but currently not (yet) realistic.

Persistency/Storage layer

This System sub-layer is introduced to visualise explicitly that information is being stored within each component of the HERACLES infrastructure but not within the HERACLES dataspace c.q. the dataspace layer. At System level the following system persistency requirement is identified:

REQ All data is only and exclusively accessible for processing via HERACLES channels. In other words, data of a data provider can only be used under the terms and conditions of a Data Permit.

Note The proposed system architecture is storage technology agnostics. This means that the design decisions for specific implementations of a HERACLES channel might result in additional system storage requirements.

Data space layer

The proposed design of this system sub-layer is not specific for the Health domain and implements generic functionality as described in section 3.2 such as identification, authentication, authorisation and usage restrictions. The purpose of the Data space layer is to connect and facilitate the use of data between registered and certified end-points.

The Data space layer can be divided into a *control plane* and a *data plane* as described in the upcoming EU Dataspace protocol[16].

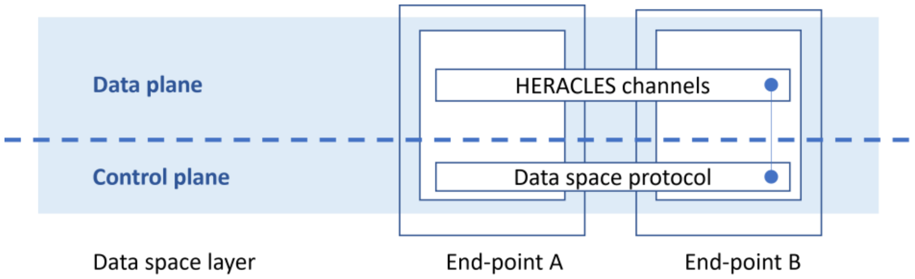


Figure 16 HERACLES Data space layer

In the *control plane* all involved end-points must first agree on access to data resources based on a standard negotiation process between the involved end-points based on an agreed data contract (Data Permit). The Dataspace protocol describes the technical API between the end-points to be implement as part of the control plane. When all involved end-points have agreed that information can be exchanged, each end-point will delegate the transfer process of data and the reception of data to the *data plane*.

In laymen terms: the Control plane checks if terms and conditions as descripted in the Data Permit are met and if so, authorizes the corresponding HERACLES channel in the Data plane. This enables local processing of and collaboration with other services connected to the same HERACLES channel and therefore under strict terms and conditions of the same data contract.

REQ The Data space layer implements the EU Dataspace protocol standard[16]

The Dataspace protocol defines several APIs such as the Catalogue API and the Transfer Process protocol. The Transfer Process protocol results in this document in a Data Usage API which has the same specification but is intentionally renamed to address that it is more then transfer of data only (Data to the algorithm) but also supports local processing (Algorithm to the Data).

For the implementation of the Data space protocol, a so called Dataspace Connector is used. An existing software component that implements the Data space protocol is the IDS connector based on IDS-RAM version 4.0. However the choice of an IDS connector is an implementation design choice made on availability of the Open Source software[x] and the expertise about this component within the HERACLES consortium. Another type of dataspace connector such as the Eclipse Dataspace Connector (EDC) can be used as well as long as it is compliant to the Dataspace protocol standard, which is at time of writing not yet the case.

Note The use of an IDS compliant connector is therefore not a requirement but a practical implementation choice.

The dataspace specific processes as identified bij IDS-RAM[4] are summarized in the figure below as interactions between the dataspace components. Please note that the Identity Provider is not shown in the figure in order to maintain readability.

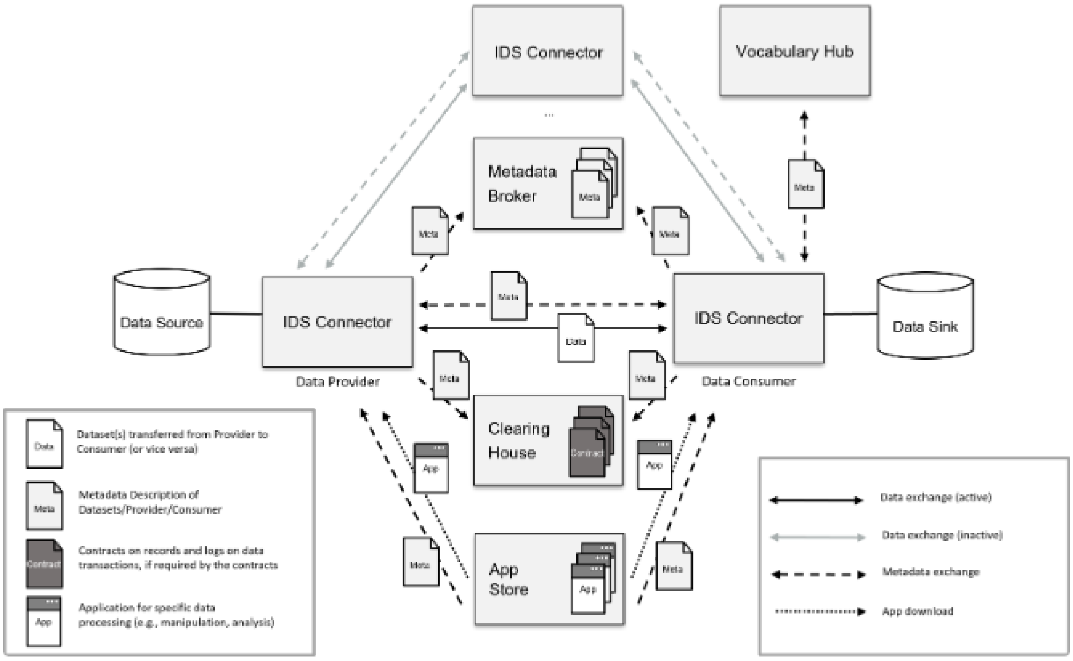

IDS Connector
Vocabulary Hub
...
Metadata Broker
Meta
Data Source
IDS Connector
Data Provider
Data
IDS Connector
Data Consumer
Data Sink
Clearing House
Contract
App
App Store
Data
Dataset(s) transferred from Provider to Consumer (or vice versa)
Meta
Metadata Description of Datasets/Provider/Consumer
Contract
Contracts on records and logs on data transactions, if required by the contracts
App
Application for specific data processing (e.g., manipulation, analysis)
Data exchange (active)
Data exchange (inactive)
Metadata exchange
App download

Figure 17 Interaction of the technical dataspace components (source IDS-RAM 4.0)

A distributed network like the HERACLES infrastructure relies on the connection of different participants where IDS Connectors or other core components are hosted (an IDS Connector comprising one or more Data Endpoints). The IDS Connector is responsible initiating the connection from and to the internal data resources and enterprise systems of the participating organizations and the International Data Spaces. It provides metadata about the network to the Metadata Broker as specified in the IDS Connector self-description, e.g. technical interface description, authentication mechanism, and associated data usage policies. A Clearing House, App store and Vocabulary hub are not foreseen in the first version of the HERACLES infrastructure.

REQ The IDS compliant HERACLES infrastructure consists of the following core components:

- the Identity Provider,
- the IDS Connector,
- the Metadata Broker,
- Data Apps and the *App Store,*
- *the Clearing House, and*
- *the Vocabulary Hub.*

A detailed description about these components can be found here: IDS-RAM v4.0[4] and the grey-out and *italic* components are not foreseen in de first implementation of the HERCLES Infrastructure but are required technical components for a controlled and sustainable solution.

Data Usage API

The Data Usage API can be used for components with verifiable attributes such as a unique and known ID, signature and certificate which can perform strict and limited operations which are known in advance based on an agreed contract (Data Permit) with policies and procedures.

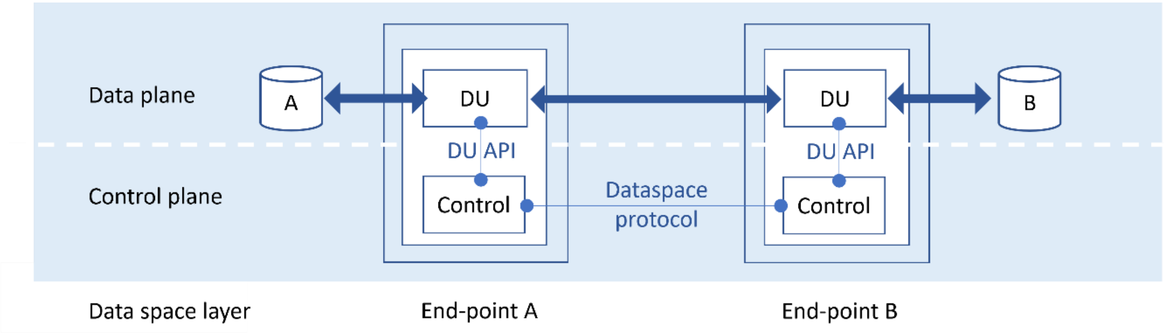


Figure 18 Dataspace protocol including the Data Usage API

The "control" components (IDS Connectors) provide verifiable tokens to the Data Usage components when all terms and conditions of a mutually agreed contract are applicable. A Data Usage component implements a specific protocol (for D2A or A2D) to facilitate predefined and verified operations on local data including collaboration with one or more other DU components of the same type and with the same verifiable identification and verifiable certification.

REQ The functional requirements for the Data Usage API for the "control" of the data plane are:

- Request the current state of the component Data Usage (DU)
- Start a component with a specific configuration on behalf of a specific user
- Suspend or abort an on-going execution
- Request logging information
- Request results

REQ The following states are applicable and compliant with the states of the Transfer Process protocol:

- REQUESTED: requested under an Agreement(contract)
- STARTED: available for access / local processing.
- COMPLETED: execution has been completed.
- SUSPENDED: execution has been suspended by the Consumer or the Provider.
- TERMINATED: execution has been terminated by the Consumer or the Provider.

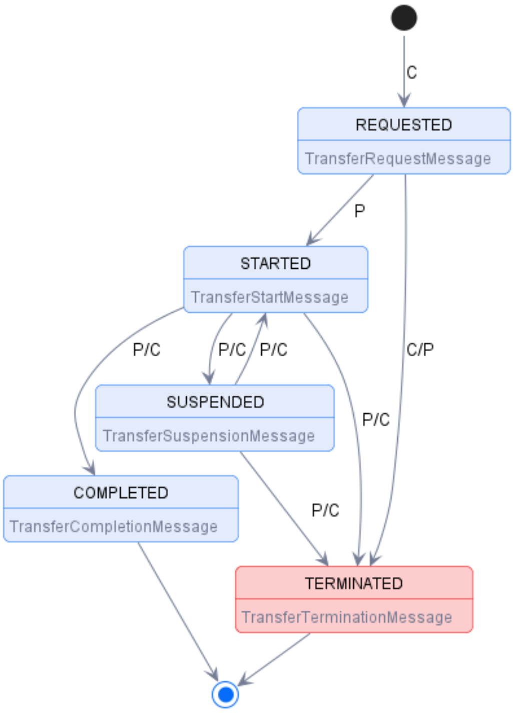


Figure 19 Transfer protocol finite state diagram (source IDS-RAM 4.0)

REQ The following specification defines a RESTful API over HTTPS equivalent to the Transfer Process Protocol as described here[6].

Overview of provider-side end-points:

- GET transfers/:providerPid (=> transfer process)
- POST transfers/request (=> process available)
- POST transfers/:providerPid/start
- POST transfers/:providerPid/completion
- POST transfers/:providerPid/termination
- POST transfers/:providerPid/suspension

With the corresponding callback path-bindings:

- POST :callback/transfers/:consumerPid/start
- POST :callback/ transfers/:consumerPid/completion
- POST :callback/ transfers/:consumerPid/termination
- POST :callback/ transfers/:consumerPid/suspension

REQ All messages must be serialized in JSON-LD compact form as specified in the JSON-LD 1.1 Processing Algorithms and API. Further Dataspace specifications may define additional optional serialization formats.

Catalogue API

As part of the upcoming Dataspace protocol, the Catalogue Protocol defines how a Catalogue is requested from a Catalogue Service by a Consumer using an abstract message exchange format. The concrete message exchange wire format is defined in the binding specifications. The implementation results in a Catalogue API which can be used to request metadata from a FAIR Data Point. It is based on generic standards such as DCAT and not specific for the Healthcare domain.

Note This means that additional semantic standards are required for the correct interpretation of the meta data and data itself. Therefore this Catalogue API needs to be extended with …

Because of its generic function, the Catalogue API is implemented as part of an IDS connector and access is controlled by usage policies to implement a HERACLES channel for metadata.

REQ The functional requirements for the Catalogue API are:

- Request a catalogue to be downloaded from a FAIR Data Point;
- Request a dataset to be downloaded from a FAIR Data Point.

REQ The following specification defines a RESTful API over HTTPS equivalent to the Catalogue functionality as described here[7].

Overview of provider-side end-points:

- POST catalog/request => DCAT catalogue
- GET catalog/datasets/:id => DCAT dataset

Note The American spelling for catalogue is used in the end-points!

[6] https://docs.internationaldataspaces.org/editorial-rework-of-dspace/transfer-process/transfer.process.binding.https

[7] https://docs.internationaldataspaces.org/editorial-rework-of-dspace/catalog/catalog.binding.https

REQ A dataset must have A Dataset must have 1..N hasPolicy attributes that contain an ODRL Offer defining the Usage Policy associated with the Catalogue. Offers must NOT contain any explicit target attributes. The target of an Offer is the associated Dataset. This is in line with the semantics of hasPolicy as defined in the ODRL Information Model, explaining that the subject (here the Dataset) is automatically the target of each Rule.

REQ A Catalogue is a DCAT Catalogue with the following restrictions:

- Each ODRL Offer must be unique to a Dataset since the target of the Offer is derived from its enclosing context.
- Each ODRL Offer linked from a Catalogue must NOT include an explicit target attribute.

REQ All messages must be serialized in JSON-LD compact form as specified in the JSON-LD 1.1 Processing Algorithms and API. Further Dataspace specifications may define additional optional serialization formats.

Detailed technical considerations such as filtering, pagination and compressed can be found as will in the Catalogue protocol specification[6].

4 Perspectives of the HERACLES reference architecture

Directly related to the five layers of the HERACLES Architecture are three cross-sectional perspectives: Security, Certification, and Governance. These are described in detail in the following sections.

4.1 Security perspective

Security aspects

Security delivers the means to establish trust in the HERACLES infrastructure which is the basis for the sovereign data exchange and processing targeted.

REQ In the Business layer the *Governance body* is responsible for setting up and maintaining a trustworthy HERACLES infrastructure. They assess risks across the entire system and for all of the five layers of the HERACLES architecture.

REQ *Vulnerability scanning* and *penetration testing* are topics to be addressed at the functional layer. Vulnerability scanning is the process of identifying and assessing security weaknesses and flaws in systems and software running on them. Penetration testing is a simulated cyber-attack on a computer system or network that evaluates the security posture of the target systems or applications, with the goal of identifying vulnerabilities that could be exploited by an attacker. The governance model should include a procedure including frequency for performing vulnerability scanning and penetration testing.

REQ For the Information layer *data integrity* of data and metadata is an essential aspect to managed by ensuring that (meta)data is accurate, complete, consistent, and reliable, and by regularly reviewing and auditing metadata processes to ensure that data management plans are effectively in place. Verifiable claims and data in general is a solution direction

REQ *Access control* in the process layer involves ensuring that secure business processes are in place, access is restricted to authorized users, and separation of duties is enforced. Secure business processes involve implementing policies and procedures to ensure that business processes are designed and executed in a secure manner, including access control, separation of duties, and secure data handling.

REQ *Access control* and *network security* in the system layer of the HERACLES architecture involve ensuring that secure hardware and virtualized platforms are in place, access is restricted to authorized users, and network security controls are implemented to prevent unauthorized access to the network.

Identity and Trust management

Identification and trust management require reliable information, a security-conscious organizational culture, and established security requirements and concepts.

REQ *Reliable Information*: Reliable information is essential to establish trust and enable participants to make sovereign and informed decisions. Trust management is a scientific process of managing trust by following a systematic approach of removing and protecting a network from untrustworthy elements

REQ *Security-Conscious Organizational Culture*: A security-conscious organizational culture is necessary to ensure that all staff members understand the importance of protecting the system and are committed to security. Security management consists of nurturing a security-conscious organizational culture, developing tangible procedures to support security, and managing the myriad of pieces.

REQ *Security Requirements and Concepts*: Security requirements and concepts for different devices and involved entities in the IDS should be established and implemented. Trust in the security of information exchanged over the Internet and other networks during transactions will play a vital role in the future of the HERACLES infrastructure.

Securing the Platform

Securing the platform requires expertise and compliance, data security, and access control and network security.

REQ *Expertise and Compliance*: It is important that the platform team has expertise in formulating quick responses to new threats and that the solutions being formulated are in compliance with regulations and best practices.

REQ *Data Security*: Comprehensive data security means that the HERACLES Infrastructure can endure or recover from failures, and building resiliency into hardware and software to avoid that events like power outages or natural disasters won't compromise security.

REQ *Access Control and Network Security*: Access control and network security in the system layer of an architecture description involve ensuring that secure hardware and virtualized platforms are in place, access is restricted to authorized users, and network security controls are implemented to prevent unauthorized access to the network. Firewalls control incoming and outgoing traffic on networks, with predetermined security rules, and keep out unfriendly traffic.

Securing Applications

Securing applications as part of the HERACLES Infrastructure requires secure coding practices.

REQ *Secure coding practices* are essential to prevent vulnerabilities in applications. Developers should be trained in secure coding practices and should follow secure coding guidelines.

Also Vulnerability scanning and penetration testing as mentioned earlier are important to identify and remediate vulnerabilities in applications.

Securing Interactions between components

Securing interactions between components requires access control, data integrity, and security controls. Access control and data integrity are already addressed earlier in this section.

REQ *Security controls* are measures that are put in place to protect systems and data from unauthorized access, use, disclosure, disruption, modification, or destruction. There are three main types of security controls including technical, administrative, and physical controls. Combining controls into *multiple layers of security* will prevent that if one layer fails to counteract a threat, other layers will help to prevent a breach in the HERACLES Infrastructure.

Usage Control

Usage control is a security mechanism that regulates how data can be used after access has been granted, and it is an extension to traditional access control.

REQ *Policy contracts* are used to verify by the data provider if the result of a local computation (use of data) *for HERACLES governance compliancy* before the results are released. At first this will be a manual process which can be (partially) automated over time with use of software that verifies the agreed policy contracts.

REQ *Audits* performed by Evaluation authorities are used to verify if the receipients of the data are compliant to data usage agreements as stated in the HERACLES governance model. The evaluation is not expected to be automated within the near future and remains a manual process to be organised by the Governance board.

4.2 Certification perspective

Certification aspects

In the Business layer the *Certification Body* and Evaluation Facilities are in charge of the certification process. Organizations assuming a role under one of the three categories*: Data providers, Researchers* and *Service providers* are potential targets of certification, i.e. may act as Applicant for the Certification.

REQ A *Certification Scheme* as part of the governance model describes for each role what level of certification is required and what the focus of the certification is.

The *functional requirements* of the HERACLES Infrastructure defined in Section 3.2 are the core requirements expected to be implemented by the technical components. Therefore, compatibility of each such implementation with these functional requirements forms the basis of the compliance part of a core component's certification. The security part of the certification focuses on security specific requirements. The security requirements are mainly related to the System Layer and the Security Perspective provided in Section 4.1.

Certification of a core component comprises also its compliance with the HERACLES Architecture regarding functionality, protocols, etc. Whenever relevant, evaluation of a core component's compliance also refers to its *compatibility with the Information Model* defined in the Information Layer (Section 3.3.).

The Process Layer (Section 3.4.) defines the certification process to evaluate submitted/candidate software components and software services.

The System Layer (Section 3.5.) defines the possible interactions between the components, detailed requirements for the Connector, and specific types of Connector implementations. The System Layer is the predominant layer regarding the security requirements with *the Component Certification.*

Roles

The realization of the HERACLES Certification process requires different roles responsible for different tasks:

- Applicants (Data providers, Researchers and Service providers),
- Evaluation Facilities, and
- Certification Body (Evaluation authority)

Note The roles described in this section are specific to the certification process for the HERACLES Infrastructure (i.e. terms such as "Certification Body" should not be misunderstood to refer to an existing organization already granting certificates). The defined roles and their main tasks are described below while their tasks and interactions are described in section 4.2.5.

The *Certification Body* is an Evaluation Authority who oversees the certification process regarding quality assurance and framework governance. It defines standard evaluation procedures and supervises the actions of the Evaluation Facilities. A certificate is granted only if both the Evaluation Facility and the Certification Body have come to the conclusion that all preconditions for certification are fulfilled.

Contracted by an Applicant (see below), the *Evaluation Facility* is responsible for carrying out the detailed technical and/or organizational evaluation work during a certification process. The

Evaluation Facility issues an evaluation report for the respective organization/individual or core component, listing details regarding the evaluation process and an assessment whether all requirements are properly fulfilled.

The term "Evaluation Facility" refers both to authorized auditors for management system evaluations (i.e., for Operational Environment Certification) as well as approved evaluators for software stacks (i.e., for Component Certification). Hence, the Certification Body oversees and cooperates with multiple Evaluation Facilities. However, only one Evaluation Facility is involved in each evaluation of an organization/individual or core component.

The *Applicant* is subject of the evaluation and certification process and needs to actively submit an application to trigger the certification process. This applies to organizations that develop software components intended to be deployed within or to be connected with the HERACLES Infrastructure . This task can be delegated to a Software Provider but the Applicant remains accountable. Also the role of Researcher as a data consumer can be subject of certification to increase the chance of correct use of the HERACLES infrastructure and the proper use of the received data.

REQ The terms and conditions for an Applicant shall be stated in the Governance model.

During the certification process, the Applicant provides all necessary material needed for the evaluation and certification of its component or organization and supports with questions or issues arising.

Operational environment certification

Participants in the HERACLES Infrastructure share valuable data. It is essential that all participant's organizational processes and operational environments are trustworthy. This trustworthiness is evaluated in the HERACLES Operational Environment Certification.

Central elements of the HERACLES Operational Environment Certification are the different *Trust Levels* and Assurance Levels to offer suitable certification profiles for different use case requirements.

On one side, the following three Trust Levels are established:

- Trust Level 1: Entry into data sharing
- Trust Level 2: Providing reliable services
- Trust Level 3: Offering trust-building services

Higher Trust Level represent the increasing amount of criteria which needs to be fulfilled for a successful certification.

On the other side, the following three Assurance Levels are established:

- Assurance Level 1: Self-Assessment
- Assurance Level 2: External evaluation of corporate policies and processes
- Assurance Level 3: External audit of measures controlling the adherence to corporate policies

Higher Assurance Level represent the increasing demand for more reliable evidence that needs to be presented in different evaluation methods to prove compliance with the certification criteria.

REQ An in-depth description of the Operational Environment Certification for each type of Applicant shall be stated in the HERACLES Governance model.

Component certification

Trustful cross-company information exchange requires secure soft- and hardware components.

REQ All HERACLES Infrastructure components have to meet a list of certification criteria to prove the provision of the required functionality, interoperability and level of security. The evaluation of these certification criteria is conducted in the HERACLES Component Certification which is part of the HERACLES Governance model. Similar is for the operational environment certification, three different levels of assurance and trust for the certification of components are defined.

The depth and rigor of a component evaluation consists of the following *three assurance levels*, independent on the type of component that is being certified:

- Assurance Level 1: Checklist self-assessment and automated interoperability testing;
- Assurance Level 2: External concept review including functional and security testing;
- Assurance Level 3: External evaluation including concept review, testing and source code audit.

The criteria that make up each of the three trust levels for a Connector are defined in such a way that they are specific enough to ensure interoperability with the functional requirements of an HERALCES Component, yet general enough, to allow the use of a component in different deployment scenarios without having to define different criteria catalogues for each separate use case. The following *three trust levels* are defined for the certification of a Connector:

- Trust Level 1: Data space interoperability;
- Trust Level 2: Feature complete for data usage control;
- Trust Level 3: Additional protection from internal attacks.

Processes

Participants and core components in the HERACLES Infrastructure shall fulfil common requirements to ensure the security of data being processed in the HERACLES Infrastructure.

REQ Therefore, the certification of operational environments and components is mandatory. Involved partners are the Applicant, Evaluation Facility and the Certification.

4.3 Data governance perspective

The Governance Perspective of the HERACLES Architecture defines the roles, functions, and processes of the HERACLES Infrastructure from a governance and compliance point of view. It defines the requirements to be met by the business ecosystem to achieve secure and reliable corporate interoperability. Two dedicated documents: D3.3 – Governance Assessment[D3.3] and D3.4 – Governance recommendations [D3.4] will describe the data governance perspective for compliant collaboration of stakeholders with use of the HERACLES Infrastructure.

REQ The HERACLES Governance model for the HERACLES Infrastructure distinguishes three levels: the *business, organisational* and *operational level* and describes at least the following aspects:

- Governance aspects for each of the five layers as described in section 3;
- Key roles and correlating data governance and management activities;
- Data ownership which is relevant on every layer of the HERACLES architecture;
- Term and conditions for minimal data quality and data provenance;
- Privacy perspective as mentioned earlier in this section;
- Governance of vocabularies (semantic models required for the data exchange).

5 Conclusions and recommendations

5.1 Conclusions

To summarize, this document describes the semantical and technical aspects of the HERACLES infrastructure as a whole based on certified services connected via a dataspace to enable data sovereignty and using real world sensitive medical data in a secure and privacy preserving way.

The HERACLES Infrastructure takes in account the legal, organizational, semantical and technical interoperability aspects. Permission to use data is specified in a *Data Contract* per research case from all four interoperability aspects. The technical implementation of such a Data contract is being supported with use of European standards for a dataspace and the creation of secure *HERACLES channels* which support the use of Privacy Enhanced technologies such Multi-party Computation and Federated Learning. FAIR Data Points (FDP), Data Permit Management (DPM) and FAIR Data Stations(FDS) are the key functional components for respectively data discovery, data contract negotiation and data usage. The figure below visualizes the three main steps to be taken from the dotted outside hexagon where discovery is available, to the safe and controlled inner space where data can be used.

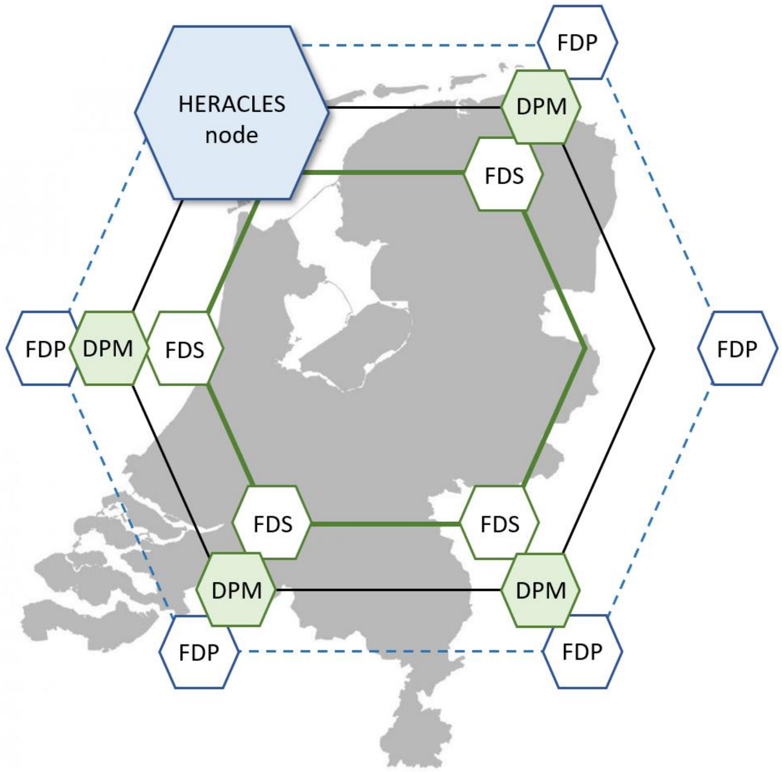


Figure 20 HERACLES infrastructure based on nodes

Based on concrete specification of protocols which includes process and interface descriptions, multiple implementations can be realized based on already existing components, to test the architecture and to demonstrate the interoperability between different implementations and the feasibility of the HERACLES infrastructure.

5.2 Recommendations

The main recommendation is to realize multiple minimal viable products (MVPs) based on the same architecture (as described in this document) but as different implementations to test and demonstrate the interoperability aspects. This avoids a single solution / vendor lock-in and maximize the use of already working components.

Deliverable “Legacy feasibility study to take in account applicable regulations” [D3.1] will provide concrete input for the structure and required content of a Data Contract. A proposal based on that input will be included in version 2 of this document. Generic terms and conditions for a Data Permit / Data Contract are also expected to be derived from deliverable “Governance assessment, to describe the procedures to access the data” [D3.3].

6 References

[1] – PPP-Allowance Agreement LSHM21060 complete 2023

[2] – New European Interoperability Framework 2017
https://inspire.ec.europa.eu/news/new-european-interoperability-framework-eif

[3] – OpenDEI Position Paper: Design Principles for Data Spaces,
https://www.opendei.eu/

[4] – IDS-Reference-Architecture-Model (v4.0)
https://docs.internationaldataspaces.org/ids-knowledgebase/v/ids-ram-4/

[5] – DIN-SPEC-27070
https://internationaldataspaces.org/ids-is-officially-a-standard-din-spec-27070-is-published/

[6] – IDSA Rule book 2023
https://docs.internationaldataspaces.org/ids-knowledgebase/v/idsa-rulebook

[7] – Designing Data Space – The Ecosystem Approach to Competitive Advantage
https://link.springer.com/book/10.1007/978-3-030-93975-5/

[8] – GAIA-X,
https://www.gaia-x.eu/

[9] – eIDAS
https://digital-strategy.ec.europa.eu/en/policies/eidas-regulation

[10] – Self-sovereign Identity (SSI)
https://en.wikipedia.org/wiki/Self-sovereign_identity

[11] – Decentralized Identifiers (DIDs) v1, W3C
https://www.w3.org/TR/did-core/

[12] – Verifiable Credentials Data Model v1.1
https://www.w3.org/TR/vc-data-model/

[13] – Data Catalog Vocabulary (DCAT) - Version 2
https://www.w3.org/TR/vocab-dcat-2/

[14] – FAIR Data Point
https://www.fairdatapoint.org/

[15] – Joint Action Towards the European Health Data Space – TEHDAS
https://tehdas.eu/

[16] – Dataspace Protocol 2024-1
https://docs.internationaldataspaces.org/ids-knowledgebase/v/dataspace-protocol

[D1.1] – Information specification v1.0

[D1.2.1] – Software algorithms and models (version 1)*

[D1.3] – Proof of Concept*

[D2.1] – Technology Radar, v1.0

[D2.2.1] – Architectural design (version1)
(this document)

[D3.1] – Legacy feasibility study*

[D3.3] – Governance assessment*

* These documents were not available at time of writing but mentioned in this document.

Annex A: HERACLES dataspace actors

This section provides background information about the identified HERACLES dataspace actors from a generic point of view including their mutual relations. From a legal point of view it is important to make the distinction between accountable and responsible actors.

Basic type of actors

At a high level, three basic types of actors are distinguished: *natural persons*, *services* and a *group* of actors . An actor can act on behalf of its *self* or perform a (*delegated*) action on behalf of another actor which becomes relevant in case of legal issues.

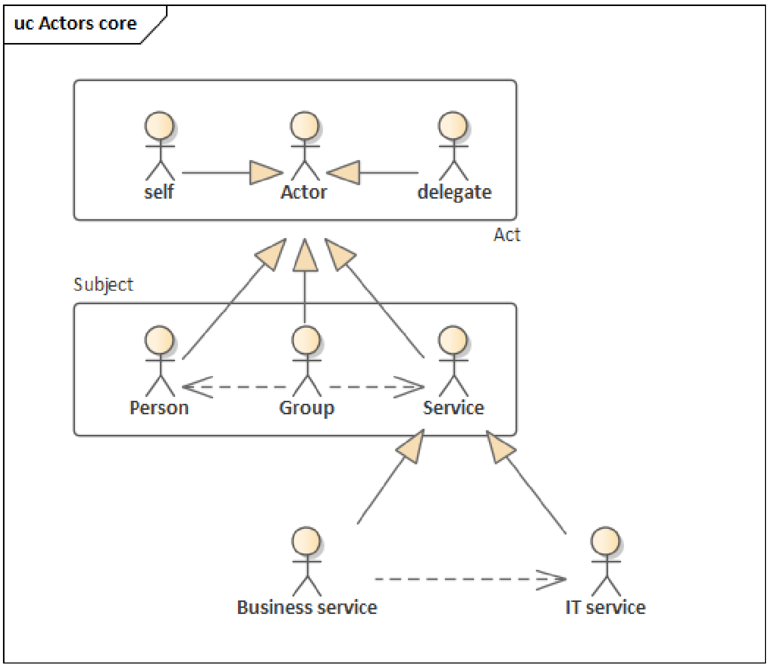


Figure 21 basic type of actors

A *service* actor can be a *business service* with a corresponding business model or an *IT-service*. A (natural) person or organisation are both *legal entities* while services are not. Legal entities can be hold accountable while delegation is by definition not, and can only be hold responsible.

For handling verifiable claims (statements about a subject), the following generic roles are relevant for dataspaces too, because verifiable identities and verifiable claims in general are the cornerstone within a dataspace:

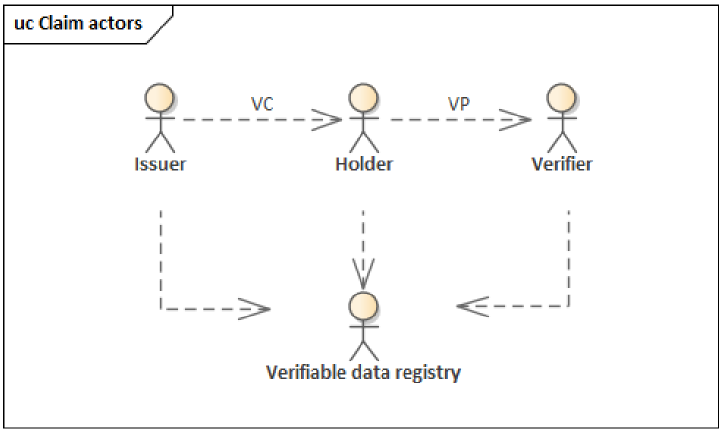


Figure 22 Claim actors

The following actors are defined[8] to enable verifiable identites/claims:

- *Issuer,* asserts claims about one or more subjects, creates a verifiable credential (VC) from these claims, and transmits them to a holder;
- *Holder*, possesses one or more verifiable credentials and generates verifiable presentations (VP) from them;
- *Verifier*, relying party or client receiving one or more verifiable credentials, optionally inside a VP for processing;
- *Verifiable data registry*, mediating service to assist the creation and verification of identifiers, keys, and other relevant data, such as verifiable credential schemas, revocation registries, issuer public keys, and so on, which might be required to use verifiable credentials.

IDS Dataspace actors

The IDS-RAM[4] describes four categories of actors: *core participants, intermediary, software & service providers* and *governance body*.

An exact definition of the core-actors and an elborate description can be found in IDS-RAM[4] and the DIN-SPEC-27070 [5]. This document only highlights the main aspects relevant for the functional use-case.

A *dataspace participant* is defined as a *legal entity* with a unique verifiable identity, authenticated and accepted within a dataspace, who has accepted and is committed to the governance model of that dataspace.

[8] Summary of the definitions stated in Verifiable Credentials Data Model v1.1 [12]

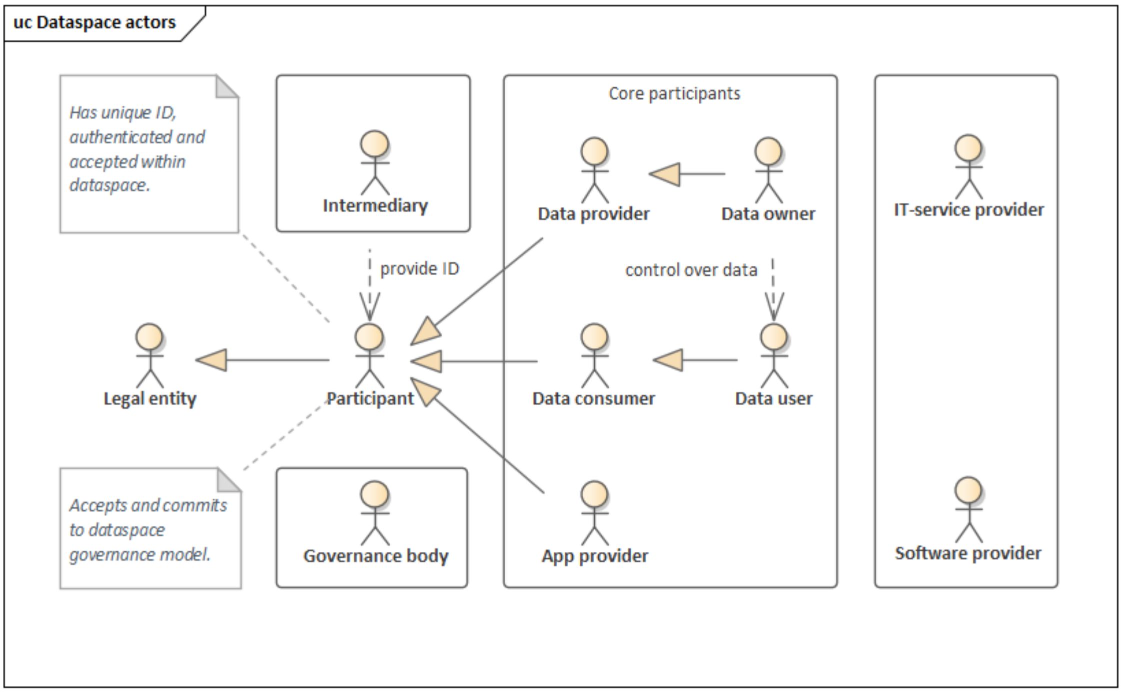


Figure 23 Dataspace actors

Core participants

All core participants are legal entities.

- *Data owner*, creates and has legal control over the data based on data access and usage policies. Can be data provider it self or delegate that role;
- *Data provider*, makes data available (on behalf of data owner) for data consumers and submits metadata for discovery purpose;
- *Data consumer*, receives data from the data provider;
- *Data user*, a data consumer with the legal rights to use data based on data access and data usage policies agreed with the data owner;
- *App provider*, develops DataApps[9] compliant to the architecture of the dataspace), submits DataApps for certification and distributes them via an App store service.

[9] DataApps are software components to provides access to data and data processing capabilities in collaboration with a Security Gateway software component.

Intermediary actors

Intermediary actors are trusted[10] *legal entities* who are establishing trust, providing metadata and supporting dataspace functionalities for the *participants*.

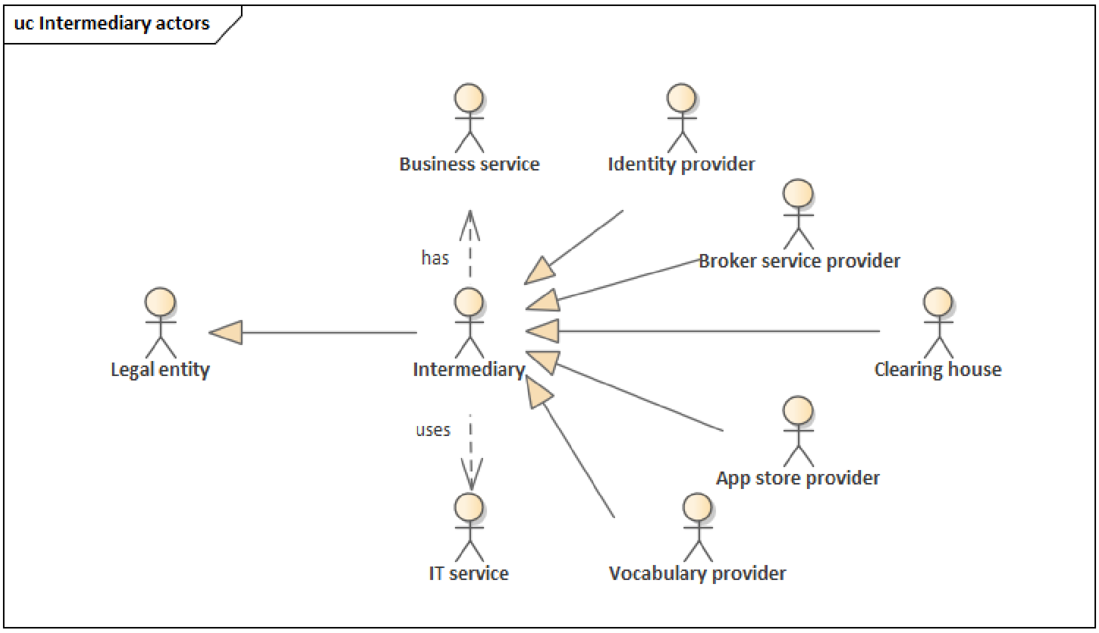


Figure 24 IDS intermediary actors

An intermediary actor has created a business model with corresponsing *business service(s)* to provide one or more *IT services* to participants.

- *Broker service provider*, stores and manages meta-information about available data;
- *Identity provider*, creates and authenticate the identity information of all participants;
- *Clearing house*, provides clearing and settlement services for (all) financial and data exchange transactions;
- *App store provider*, provides trusted Data Apps and facilitates publishing and retrieving of Data Apps including corresponding metadata;
- *Vocabulary provider*, manages and offers ontologies, reference data models, or metadata elements being used with the dataspace.

Software and IT-Service providers

Software and *IT-service providers* are IT companies and therefore also legal entities which perform delegated actions and have a technical supporting role within a dataspace in contrast to *intermediate actors* which have a functional supporting role.

The roles are defined as:

- *Software provider*, provides software for implementing the functionality required for the dataspace. In contrast to Data Apps, the software will not be distributed via the App Store service;

[10] Being a trusted participant is a key aspect of realizing “implicit” trust within a dataspace c.q. among the other participants within the HERACLES dataspace. This concept of trusted participants is crucial when connecting with other Health dataspaces.

- *IT Service provider*[11], provides the deployment of the required infrastructure for participation in a dataspace.

Governance body

The *governance body* is an organisation to facilitate the dataspace with controlling functionality such as certification, evaluation and governance of the dataspace, based on a governance model agreed by the participants of that dataspace.

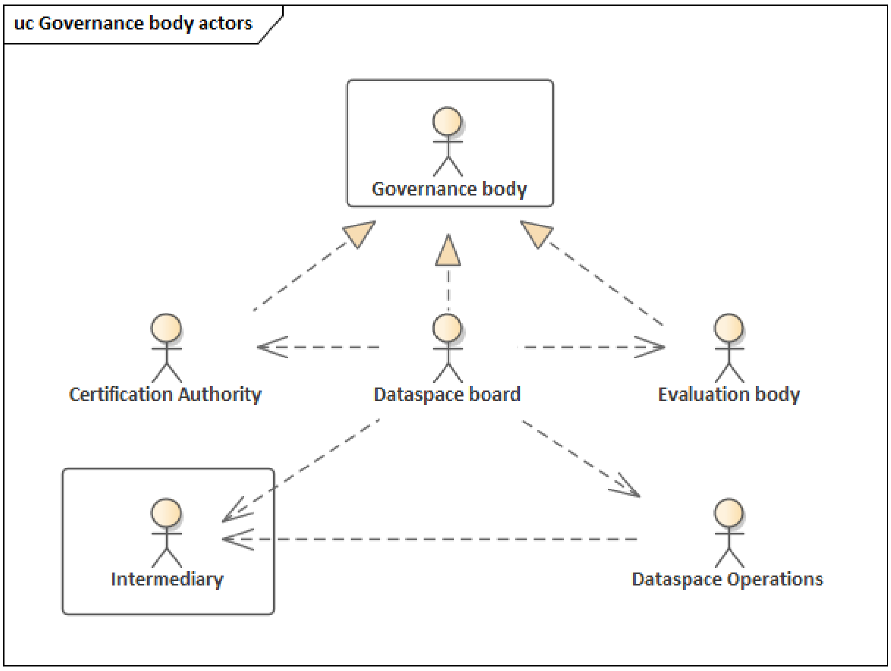


Figure 25 Governance body actors and relations with operations

The following roles are identified:

- *Certification authority* for authentication of participants and certification of services and software;
- *Evaluation body* verifies with audits if dataspace operations are compliant with the governance model and regulations[12];
- *Dataspace board* represents the dataspace on behave of the participants and is accountable for compliance with the governance model;
- *Operations* is responsible for daily operations in collaboration with intermediary services and optionally provides a servicedesk to assist new and existing participants.

The *certification* and *evaluation authorities* are both part of the *governance body* but can be delegated to external specialized organization capable of providing professional, specialized and independent/objective services. An example of an *evaluation authority* is for instance the *GDPR supervisory authority*. (see below)

[11] The name "*service provider*" used by IDS-RAM[4] is made more specific by using the term "IT-service provider" to distinguish it from a "business service provider".

[12] The governance model should be compliant with regulations to begin with and periodic checks on regulatory compliancy are required because the world is in motion.

The General Data Protection Regulation (GDPR) and corresponding actors are mentioned here for completeness reasons and the definitions are not copied into this document because they are standardized and clearly specificied.

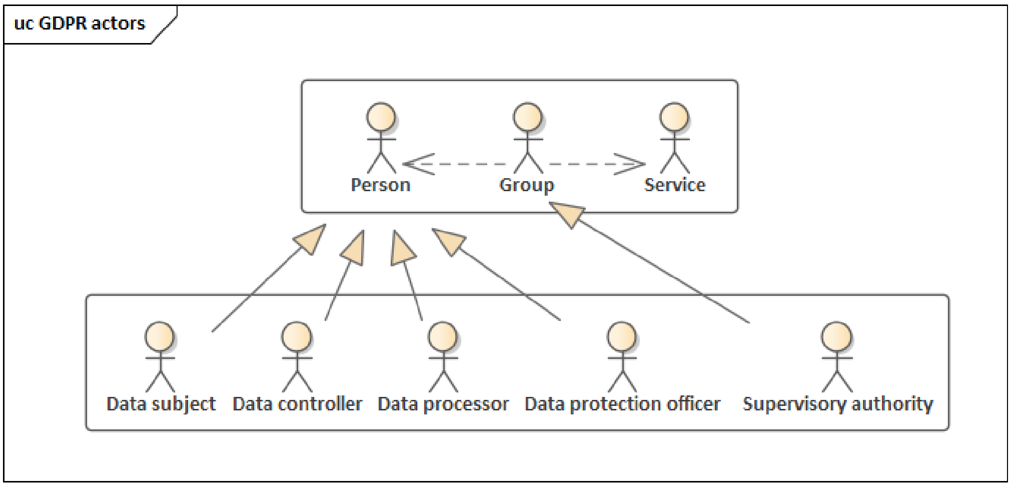


Figure 26 GDPR actors