TEH
DAS
Towards
European
Health
Data
Space

Milestone M7.5

Catalogue of EHDS services for secondary use of health data

9 December 2021

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
Helena Lodenius	CSC, IT Centre for Science, Finland
Carlos Tellería-Orriols	IACS, Aragón Institute of Health Sciences, Spain
Erkki Soini	ESiOR Oy, Finland (Working Package Advisory Group)
Flavio Soares	SPMSS, Shared Services of Ministry of Health, Portugal
Philip Schardax	ATNA, Ministry of Healthy, Austria

0.2 Keywords

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

Accepted in Project Steering Group on 23 November 2021.

Disclaimer

The content of this milestone represents the views of the author(s) only and is his/her/their sole responsibility; it cannot be considered to reflect the views of the European Commission and/or the Consumers, Health, Agriculture and Food Executive Agency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use of its contents.

Copyright Notice

Contents

Summary 3
1 Introduction 4
2 Initial architectural concepts for the EHDS 4
2.1 Business cases (See EHDS Pilots and WP5) 7
3 EHDS User Journey 7
4 EHDS Services 8
4.1 Data discovery 9
4.2 Permit applications, contracts, and training 9
4.3 Consents collection (optional) 10
4.4 Data preparation for use 10
4.5 Data access provision 10
4.6 Data use 11
4.7 Results dissemination 11
4.8 Infrastructure and other services 12

Summary

Within the joint action (JA) Towards the European Health Data Space (TEHDAS), the work package 7 (WP7) “Connecting the dots” will detail the technical options to provide effective secondary use of health data through the European Health Data Space (EHDS). According to the European Interoperability Framework, these solutions that are to be explored in the WP7 represent the technical interoperability elements of the EHDS, that should be aligned with the legal and organisational interoperability, WP5 “Sharing data for health”, and semantic interoperability, WP6 “Excellence in data quality”.

Objective 7.3 of WP7 aims to “Define the options for EHDS services for secondary use of health data”. This document is the mean to fulfil the first milestone (Milestone 7.5) towards this overarching objective. As a result, this document detail the possible services that will be required for a proper EHDS operation, following the envisioned architectural solutions and the user journey definition (i.e., the interaction of the final users of the EHDS with the EHDS itself) identified by the Work Package Advisory Group (WPAG), already presented in Milestone 7.2 and further described in Sections 2 and 3.

The experience of previous data sharing initiatives surveyed in Task 7.1 and reported in Milestone 7.1 was also considered for the services definition. This document also received inputs from discussions and materials produced in other sources, such as the EHDS2 pilots planning initiative, InfAct JA and the PHIRI Horizon 2020 project and the HealthyCloud Horizon 2020 project, among others.

Finally, this document was reviewed by the WP7 participated in the document elaboration.

1 Introduction

The TEHDAS joint action (JA) Towards the European Health Data Space, helps EU Member States, and the European Commission (EC) to develop common practices for the cross-border secondary use of health data to benefit public health and health research and innovation in Europe. The goal of the JA 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 (EHDS). According to the European Interoperability Framework, the solutions that are to be explored in the WP7 represent the technical interoperability elements of the EHDS.

As described in the TEHDAS grant agreement, the specific objectives (O) of the WP7 are the following:

- 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 EHDS and EHDS 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 EHDS 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 aims to meet the first milestone to reach the objective O7.3. This document provides a detailed set of possible services required for a proper EHDS operation (Section 4), aligned with the initial architectural concepts for the EHDS (Section 2) and the User Journey for the EHDS (Section 3), both proposed by the WP7 Advisory Group.

2 Initial architectural concepts for the EHDS

During the initial discussion of the TEHDAS JA proposal and in further conversations and meetings taken place in multiple stakeholder's forums, the currently envisioned architecture for the future EHDS is a peer-to-peer (P2P) system that interconnect nodes that support various functions for health data access and processing, and offers a certain number of services to the users it serves. **Error! Reference source not found.** contains an overarching schema of the envisioned architecture.

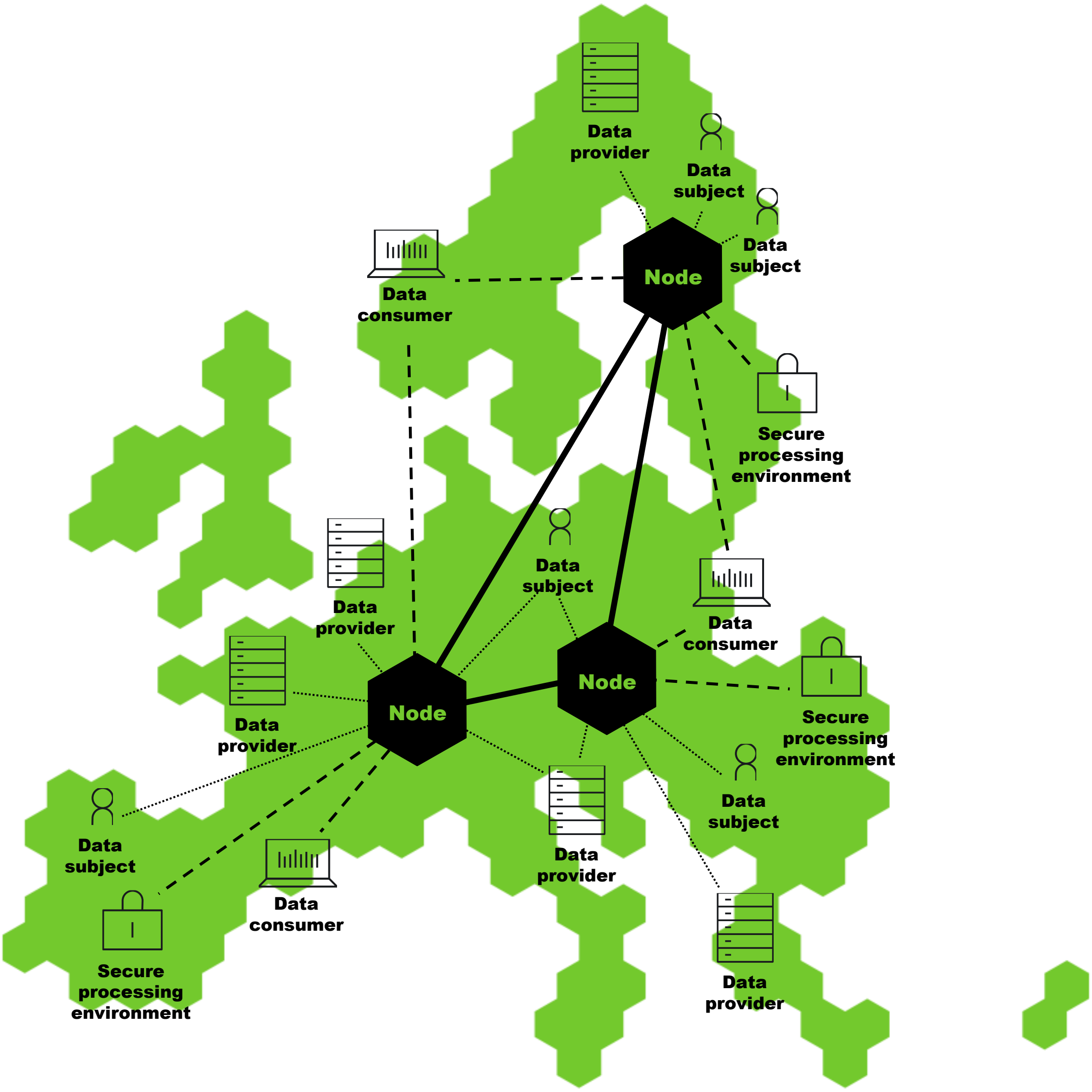

Figure 1 EHDS initial architecture concept

The role played by different elements in Figure 1 is the following:

- **Nodes** are the interface for the final users of the EHDS, encapsulating all the required services to ensure a proper secondary use of health data across Europe.
- **Data consumers** are those EHDS users that will use the data accessible through the EHDS in their day-to-day work, e.g., researchers, policymakers or regulators.

- **Data providers** are those institutions that control existing data sets or registries and put them in common for its secondary use through the EHDS. The data providers' functions include for example:
    - Capturing original data generated directly by data subjects, or when data subjects have been in contact with the health system. Include a wide range of actors from hospitals or primary care settings to pan European registries or research infrastructures and also researchers that collect data for their own studies and want to share it within the EHDS.
    - Guaranteeing the quality of the data to be shared in the EHDS. May overlap with data producers, or may consider actors exclusively dedicated to data curation, or may include data managers (preparation) for analysis.
    - Granting the access to the existing data.
    - Providing technical infrastructures or entry points to single or multiple data sets.
- **Data subjects** are the individuals whose data is gathered for secondary use, that interact with the EHDS to exercise their data control rights or give their consent.
- **Secure Processing Environments** (named after their definition in the Data Governance Act[1]) facilitate secure processing of data in the secondary use context.

Note that the high-level architecture of Figure 1 describes the main legal entities of the EHDS, but does not include the various actors needed in setting up and maintaining the infrastructure, such as:

- Networking: actors in charge of setting up and maintaining the interconnection mechanisms between nodes.
- Computing power: actors in charge of setting up and maintaining the computing resources needed to access, process and analyse the data available.
- Storage: actors in charge of setting up and maintaining the data storage capabilities within data hubs.
- Software solutions: actors in charge of setting up and maintaining the full software stack to be used in the EHDS nodes, from the operating system, authentication, data access permission and rest of services detailed in the present document.

[1] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act). COM/2020/767 final https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

In this complex scenario, EHDS nodes need to provide a harmonised set of services to guarantee the proper functioning of the full EHDS. The services should be aligned with specific and general regulations, at European and national/regional level, and ensure a safe and trustworthy working environment for the different actors involved in the EHDS operation.

2.1 Business cases (See EHDS Pilots and WP5)

In different conversations during the project development, and especially within the “EHDS2 pilots” informal meeting series, an abstract summarising some key EHDS access patterns, also understandable as abstract use-cases, have been defined. Those are introduced in the working document, EHDS2 User Journey, and are the following:

- Answering a time-limited research/policy-making question: A researcher/ policy-maker is authorised to use a certain number of the assets available (data, processing capabilities) in the EHDS to answer a research question formulated in their field. The use case finishes when the researcher has determined that their answers or findings are sufficient.
- Answering a continuous research/policy-making question: A researcher/ policy-maker or institution is authorised to use a certain number of the assets available in the EHDS to answer a research/policy-making question that requires a continuous monitoring of the status and *evolution of data* during a long (possibly unlimited) timespan. This use case requires a data update mechanism to be implemented to deliver data updates to the data requester.
- Development of a service: a researcher requests the data in order to develop a service, e.g., diagnostic support tool for clinical practice based on machine learning or development of a new image analysis algorithm. This can be done federated (e.g., federated machine learning or federated application of generalised linear models for statistics) but may also require pooling the data (e.g., for manual data analysis prior or during the service development).

3 EHDS User Journey

The EHDS User Journey depicted in Figure 2 has been defined in the WP Advisory Group combining inputs also from discussions with other external experts. The user journey is used to guide the work of TEHDAS WP7 in defining the EHDS architecture and service options to be delivered as the WP results. It is the high-level description of the steps that a data consumer will follow for a successful interaction with EHDS, according to the business cases detailed on Section 2.1 for secondary use of health data.

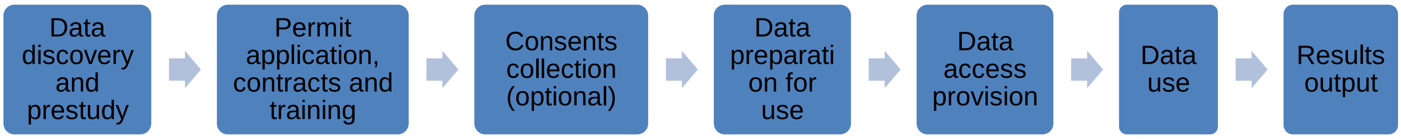
Data discovery and prestudy
Permit application, contracts and training
Consents collection (optional)
Data preparati on for use
Data access provision
Data use
Results output

Figure 2 EHDS User Journey

A brief description of these seven steps (Figure 2) is as follows:

1. **Data discovery and pre-study**. Search and find of data among the EHDS infrastructure, as well as evaluate the data quality and to measure the potential power to perform the required analyses (e.g., number of individual registries).
2. **Permit application, contracts and training**. Includes filling the application for data access, the acceptance of permit application and signing the contracts that define the data access conditions including ethical reviews. In some cases, this step may also imply assisting to online-training courses or study registration.
3. **Consents collection**. If needed, interaction with data subjects for them to opt in/out to the specific projects that use the data from the EHDS (applies only to those data sets, subject to consent).
4. **Data preparation for use**. Pre-processing, for example integrating and harmonizing datasets (e.g., clinical codifications) as well as anonymising/pseudonymising records to make data ready to be used.
5. **Data access provision**. Includes the specific means provided to data consumers to access the data they require, remotely or on premises.
6. **Data use**. Actual processing and analysis of the data in the scope of the secondary use.
7. **Results dissemination**. Actions to guarantee a proper use of the results of the research, including the ensuring of privacy, anonymity, reusability, and appropriate publication of the results.

Note that Figure 2 describes only the main phases of the user journey. There are several other processes and activities that needed to support the EHDS.

4 EHDS Services

This section includes a high-level description of the possible services associated to each step of the EHDS User Journey, following from the analysis of different discussions within the WP Advisory Group (Task 7.2), among other forums. The service descriptions are also based on information collected in the survey of data sharing initiatives in Task 7.1.

4.1 Data discovery

For the data discovery, the possible services are the following:

- Data search: an interface for data consumers to describe the existing data he/she needs and find it in the node registries. In general, the data described is expected to be a cohort of patients with a specific inclusion criterion, such as a given diagnosis or certain type of intervention (codified using a standard encoding system), other characteristics of interest (age, sex, etc.), etc.
- Data search broadcast: a service to send the searches among the EHDS nodes. Desirably, the broadcast should be done in a manner that is transparent to the data consumer doing the search.

In both cases, the results of the data search should provide an informative summary regarding the number of records found, high-level quality measures and other information useful to the data consumer to decide the potential feasibility of the further actions.

In the data sharing initiatives analysed, half of the sites provided a search portal with the capacity to launch queries with some level of complexity, including concept browsers, topic or keyword search, actual data scanners or info about publications based on archived data.

4.2 Permit applications, contracts, and training

For the data permit applications, contracts and training, the possible services are the following:

- Permit application form: an interface or form detailing the specificities of the data request, including the data specification, the end-use or purpose, data consumer information, study protocol, etc.
- Permit acceptance and management: the services to forward the permit application among the EHDS nodes as needed and to coordinate the different boards or bodies in charge of accepting the application. May include data hubs representatives as well as ethics committees' approval.
- Contract signing: after the acceptance of the permit, a service to facilitate the signing of contractual commitments between the appropriate signature parties. Optionally, it may include the registration of the study.
- Training: a service/services of online courses for a proper use of the EHDS services and data. Training may be required as a precondition for the data access permit.

In the survey developed in Task 7.1 the data application processes of the data sharing initiatives were not studied.

4.3 Consents collection (optional)

For the consents collection, if needed, the possible services are following:

- Data subject consent notification: a service to notify data subjects whose data is requested for a particular project that should opt in/out to share the data.
- Data subject sharing control: a service for data subjects to effectively opt in/out for a specific request and also for a wider set of topics or areas.

Note that this is an optional step, that may depend on the data request as well as the specific data set or registry.

In the survey developed in Task 7.1 the consent collection processes of the data sharing initiatives were not studied.

4.4 Data preparation for use

For data preparation for use, the possible services are the following:

- Data retrieval: a service to copy, move or link the data from its original location to the place where the data requester will be able to access it. It might include accessing a single node or forward the requests for retrieval to multiple EHDS nodes.
- Data integration: a service to combine the data from multiple sources to be made available as single coherent dataset, when possible. This includes the data linkage, harmonisation of clinical coding systems, vocabularies and/or data models, etc. and cleanse of potential data artifacts.
- Data anonymisation/pseudonymisation: a service to perform the required modifications to the data to guarantee the anonymity/pseudonymity of the contents, and thus, the privacy of the data subjects according to legal frameworks. May include removal or modifications of variables, the inclusion of noise, or the removal of full records among other techniques (*k-anonymity*).

The data sharing initiatives inquired typically operate as single nodes, so the data retrieval consists of accessing the databases/files where the data is stored and previously integrated. One of the data sharing initiatives surveyed presented a federated approach, where data from multiple sources can be analysed, but the integration is done *a priori*, by ensuring a common data model among all nodes in the federation. Regarding the anonymisation/pseudonymisation processes, in most cases it is performed where the data is collected, e.g., in the information systems of hospitals.

4.5 Data access provision

The possible services for the data access provision are the following:

- Data download: a service where the requested data is placed temporarily, and the data requester can access and download it to a processing environment under their own control.

- Application programming interfaces (API) to enable machine-to-machine access to the requested data programmatically or to remotely execute algorithms on the data, facilitating the development of rich and complex analysis applications.
- Secure processing environment: a service that provides data requesters an interactive environment to access, process and analyse the requested data in a secure way, for example via a remote desktop server, enabling its safe storage and archiving.

In most of the data sharing initiatives surveyed, the data access provision was based on secure processing environments. Alternatively, many initiatives provided a data download service, for example through secure file transmission protocol (SFTP) servers or authenticated web portals. No data sharing initiative detailed the possibility of accessing through APIs.

4.6 Data use

For the data use, the possible services are the following:

- Statistics software: a service for secure processing environments and for those users that downloaded the data is to give access (via licenses or open-source tools) to statistics software (e.g., R, Stata, SPSS, SAS) to analyse the data.
- Analysis runtimes: a service for secure processing APIs and for those users that downloaded the data to give access (via licenses or open-source tools) to data analysis runtimes and libraries (e.g., Pandas, Mathlib or sklearn for Python language) to analyse the data.
- Distributed runtimes: a service for secure processing APIs to give access to tools and libraries to developed distributed/parallel/federated applications. The purposes could be to analyse large volumes of data in a single node (e.g., performing the analyses in a large computer cluster/supercomputer using Spark), or to avoid the data motion between nodes (e.g., using TensorFlow Federated for federated learning approaches).

In the data sharing initiatives surveyed, the services provided for data use have been mostly statistics software (R, Stata, SAS or SPSS), followed by those that offer analysis runtimes or programming environments (mostly Python) and, in a few of them, distributed runtimes (Spark). In some initiative, it has been stated that data users may ask to install those tools they need to perform their analyses in the secure processing environment. Nonetheless, the tools to be installed will need to fulfil all the security requirements stated by the institution running such secure processing environment.

4.7 Results dissemination

For the results dissemination, the possible services are the following:

- Results storage: a service that enables results with related metadata and materials to be stored and archived for potential re-use in further projects. Data and metadata, and any other relevant project documentation, should be included to guarantee the reproducibility of the analyses by other researchers.

- Results cataloguing: a service of harmonisation, cataloguing and indexing of the long-term archived results, to ensure its findability for future users with similar interests.
- Document repository: a service to store the documents created while using the requested data, and, optionally, for long term archive. It might also serve as a preprint archive for research papers.
- Results communication: a service to communicate the results obtained while using the data from the EHDS. It might consist of a dedicated space in a web page for interactive web apps or the inclusion links to the results in EHDS related social media accounts.
- Notification of clinically relevant findings: a service to notify data subjects (via the relevant clinical processes) about outcomes of the research, especially in the case of actionable findings.

In the survey developed in Task 7.1 the results dissemination processes of the data sharing initiatives were not studied.

4.8 Infrastructure and other services

Some of the identified services are not directly related to a given step of the user journey, but are needed in the infrastructure:

- Authentication and Authorisation and Identification: a federated identity management and role-based authorisation service to support the EHDS processes.
- Data donation: a service that provides the possibility to include health or health-related data donated by individuals directly or via research projects. Such services are foreseen to be implemented by specific data providers.
- Node network management: set of services needed to maintain and extend the node network. The services cover, for example, the process of annexing a new node with the necessary auditing procedures.
- EHDS communication: services for general awareness and communication of the EHDS and its services to the stakeholders and general public.
- Financial management: services to manage tenders, billing and invoicing of EHDS.