Towards European Health Data Space

Deliverable 7.1

Options for the minimum set of services for secondary use of health data in the EHDS

5 April 2022

This project has been co-funded by the European Union's 3rd Health Programme (2014-2020) under Grant Agreement no 101035467.

0. Document info

0.1 Authors

Author	Partner
Juan González-García	IACS, Aragón Institute of Health Sciences, Spain
Jaakko Lähteenmäki	VTT, Technical Research Centre of Finland, Finland
Juha Pajula	VTT, Technical Research Centre of Finland, Finland
Helena Lodenius	CSC, IT Centre for Science, Finland
Carlos Tellería-Orriols	IACS, Aragón Institute of Health Sciences, Spain
Enrique Bernal-Delgado	IACS, Aragón Institute of Health Sciences, Spain
Francisco Estupiñán-Romero	IACS, Aragón Institute of Health Sciences, Spain
Ramón Launa-Garcés	IACS, Aragón Institute of Health Sciences, Spain
Lars Simesen	RM, Region Midtjylland, Denmark
Rosie Richards	NHSC, The NHS Confederation, UK
Coen van Gool	RIVM, National Institute for Public Health and the Environment, Netherlands
Michel Silvestri	SEHA, Swedish eHealth Agency, Sweden

0.2 Keywords

Keywords	TEHDAS, Joint Action, Health Data, Health Data Space, Data Space, data permit, secondary use, services

Accepted in Project Steering Group on 29 March 2022.

Contents

Executive summary

The joint action (JA) Towards the European Health Data Space (TEHDAS) is the policy development tool responsible on developing options for the common frameworks to support the cross-border secondary use of health data. The goal is to inform the European Commission (EC) in their legislative proposal for the European Health Data Space for secondary use (EHDS2). The goal of the EHDS2 is that, in the future, European citizens, communities and companies will benefit from secure and seamless access to health data regardless of where it is stored. The TEHDAS JA started in February 2021 and runs until 1 August 2023.

Within the TEHDAS JA, the work package 7 (WP7) "Connecting the dots" will detail the technical options to provide an effective secondary use of health data through the European Health Data Space for secondary use of health data (EHDS2). As defined in the TEHDAS glossary[1], the secondary use of data occurs "when data is used for a purpose different from the purpose for which the data was initially collected." It is important to reinforce that this is definition agreed within the TEHDAS JA context, and may differ to the final EHDS legislative proposal.

This document presents the options for the minimum services, understood as computing systems and software, required for the proper operation of the EHDS2. This deliverable also sets the basis for the initial discussions on services implementation as well as the EHDS2 infrastructure architecture that will be disclosed in the final deliverable of the work package: D7.2 "*Options for the services and services' architecture and infrastructure for secondary use of data in the EHDS*".

[1] https://tehdas.eu/results/tehdas-glossary/

1 Introduction

Within the TEHDAS JA, the work package 7 (WP7) "Connecting the dots" has the objective of detailing the technical options to provide an effective secondary use of health data through the European Health Data Space for secondary use of health data (EHDS2). As collected in the TEHDAS glossary[1], the secondary use of data occurs is defined as "when data is used for a purpose different from the purpose for which the data was initially collected."

According to the European Interoperability Framework (EIF)[2], the solutions to be explored in WP7 represent the technical interoperability elements of the EHDS2. As defined in EIF technical interoperability covers "*[…] the applications and infrastructures linking systems and services. Aspects of technical interoperability include interface specifications, interconnection services, data integration services, data presentation and exchange, and secure communication protocols.*". Organisational and legal interoperability are developed in work packages 4 and 5, while semantic interoperability is addressed in work package 6.

The work on the technical interoperability described in the TEHDAS grant agreement is organised around four specific objectives (O):

- O7.1 Study existing initiatives on secondary use of health data focusing on the requirements for their deployment.
- O7.2 Foster the participation of future users of the EHDS2 and EHDS2 implementers, institutions or industry, to participate in the co-design of the services for secondary use of health data as well to provide architecture and infrastructure options.
- O7.3 Define the options for the EHDS2 services for secondary use of health data.
- O7.4 Detail the architecture and infrastructure options of the EHDS services for secondary use of health data, fully compliant with legal frameworks and with total guarantee of privacy and security.

The present document constitutes the first deliverable of WP7, which addresses O7.3, using as inputs the results described in the milestones that addressed O7.1 and O7.2.

This deliverable also represents a synthesis of the Milestone 7.5 report "*Catalogue of EHDS services for secondary use of health data*"[3], that included a broad catalogue of the possible services, defined as the software elements in a computer system that act as the "logical representation of a set of activities that has specified outcomes, is self-

[2] European Commission, Directorate-General for Informatics, New European interoperability framework: promoting seamless services and data flows for European public administrations, Publications Office, 2017, https://data.europa.eu/doi/10.2799/360327

[3] TEHDAS Milestone M7.5 "*Catalogue of EHDS services for secondary use of health data*" https://tehdas.eu/results/tehdas-proposes-european-health-data-space-services/

contained, may be composed of other services, and is a “black box” to consumers of the service”[4]

This document provides an in-depth description of the options for the essential services required for a proper EHDS2 operation (Section 4), aligned with a re-visited EHDS2 high-level architecture (Section 2.1) and the User Journey for the EHDS2 (Section 2.2), both based on discussions maintained with the WP7 Advisory Group as well as in conversations with other WPs of the joint action.

The options of the services are mostly focused on initial architectural concepts and discussions. This deliverable sets the base of the initial discussions that will be finally disclosed in the final deliverable D7.2 “*Options for the services and services’ architecture and infrastructure for secondary use of data in the EHDS*”, that will address the O7.4 and also an overall view of the whole WP work.

[4] ISO/IEC 18384-1:2016, *Information technology — Reference Architecture for Service Oriented Architecture (SOA RA) — Part 1: Terminology and concepts for SOA*

2 Background

There are two main aspects required to fully understand the discussion on the services presented in this document. The first aspect is the high-level architecture envisaged for the future EHDS2. This high-level architecture contains the relation between computational elements and EHDS2 actors. Second aspect is the "Users' journey", the definition of the process that a data user (see definition in Section 2.2) must follow to access and use the data available in the EHDS2.

These two aspects are under constant discussion, review and improvement as part of the WP7 activities as well as the cross-cutting WP activities. For this reason, they should be taken with caution. Final versions of the high-level architecture and the Users' Journey will be presented in the last deliverable of this WP, D7.2 "*Options for the services and services' architecture and infrastructure for secondary use of data in the EHDS*".

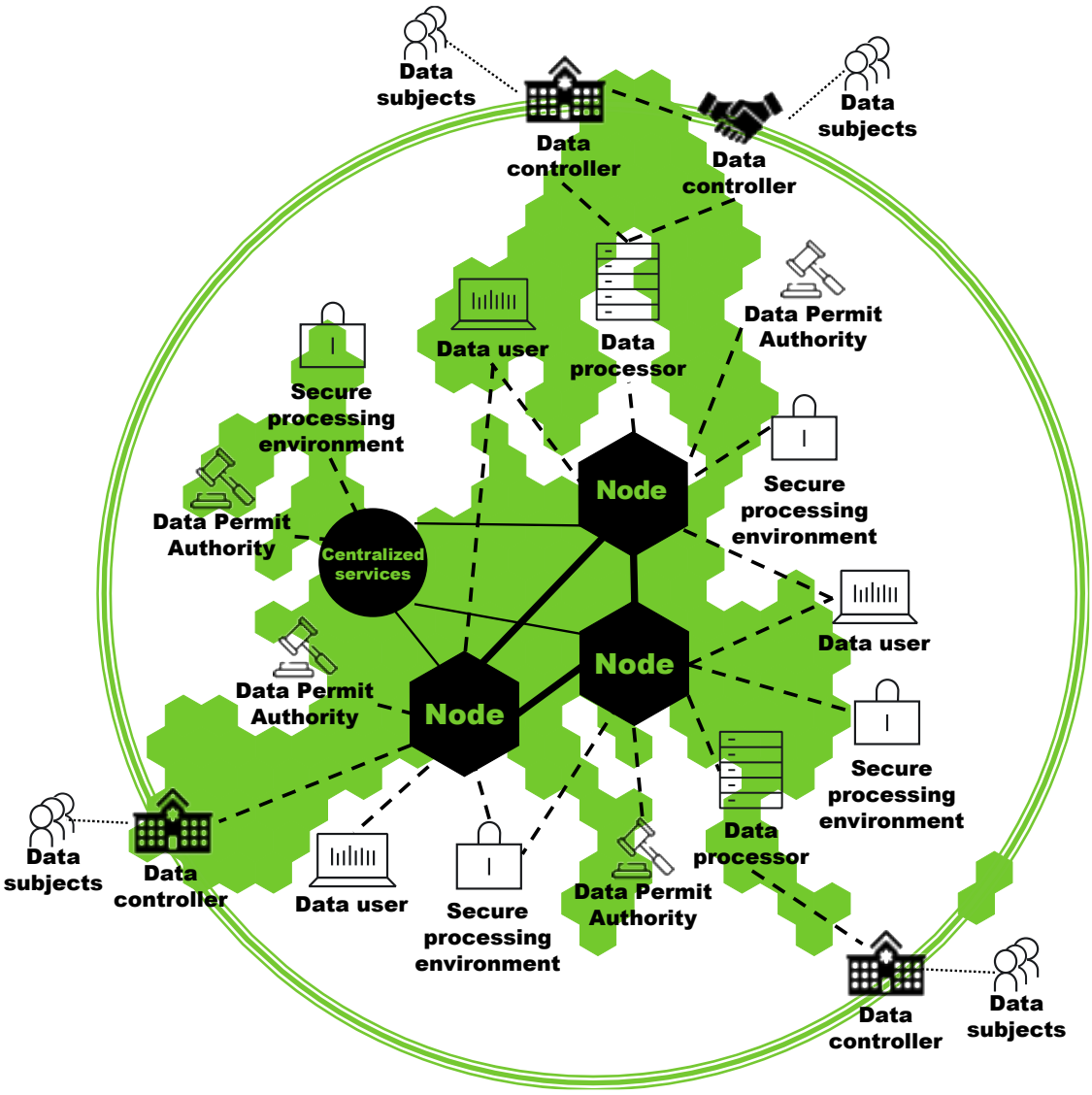

Figure 1 EHDS2 high-level architecture

2.1 High-level architecture revisited

As a conclusion of the discussions of the TEHDAS JA proposal and the further conversations and meetings taken place in multiple stakeholder's forums, the selected architecture for the future EHDS2 is based in a federated peer-to-peer (P2P) network[5].

A P2P network is an architecture of a computer network where the information, i.e., the health data in the EHDS2 context, is distributed among the member nodes, i.e., each node stores its part of the data, as opposed to a client-server scenario where all information is stored in a single node, the *server*, and the rest of nodes, the *clients*, access to this node to retrieve such information. In the specific case of the EHDS2, we defined a *federated* P2P because each node can operate isolated, providing a certain number of services to their users' community, e.g., access and analyse the data available within the node.

Figure 1 contains an overarching schema of this high-level P2P architecture. It has noticeable evolved from the one introduced in Milestone 7.5 report[3]. It clarifies the actors and roles, with an aim to ease its mapping with the legal framework and further governance model. Note that, in the Figure, the extent of the EHDS2 is the area framed with the green circle.

2.1.1 Elements of the architecture

The different elements present in the architecture can be described as follows:

- **Nodes,** are the main elements in the infrastructure, they are interconnected to other nodes in the P2P network system. Nodes serve as interfaces to the data users of the EHDS2, encapsulating all the required services to ensure a proper secondary use of health data across Europe, and orchestrating the processes between data users and the rest of the architectural elements. Note that, as it is a federated architecture, nodes may serve their data users independently.
- **Centralised services**, is a special type of node that interact with regular nodes to help in providing their services to data users across Europe. In principle, centralised services are not expected to interact with data users. These services may be for example metadata catalogues and other types of search engines for metadata.
- **Data users**, are those individuals or institutions that will use the data accessible through the EHDS2 in their day-to-day work, e.g., researchers, innovation managers, policy makers or regulators.
- **Data controllers**, as defined in the GDPR Art4(7)[6], are "*the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data*". In the

[5] M. Parameswaran, A. Susarla and A. B. Whinston, "P2P networking: an information sharing alternative," in *Computer*, vol. 34, no. 7, pp. 31-38, July 2001, doi: 10.1109/2.933501.

[6] *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

context of the EHDS2, the data controllers are the actual data custodians, where data from data subjects is collected and stored for its primary use, for example hospitals or health services (hospital icon). Additionally, data controllers may be other institutions where data subjects have ported their personal data in order to guarantee a finer control on its use (handshake icon). In the Figure 1, data controllers are located on the border of the EHDS2 as the personal data they store for its primary use is then "transformed" to its secondary use within the EHDS2.

- **Data processors**, as defined in the GDPR Art4(8), are "*Natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller*". Note that nodes will also classify as data controllers in the GDPR context. Data processors' functions may include:
    - Controlling the quality of the data to be shared in the EHDS2.
    - Providing technical infrastructures or entry points to single or multiple data sets.
    - Co-operate across Europe to maintain data catalogues and metadata of services and APIs.
- **Secure Processing Environments**, are a special case of data processors that facilitate secure processing of data.
- **Data permit authorities**, are those designated bodies that have the mandate to grant or reject the access to specific health data requested by data users. The functions of the data permit authorities are not developed in the GDPR.

Data subjects are also depicted in the schema as their contacts to the health services represent the actual source of data, in its primary form. Data subjects interact with data controllers to exercise their rights covered in Chapter III of the GDPR[6], especially those related to the secondary use, e.g., Art18 "*Right to restriction of processing*" or Art21 "*Right to object*".

2.1.2 On the legal framework of the architecture

The descriptions and considerations detailed in the previous section are related to GDPR. The exception of the data permit authorities is due to non-existence of such role within the GDPR, but its inclusion clarifies the organisation the architecture and the services associated. This role has been developed in Finnish and French health data sharing initiatives as described in the report "*Assessment of the EU Member States' rules on health data in the light of GDPR* "[7]. For example, Findata, the national data sharing facility also acts as a data permit authority, while in France, the data permit authority is the *Commission nationale de l'informatique et des libertés*.

[7] Hansen J, Wilson P, Verhoeven E, Kroneman M, Kirwan M, Verheij R, van Veen EB. *Assessment of the EU Member States' rules on health data in the light of GDPR*. Specific Contract No SC 2019 70 02 in the context of the Single Framework Contract Chafea/2018/Health/03.

Further mapping with other legal frameworks within the European data strategy[8], for example with the Data Governance Act[9] proposal, Data Act proposal[10], or the AI act proposal[11], is under review as a cross-cutting activity with work package 5, devoted to the legal and organisational interoperability, as well as with rest of technical work packages.

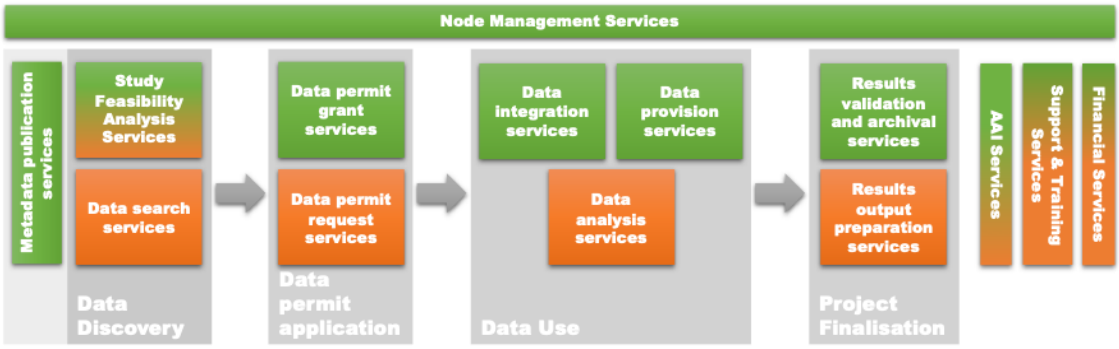

Figure 2 Revisited Users' Journey (AAI: Authentication, Authorisation & Identification)

2.2 EHDS2 Users' Journey Revisited

The Users' Journey is the high-level description of the typical steps that a data user will follow to carry on a data analysis project within the EHDS2 infrastructure. It is used to guide the work of TEHDAS WP7 in defining the EHDS2 technical infrastructure in terms of service options and architecture to be delivered as WP results.

The revisited User Journey, depicted in Figure 2, is richer in terms of separation of concerns than the one presented in the Milestone 7.5[3]. In other words, it clearly separates the specific services that compose each Users' Journey phase from the infrastructure point-of-view and the data users' point-of-view. The separation of concerns facilitates the understanding of the phases and eases the further location options, detailed in Section 4. The revision of the Users' Journey also makes explicit some of the services that were not depicted in previous version in Milestone 7.5.

In the schema, green boxes represent those services that are related to the EHDS2 point-of-view, i.e., the services that are conceived to involved data controllers, data permit authorities and other actors than data users. The orange boxes represent those

[8] Communication from The Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions *A European strategy for data* COM/2020/66 final

[9] Proposal for a Regulation of the European Parliament and of the Council of European data goveranance (Data Governance Act) COM/2020/767 final

[10] Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act) COM/2022/0047 final

[11] Proposal for a Regulation of the European Parliament and of the Council laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending Certain Union Legislative Acts COM/2021/206 final

services purely related to the EHDS2 data users' point-of-view, i.e., where data users interact with the EHDS2. The grey boxes represent the actual phases of the User Journey itself. A brief description of the phases and services is the following:

1. **Data discovery phase**: the data discovery phase is the phase where the data user looks for the data, he or she needs to perform their work (answer a research question and/or take decisions regarding new or existing policies or regulations). Once the search is performed, he or she decides on the feasibility of carrying on their study according to the data found, possibly with the advice of data experts from the nodes. Please note that in the Figure there is an attached block regarding the metadata publication services, this is due to the fact that the metadata publication services, as described in Section 4.1.1, are not essentially part of the User Journey, but a pre-requisite to it: metadata should be *published* so as to be discovered but as independent process to the Users' Journey.
2. **Data permit application phase**: the data permit application phase is the phase where the data user asks for permit to access the data, he or she has found of utility for its purposes to those competent bodies in the EHDS2.
3. **The data use phase**: the data use phase is the phase where the data user finally performs the data analyses, he or she needs to perform the work, thus answering the research questions or finding the evidence to support new or existing policies or regulations.
4. **The project finalisation phase**: the project finalisation phase is the phase where the data requester needs to ensure a proper disclosure of its findings back to EHDS2 infrastructure, following the FAIR principles[12] for the results. It may imply a notification of the incidental findings to the data controllers.

3 Overall considerations on services options

There are two considerations that should be discussed, or, at least mentioned so as to understand the contents of this deliverable. First, the front-end and back-end locations, i.e., how the services that interact with data users and services that operate from the system point of view are placed and how they interact, and their legal implications. Second, how the data mobilisation, i.e., the possibilities of moving data across nodes even crossing borders, interfere with the actual analysis locations, the secure processing environments (SPEs).

These two considerations are related to the architecture and the legal frameworks. A deeper analysis on them will be deeply detailed in last deliverable of this work package as well as in WP5 deliverables.

[12] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

3.1 Front-end and back-end location

A first consideration to be made is how the front-end and the back-end of the system are located and their interaction. By front-end we define those elements in a computer system that act as interface (e.g., a web page to search data), where information is presented to users and provides the interaction point to navigate through the features such computer system offers. On the other hand, as back-end we define those elements in a computer system in charge of handling the business logic that implement the services of such a system (e.g., the algorithms to perform the search within the system).

For the sake of clarity, we consider that the "elements" of the front-end and back-end are the "services" described across the deliverable. In a simplification of the actual technological substrate, the front-end services correspond to those related to the user's point of view (in orange in Figure 2), while the back-end services correspond to those services related to the system point of view (in green in Figure 2).

In terms of the architecture, front-end and back-end share a common property related to its location: centralisation or distribution. This location alternative will influence the options for the services, and in general terms, have an important impact on the options for the services. We introduce here a short discussion on the location of the front-end and back-end, which is important to understand the implications suggested.

3.1.1 Centralised and distributed front-end axis

The location of front-end will define how the EHDS2 data users will interact with the system, and, in general, their perception on how the EHDS2 works.

A centralised front-end will aggregate in a single point of access the services described on each phase. A centralised front-end means that there will be a single-entry point for the EHDS2, for example, a single web application that represents the overall EHDS2. In this case, the perception of data users will be that they are working within a single system that provides cross-border data access. This option eases the interaction but may hide the effective contribution of the different member states or other infrastructures in the operation of the EHDS2.

A distributed front-end will suppose the existence of multiple entry points to the EHDS2, for example, different web applications per participating member state, from where data users will perform the interaction. In this case, the perception of data users will be that each member state (or any other infrastructure connected to the EHDS2) has an explicit participation in the overall EHDS2.

Regarding the front-end location, it is possible to present also a hybrid approach. Some member states or institutions provide their entry points to the EHDS2, i.e., their own front-end, while some others rely on a central front-end, operated by a pan-European agency. That might be the case when no resources are available to build such front-end or in the interim while setting up their own infrastructures.

3.1.2 Centralised and distributed back-end axis

Most of the services listed can be placed in a centralised or a distributed location. Following the schema depicted in Figure 1, centralised location means that the services

will be deployed in the “Centralized Service”, while a distributed location means that the services are being deployed by each “regular” node.

The centralisation of the services eases its further implementation because there is no need to orchestrate or synchronise the participation of the participating nodes to provide such a service, thus simplifying the processes. Distributed systems are much more complex to design and operate but, in some cases, may result the only option to process large amounts of data, or when the data mobilisation across borders is not possible.

On the other hand, if we consider a basic simplification of the architecture where the “regular” nodes of the EHDS2 federation are located in a member state and “Centralised Services” node is expected to be run by a European-level body or agency, it is important to note that the centralisation of services goes at the expense of cession of responsibilities from national level to a supra-national level. This cession constitutes a major governance element in the EHDS2 federation and should be clearly defined for a seamless organisational and legal interoperability, addressed in WP5, and then supported by the necessary technical interoperability.

Note that there is also the possibility of a hybrid location of the back-end services, where part of the service is centralised and part of the service is distributed. For example, it is possible to have a central metadata catalogue containing high-level description of the data available in the different nodes, while each node contains a fine grain description of its data sets on its metadata catalogue. In this example, the central catalogue may serve to help in the search, quickly pointing the data user to search those nodes that are known to contain the required data.

3.1.3 Front-end and back-end location interdependences

It is clear that the location of both sides of such system (front-end and backend) impose some challenges to its further implementation and legal support. In Section 4, where we define the options for the minimum set of services for the secondary use of health data in the EHDS context, we define the possible implications of these interdependencies.

3.2 On the data use: data mobilisation and Secure Processing Environments

There is a general agreement among the different communities (WPAG, policy makers, or health authorities) contacted while developing the present deliverable: downloading data to data user’s premises as a method to access to health data is something to be avoided, only relegated to non-personal data. Current data sharing facilities operating in this manner should move to a much more secure and privacy-preserving approach of providing Secure Processing Environments (SPEs). In this context, the actual placement and ways of use of the SPEs is an architectural related discussion that will be covered in the second and final deliverable of the current work package (D7.2), devoted to define architecture and infrastructure options for the EHDS2.

The SPEs placement is tightly related to the data mobilisation capabilities, i.e., the possibility to move the data from the actual data controllers or processors to SPEs at different levels of closeness to it: same premises of the data holder, same jurisdiction of the data holder (“*same country*”), other jurisdiction (“*another country*”) or in central

services node. The data mobilisation is something that relies in a clearly defined organisational interoperability and, fundamentally, legal interoperability compliance. Once, these two levels of interoperability are defined, then SPEs placement, as a technical interoperability element, will accommodate to support the final data mobilisation options.

Last but not least, legal and architectural decisions interlink with the options for the different services presented here. For example, if data mobilisation is restricted to SPEs located in the same country as where the data controllers are located, the options to integrate the data will be limited to distributed approaches, whereas if data mobilisation is allowed between SPEs located in different countries than the data controllers, the data integration services may be distributed, or centralised in a single SPE.

4 Description of the services per phase

4.1 Data discovery

The data discovery phase is the initial point of the overall Users' Journey process, and thus, the initial contact point between a data user and the EHDS2. In this phase the data user searches for the data he or she needs to carry out a specific project of their day-to-day work (answer a research question and/or take decisions regarding new or existing policies or regulations). To perform this search, it is a requirement that the data available among the different data controllers or processors has been properly processed and published in the EHDS2 through searchable metadata catalogues. Once the search has given the results, the data user should decide on the feasibility of carrying out their study according to the data found.

4.1.1 Metadata publication services

The publication services gather those software elements used by the data processors and the nodes so as to curate the metadata descriptors of the original data, its organisation in a metadata catalogue and the placement of such catalogue in a certain location that will be later accessible through the search services. The metadata publications services have a strong interdependence with the Data Quality Assessment Framework (DQAF) to be defined in WP6, that will include the semantic interoperability options.

It is important to reinforce that, as previously mentioned in the document, the metadata publication services are not part of the data discovery phase in purity, but a pre-requisite of the data discovery phase. Without the publication of the metadata descriptors by data processors or data controllers, the discoverability of the existing data will be nearly impossible, as only data users with pre-existing knowledge of the existing data will be able to find it. This situation will prevent the start of the Users' Journey, leading to an unusable system. Finally, the publication services location has a fundamental interdependence with the data search services, further analysed in Section 0.

Actors involved

Nodes/central services node, data controllers, data processors

Location options

- **Centralised:** all the metadata required to support is stored in a single metadata catalogue.
- **Distributed:** the metadata describing the actual data from a given data controller or processor is stored in each node they are connected to.
- **Hybrid:** a high-level metadata descriptor, e.g., number of records in a collection and the pathologies of such collection, is stored in a central services node catalogue. Further detailed metadata, e.g., age of patients and comorbidities, is stored in the nodes that connect the data controllers or processors storing such data.

4.1.2 Study Feasibility Analysis Services

The study feasibility analysis services are a subset of the "Support and Training Services" (see Section 4.5.3) with a specific purpose of assisting or guiding the data users on evaluating the feasibility of their projects with the data available in the different nodes. These services may include a ticketing system as well as a communication platform to assist the communication between data users and data experts from the nodes.

Actors involved

Data users, nodes/central services node, data controllers

Location options

- **Centralised:** all the required elements to evaluate the feasibility of studies or projects are located in the central services node
- **Distributed:** the different nodes provided to evaluate the feasibility of studies or projects tailored to such node requirements or capabilities.
- **Hybrid:** central services node holds part of the feasibility evaluation, e.g., general view on the data available, while nodes provide the deeper specificities on the data available through them.

4.1.3 Data search services

The data search services include those software elements provided by the nodes or the centralised services node to let the data users find the required data, including a vocabulary and verbs to articulate the searches, i.e., a query language, and its effective communication to the metadata catalogues. Data search may include results of previous projects that used data available in the EHDS2 (see Section 4.4.1).

Actors involved

Data users, nodes/central services node

Location options

- **Centralised:** a single front-end search point provides access to search in all the available metadata repositories or catalogues.
- **Distributed option:** each node provides a search engine front-end to the available metadata on its catalogue. The user needs to enquire every node manually or the node may forward searches to the rest of nodes.

4.1.4 Interactions between metadata publication and data search services

		Metadata publication services		
		Centralised	**Distributed**	**Hybrid**
Data search services	**Centralised**	Search services only need to enquire a central metadata repository	Search services contact to the different metadata catalogues to provide a single answer.	Search services need to operate in two steps. A first step contacting the central catalogue then a second step to refine the search contacting each distributed catalogue
	Distributed	Each search service inquires the central metadata repository	Each search service contacts its own metadata catalogue. Contact to the rest of catalogues may be automatic or manually driven	Same operation as centralised search, but second step may be automatic or initiated by data user

4.2 Data permit application

The data permit application phase starts when the data user has positively evaluated the possibility of performing the analyses he or she requires for their purposes, with the data found in the EHDS2. In the data permit application phase, the data user requests permit for access/using the data. To do so, data users should provide all the necessary information, e.g., project protocol or data management plan, to the data permit authorities designated, e.g., ethical or scientific committees. With this information and the specific restrictions on the data usage, e.g., dissemination level or consents, the permit authorities will grant or reject the request.

4.2.1 Data permit grant services

The data permit grant services gather those software elements deployed by the nodes so as to manage the process of giving access to the data that the data users request. It implies the software elements to provide the designated data permit authorities the means for making their judgements including the information provided by the data user, e.g., a place to access such documents, as well as to access the requested data access restrictions in a formal way, e.g., metadata on data access limitations.

Actors involved

nodes/central Services node, data permit authorities

Location options

- **Centralised:** the data permit grant management software is operated in the central services node, that contacts to "central" data permit authority, i.e., the data permit authority attached the central services node in Figure 1.
- **Distributed:** each node operates the data permit grant management software interacting with its "own" data permit authority, i.e., the one attached to each node in Figure 1.
- **Hybrid:** the data permit grant management software is operated in the central services node, that contacts the permit authorities distributed among the different nodes of the EHDS2.

4.2.2 Data permit request services

The data permit request services are the counterpart of the data permit grant services from the data user point of view. The services include the software elements to start data access petition to the EHDS2, including the collection and submission of all the necessary information, e.g., project protocol or data management plant, a data user must supply to the permit authorities to obtain the necessary permissions.

Actors involved

Data users, nodes/central services node

Location options

- **Centralised:** the data access request software is operated in the central services node, providing a single-entry point to gather the documentation included in a data user request.
- **Distributed:** each node provides an entry point for the data user requests to contact the permit authorities attached to such node. The user needs to request access manually in every node or the node may forward searches to the rest of nodes.

4.2.3 Interactions between data access negotiation services

		Data access grant services		
		Centralised	**Distributed**	**Hybrid**
Data access requests services	**Centralised**	All the data permit application phase is placed in a central services node. Data permit requests are gathered and forwarded to permit authorities within this node	Data access requests are gathered in the Central Services node and forwarded to each Node that then forwards it to the local permit authorities.	Same as centralised "Data permit request services" + "Distributed data permit grant services"

	Distributed	Data permit requests are gathered on each node and forwarded to the central services node, that makes them available to a central permit authority.	Data permit requests are gathered on each node and forwarded to each node whose data required in the request, automatically o manually	Data permit requests are gathered on each node then forwarded to the central services node that forwards to the nodes whose data is required in the request.

4.3 Data use

The data use phase starts when the data access/use has been granted to the data user. In this phase, the data user finally performs the data analyses he or she needs to as part of their work. The data use phase finishes when the data user has finished its research project or have found the evidence to support new or existing policies or regulations. The finalisation of the data analysis phase may be also subject on contractual arrangements, for example, limiting the amount of time a data user has access to the data.

4.3.1 Data integration services

Data integration includes all software elements to provide a harmonised view of the data, including: the retrieval of the requested data and its linkage among data sources; harmonisation, i.e., use of common data models and/or codifications for those variables with well-known semantics, e.g., use of ICD-10 [13]to describe diseases and related health problems; and, pseudonymise or anonymise the data (when necessary), to guarantee the security and privacy. The data integration services have a strong interdependence with the Data Quality Assessment Framework (DQAF) to be defined in work package 6, where the semantic interoperability is being defined.

Actors involved

Nodes/central services node, data controllers, data processors, SPEs

Location options

- **Centralised:** all data from all data controllers or processors resulting from the initial search is gathered to a central services node where all the integration is performed, following the DQAF.
- **Distributed:** each node integrates the data from the data controllers attached to it. Data sets linkage between nodes is not available, e.g., it is not possible to track the data from a single individual across borders. Nodes apply locally the DQAF to guarantee the harmonisation.

[13] World Health Organization(WHO). 1993. *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. Genève, Switzerland: World Health Organization.

- Hybrid: each node integrates data from the data controllers attached to it. central services node aids to link those data sets that require cross-border integration and harmonisation.

4.3.2 Data provision services

The data provision services include all software elements to make the requested data available in a place where the data users can further perform their analyses. As previously introduced in Section 3.2, data provision depends on the data mobilisation capabilities, as the data provision services mainly include the mobilisation ("copying") of the integrated data to a SPE or multiple SPEs. In the specific cases of using aggregated data or anonymised data, a download option may be available.

Actors involved

Nodes/central services node, data processors, SPEs

Location options

- **Download:** the requested data is made available to download to data users' premises.
- **Centralised:** the requested data is mobilised from each data processor to a single SPE to be placed in the central services node or in the premises of a single nodes that will provide access to the requested data.
- **Distributed:** each data processor mobilises the requested data to its trusted SPE within the premises of the node such processor is attached to.

4.3.3 Data analysis services

The data analytics services include all software elements required to perform the actual analysis using the requested data. This includes the analysis environment (web application, remote desktop access or API access) as well as the analysis tools (analysis software, statistics and machine learning libraries for programming languages and, if necessary, distributed runtimes for distributed analysis). Note that the data analytics services do not apply when the data is downloaded.

Actors involved

Data users, nodes/central services node, SPEs

Location options

- **Centralised:** a single front-end for the data analysis services located in the central services node provides access to the SPE or SPEs where data has been mobilised.
- **Distributed:** each node or SPE where the data has been mobilised has its front-end for the data analysis services.

4.3.4 Interactions between data use services

		Data integration services		
		Centralised	**Distributed**	**Hybrid**
Data access provision services	**Centralised**	Requested data is placed and integrated in a central services node SPE or the designated SPE	Requested data is integrated in the different data processor and nodes and then mobilised to a single SPE. In the mobilisation to the designated SPE a round of cross-border linkage may be possible.	Requested data is integrated in different data processors with the help a central services node, then mobilised to a single SPE.
	Distributed	Requested data is mobilised to central services node, integrated and then redistributed to the SPEs on the different node premises.	Requested data is integrated in the different nodes involved then placed in the SPEs close to such nodes. Cross-border linkage is not available.	Requested data is integrated in the different nodes with the aid of central services node to help in the linkage. Then it is placed in the SPEs close such nodes.
	Download	Requested data is integrated in a central services node that provides a unique point to download.	Requested data is integrated on each node. Multiple download points are provided to the user. Linkage not possible (anonymous data) or not necessary (aggregated data).	Requested data is integrated on each node with the aid of the central services node. Multiple download points are provided to the user. Linkage, if required, may be done at pre-anonymisation stage.

		Data provision services		
		Download	**Centralised**	**Distributed**
Data use services	**Centralised**	N/A	The designated SPE where data has been mobilised provides the analysis environment and the associated analysis tools	One of the SPEs designated or the central services node provides an analysis environment able to orchestrate the analysis between SPEs
	Distributed	N/A	N/A	Each designated SPE provides an analysis environment and the associated analysis tools. The orchestration between SPEs relies on the data user ability.

4.4 Project Finalisation

The project finalisation phase is the last phase in the Users' Journey. It starts when the research question is answered, or the evidence required to support a legislative proposal or regulation has been found. In this phase the data user needs to ensure a proper disclosure of its findings to the rest of the EHDS2 users, following the FAIR principles for results data. The findings should also be notified to data controllers to finally inform data subjects.

4.4.1 Results validation and archival services

The results archival and validation services include all the software related to enable the effective storage and cataloguing of the projects results, including the related metadata and other for its potential re-use in further projects, including the safeguards to validate the dissemination levels authorised by the involved parties (data controllers of the original data, data permit authorities). Data and metadata, and any other supplemental material (analysis scripts, manuals, others), should be included to guarantee the reproducibility of the analyses by other data users, following the FAIR principles. Results cataloguing is expected to facilitate the further re-use in connection to the data search services of the data discovery phase of the Users' Journey.

Actors involved

Nodes/central services node, data processors, data controllers, data permit authorities

Location options

- **Centralised:** the results data and metadata are stored and catalogued in a centralized services node for further re-use.
- **Distributed:** each node provide storage for data and metadata for those projects initiated by data users of such node.
- **Hybrid:** the results data and supplemental material are stored in the nodes where the project was started. A central metadata catalogue is maintained at the central services node pointing to the nodes containing the results.

4.4.2 Results output preparation services

The results outputs services are the data user's counterpart in the project finalisation phase. These services include the applications required to facilitate the organisation of the results data and generate the metadata required for its further validation, cataloguing and archival.

Actors involved

Data users, nodes/central services node

Location options

- **Centralised:** there is a single application to prepare and submit the project results, located in the central services node.
- **Distributed:** each node provides an application to prepare and submit the project results storage for data and metadata for those projects initiated by data users of such node.

4.4.3 Interactions between project finalisation services

		Results validation and archival services		
		Centralised	**Distributed**	**Hybrid**
Results output services	**Centralised**	The data user gets access to a fixed application that catalogues the metadata and stores results to a central services node	The data user gets access to a fixed application that distributes the metadata and data results to the nodes involved in the project	The data user gets access to a fixed application that stores the metadata in a centralised point and the results data to the nodes involved in the project

	Distributed	The data user gets access to results application in the node he or she used that transmits then metadata and results to central services node	The data user gets access to results application in the node he or she used that catalogues the metadata and stores the results data locally	The data user gets access to results application in the node he or she used that stores the results locally and transmits the metadata to a central catalogue

4.5 Transversal services

The transversal services are those services not directly associated to a single phase in the Users' Journey but are required for the proper operation of the EHDS2, ranging from the management of the nodes themselves, the authentication, authorisation and identification of the users, the support and training services and the financial services.

4.5.1 Node Management Services

The node management services include all the software elements to ensure that operations of a node are performing according to the expected EHDS2 behaviour. These services can be seen as a set of auditing elements to guarantee the security of Nodes that participate in the EHDS2, and, thus, the privacy of the Data subjects whose data is being used for secondary purpose. At the same time, these services aim to build trust among the data users and between nodes. Node management services may also include the auditing of data quality, according to the Data Quality Assurance Framework.

Actors involved: nodes/central services node

Location options:

- **Centralised:** central services node operates the node management services, and periodically audits the participating nodes then storing audit results.
- **Distributed:** each node is able to execute the node management services to audit its operations, and then forwards the results to the rest of the EHDS2 nodes.
- **Hybrid:** each node is able to execute the node management services to audit its operations. Audit results are then retransmitted to the central services node.

4.5.2 Authentication, Authorisation and Identification (AAI) services

The authentication, authorisation and identification services are a role-based authorisation service to support the EHDS2 operation and security, guaranteeing (or limiting) the access to the different actors involved. In general, those actors will be data users, but it is also important to consider data controllers, data processors or SPEs identification to communicate with nodes, or the identification of the nodes themselves to communicate with other nodes.

Actors involved

nodes/central services node, data users, data processors, data controllers, data permit authorities

Location options

- **Centralised:** actors involved identify themselves in a central AAI authority, located in the central services node, with specific EHDS2 credentials.
- **Distributed:** actors involved identify themselves using the local node AAI systems, using local node credentials. Then the node is able to forward the credentials among the different nodes
- **Hybrid:** actors involved identify themselves in a federated AAI, where they use their local node credentials that are retransmitted through a central services node to the rest of nodes.

4.5.3 Support & Training Services

The support services include all the software elements to manage the support teams in their activities to help data users, data processors, data controllers or any other EHDS2 actors, serving for example the software to manage incidences or the software to assist users remotely (conferencing systems or remote desktop systems). The training services, closely related to support services, gather the software elements to share with EHDS2 users the formative documentation and applications, as well as the applications to give online seminars and others.

Actors involved

Nodes/central services node, data users

Location options

- **Centralised:** all the required elements to provide the support and training services are located in the central services node
- **Distributed:** the different nodes provide the support and training services software elements, possibly tailored to such node requirements or capabilities.
- **Hybrid:** central services node holds part of the general support and training software, while tailored solutions are provided on each of the nodes to their user communities.

4.5.4 Financial Services

The financial services include all the software elements to manage the financial transactions that might be associated to the EHDS2 fees and operating costs. The financial model to ensure the EHDS2 sustainability is being defined in WP4[14]. Possible options include charging the data users for accessing and using the data, thus requiring the software elements to manage the billing and invoicing or, alternatively, a free-at-

[14] TEHDAS. Preliminary study of funding sources for and costs of the secondary use of health data in the EU. TEHDAS M4.3, 29 March 2022.

point-use model may require services to manage the financial compensation between nodes.

Actors involved

nodes/central Services node, data users, data controllers

Location options

- **Centralised:** all the required elements to provide the financial services are located the central services node.
- **Distributed:** the different nodes provide the financial services, orchestrating the operation across them.
- **Hybrid:** the different nodes provide the financial services while the orchestration is performed through the central services node.

5 Conclusions and next steps

This deliverable presented the options for the minimum set of services required for a proper operation of the EHDS2. The services are structured among the Users' Journey (Section 2.2), the process a data user will follow so as to conduct a health data-based project. Each service is described in terms of its functionality, the location option for its deployment according to the high-level architecture envisaged for the EHDS2 (Section 2.1) and the implications this location may have between services.

The current deliverable serves as input to discuss important elements on other Work Packages of the TEHDAS JA, for example, the funding models for the EHDS2 sustainability in WP4, the governance options and legal requirements, as part of the legal and organisational interoperability in WP5; the data quality elements to guarantee the semantic interoperability in WP6; and, finally, the citizen involvement within the EHDS2 being analysed in WP8.

Finally, the work done on the services definition sets the base to the final deliverable of the current WP7, D7.2 "*Options for the services and services' architecture and infrastructure for secondary use of data in the EHDS*", that will contain an in-depth analysis of architecture and infrastructure options that will support the implementation of such services. Both the current deliverable and the final deliverable will enable readers to envisage how the EHDS services architecture and infrastructure could look like once the EHDS2 legislative proposal will be published.