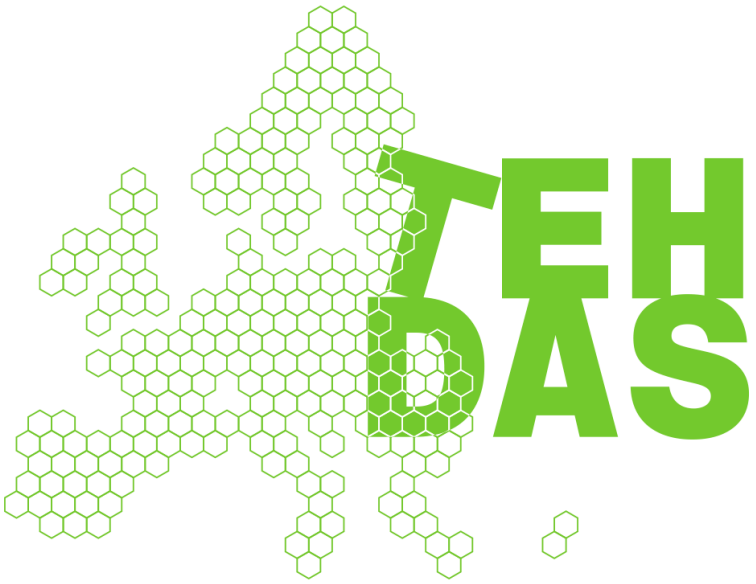

Milestone 7.6

Report on architecture and infrastructure options to support EHDS services for secondary use of data

24 March 2023

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

Document info

0.1 Authors

Author	Partner
Enrique Bernal-Delgado	Aragón Institute of Health Sciences, Spain
Fidelia Cascini	Ministry of Health, Italy
Sergio Dinis	Serviços Partilhados do Ministério da Saúde, Portugal
Francisco Estupiñán-Romero	Aragón Institute of Health Sciences, Spain
Yasmin Fonseca	Serviços Partilhados do Ministério da Saúde, Portugal
Juan González-García (JGG)	Aragón Institute of Health Sciences, Spain
Lionel Grondin	Health Data Hub, France
Truls Korsgaard	Directorate of e-health, Norway
Vanessa Lima	Serviços Partilhados do Ministério da Saúde, Portugal
Helena Lodenius	CSC, IT Centre for Science, Finland
Klara Lundgren	Directorate of e-health, Norway
Jaakko Lähteenmäki	VTT, Technical Research Centre of Finland, Finland
Juha Pajula	VTT, Technical Research Centre of Finland, Finland
Marja Pirttivaara	Sitra, Finnish Innovation Fund
Katharina Schneider	Health Data Lab, Federal Institute of Drugs and Medical Devices, Germany
Anne Heidi Skogholt	Directorate of e-health, Norway

Carlos Tellería-Orriols	Aragón Institute of Health Sciences, Spain

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 28 February 2023.

Contents

Executive summary

The Joint Action (JA) Towards the European Health Data Space (TEHDAS), helps EU Member States, and the European Commission (EC) to develop a common framework 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 for secondary use of health data (HealthData@EU, informally "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."

This document presents a synthesis and refinement of the Deliverable D7.1 "*Options for the minimum set of services for secondary use of health data in the EHDS*"[2], delivered in March of 2022. where the catalogue of the possible services as well as the deployment options was presented. The synthesis and refinement presented here is based on the analysis of the evolution of the HealthData@EU architectural descriptions, starting from the EHDS legislative proposal, presented in May 2022, as well as the rest of advancement around the HealthData@EU infrastructure, for example, the prospection work being done in the HealthData@EU pilot project[3]. This milestone will serve as the basis of the final Deliverable of WP7 D7.2 "*Options for architecture and service infrastructure and services for secondary data use in the EHDS*", to be delivered in May 2023.

In addition, in this document it is also presented a dissertation in the implementation options of in three components that will be required in the HealthData@EU infrastructure: the information systems to manage metadata catalogues, the information systems to manage the cross-border data access requests, and the secure processing environments. This dissertation will constitute the respective guidelines that will accompany the final deliverable D7.2, as per request of the European Commission.

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

[2] TEHDAS Milestone M7.5 "*Options for the minimum set of services for secondary use of health data in the EHDS*" https://tehdas.eu/results/tehdas-suggests-minimum-technical-services-for-the-european-health-data-space/

[3] https://www.ehds2pilot.eu/

1 Introduction

Within the TEHDAS Joint Action, 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 (HealthData@EU, informally "EHDS2"). As collected in the TEHDAS glossary[1], the secondary use of data is defined as "using data for a purpose different from the purpose for which the data was initially collected."

According to the European Interoperability Framework (EIF)[4], the solutions to be explored in WP7 represent the technical interoperability elements of the HealthData@EU infrastructure. 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 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 constitutes the fifth milestone achieved within the WP7[5], which addresses O7.4, using as inputs the results described in the previous milestones were objectives O7.1 to O7.3 were addressed. In particular, this milestone represents a synthesis and refinement of the Deliverable 7.1 "*Options for the minimum set of services for secondary use of health data in the EHDS*"[6], that included the catalogue of the possible services as well as the deployment options. This milestone provides a further

[4]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

[5] As a curiosity, due to project management issues, the order of achievement of milestones has bee M7.1, M7.2, M7.5, M7.3 and the present M7.6

[6] TEHDAS Milestone M7.5 "*Options for the minimum set of services for secondary use of health data in the EHDS*"
https://tehdas.eu/results/tehdas-suggests-minimum-technical-services-for-the-european-health-data-space/

view on the evolution of the Users' Journey and architecture proposals made during the Joint Action, a further dissertation regarding the possible implementations for the described services, putting some stress in the possible implementation of in three components that will be required in the HealthData@EU infrastructure: the information systems to manage metadata catalogues, the information systems to manage the cross-border data permits requests, and, the secure processing environments. The analysis of requirements and specific solutions for these three components were discussed in dedicated workshops with internal and external stakeholders.

It is especially notable the work done around the secure processing environments, as its conception plays a central role in the HealthData@EU infrastructure. As all the individual level data will be legally obliged to be analysed in such systems, it is required to reach a large consensus on the requirements of secure processing environments. The requirements will be not only in technical terms, but also in semantical and organisational terms. This report presents a large and detailed discussion about these systems.

To conclude, just to mention that all the discussion presented in this document serves to prepare the final proposals for architecture, infrastructure and service deployment to be included in the last deliverable of this work package: D7.2 "*Options for architecture and service infrastructure and services for secondary data use in the EHDS*". In addition, the final deliverable will include structured guidelines for deploying three key components already mentioned. These three guidelines are part of a direct request given by the European Commission to this work package.

2 TEHDAS Users' Journey

2.1 WP7 analysis framework evolution

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 (HealthData@EU, informally "the EHDS2").

Two main aspects have been already addressed: the first is the high-level architecture envisaged for the future HealthData@EU. This high-level architecture contains the relation between computational elements and HealthData@EU actors (covered in the next Section); the second is the "Users' journey", the definition of the process that a data user must follow to access and use the data available in the HealthData@EU. These two aspects have been under constant discussion, review and improvement as part of the WP7 activities, the cross-cutting WP activities and further interactions with external stakeholders. This Section presents how this work has been reflected in the evolution of the User's Journey.

2.2 Updates on the Users' Journey

The TEHDAS user's journey is the process describing the interaction of different actors with different roles (as the EHDS regulation - currently under discussion - will establish) to make data available for secondary uses through the HealthData@EU. Based on different steps, the institutions acting as health data access bodies (HDABs) may grant the access to data of interest to the end user who asked for them after the data discovery and the permit application. The user's journey is about how to access and use the actual data, and how to finalise the use of data including devolution of intermediate outputs and enriched dataset.

The Users' Journey is also used to guide the work of TEHDAS WP7 in defining the HealthData@EU technical infrastructure in terms of service options and architecture to be delivered as WP results.

2.2.1 *The original* TEHDAS User's Journey

The original TEHDAS User's Journey was designed as a high-level service process for secondary use of health and social data including 7 steps (*Figure 1*) and in particular:

1. **Data discovery and prestudy**. This step was conceived for: searching and finding data; evaluating the availability of needed data types, data quality and number of subjects (available statistical power); open service carefully designed not to leak sensitive information.
2. **Permit application, contracts and training.** This is the step concerning: application for data access; application processing including ethical review; contracts specifying conditions for data use (e.g., definition of data processing

environment) and training the user for responsible use of data (both e-learning and helpdesk services).

3. **Consents collection (optional).** The third is an optional step, in case informed consent is needed, and the data subjects are invited to provide their consent for the study. It has to be noted that this consent is related to the secondary use of the data (not the consent that is required in the context of clinical trials). Further, the need for consent in the secondary use context varies among countries (interpretation of legislation) and use cases.
4. **Data preparation for use.** This is the step related to the pre-processing and other actions to make data ready for use, e.g., integration of registers ("real" or "virtual"), filtering, ensuring data quality and security. As an optional, it is the provision of synthetic data.
5. **Data access provision**. The fifth step of the process includes three options: (a) online access to secure processing environment (in control of EHDS), (b) online access to download data to a user-controlled secure processing environment, (c) online access to upload (or choose) algorithms for data processing in a secure processing environment (in control of EHDS or original data controller)
6. **Data use**. This is the step for data analysis and processing in the scope of secondary use of health and social data.
7. **Results output**. The last, it's the step for actions to ensure anonymity, reusability and appropriate publication of results. For example: verifying that identities of study subjects cannot be recovered; enabling results to be reproduced and verified by independent groups; archiving of results; sharing of study protocols, analysis SW and data queries. It includes actions to ensure personally targeted feedback, information of usage of personal data and reporting of incidental findings (as appropriate and as accepted by the data subject).

Figure 1: Original TEHDAS' Users' Journey

2.2.2 The revised User Journey

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[7]. 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. The revision of the Users' Journey also makes explicit some of the services that were not depicted in the previous version in Milestone 7.5.

[7] TEHDAS Milestone 7.5 "*Catalogue of EHDS services for secondary use of health data*" https://tehdas.eu/results/tehdas-proposes-european-health-data-space-services/

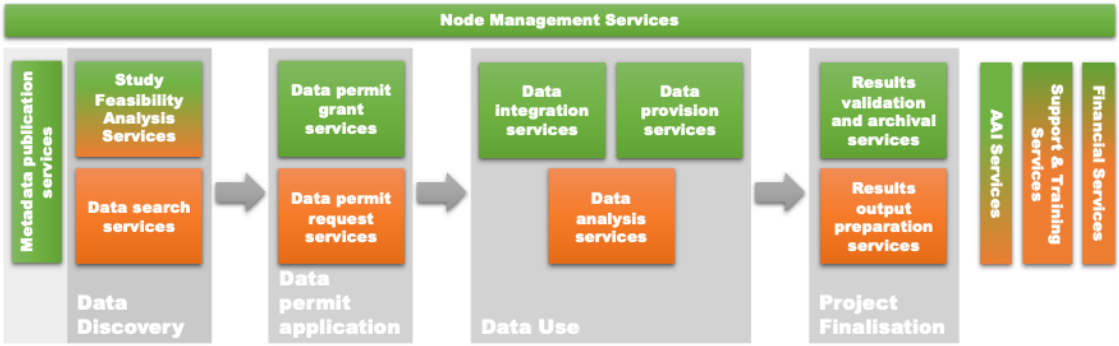

Figure 2: Second version of TEHDAS Users' Journey

In the schema *(Figure 2)* green boxes represent those services that are related to the EHDS2 point-of-view, i.e., the services that are conceived to involve data controllers, data permit authorities and other actors than data users. The orange boxes represent those 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 2* there is an attached block regarding the metadata publication services, this is due to the fact that the metadata publication services, are not essentially part of the User Journey, but a prerequisite 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 permission 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[8] for the results. It may imply a notification of the incidental findings to the data controllers.

2.2.3 The TEHDAS' data lifecycle

[8] 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

The TEHDAS' data lifecycle is depicted in *Figure 3*. It is an extension of the User's Journey that includes the data holder phases required to make the data available for its further analysis, grouped as data preparation in the Figure, but sometimes informally named as the "Data holder's journey". The proposed data lifecycle incorporates the Publication phase as part of the duties of the data holders. This phase was previously included in the second loop of the TEHDAS User's Journey as a prerequisite for the Data Discovery, depicted in *Figure 2* as the Metadata publication services.

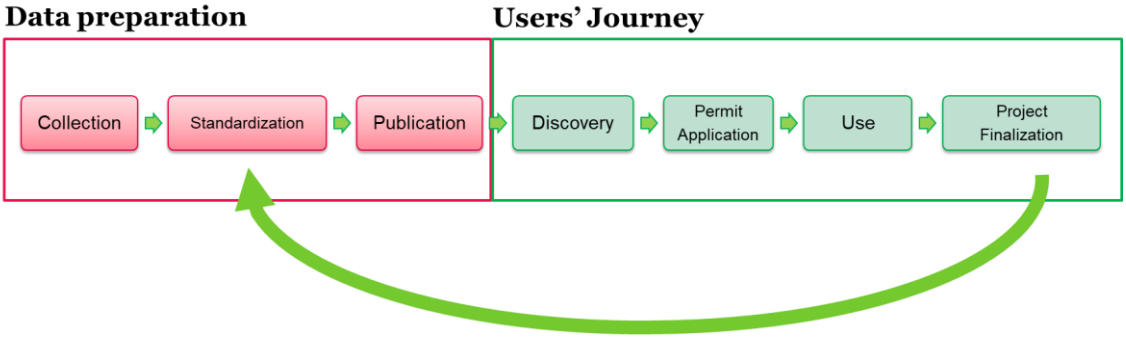

Figure 3: TEHDAS' proposed data lifecycle

2.2.4 The Users' Journey for the HealthData@EU pilots

Finally, in the HealthData@EU pilots, the European Commission provided a pre-work, proposing a Users' Journey with slight modifications over the TEHDAS proposal, depicted in Figure *4*.

Figure 4: HealthData@EU Users' Journey

HealthData@EU includes the cataloguing phase, as the legislative proposal includes an EU level catalogue, while it was an embedded service of the TEHDAS 'Data Discovery' phase. The data discovery and prestudy phase is equivalent in both User's Journey. The TEHDAS Data Use phase is divided in two phases the Data preparation and provision, which corresponds to Data integration services and Data provision services, depicted the top services in the Figure *2* 'Data use' phase, where the user has no intervention, and the Data use which corresponds to the Data analysis services in the TEHDAS User's Journey. The Results output phase of the HealthData@EU User's Journey encapsulates the services of the Project Finalisation phase of the TEHDAS' User's Journey.

There is no specific mention in the HealthData@EU User's Journey of the Node Management Services, AAI Services, Support & Training Services and Financial Services introduced in the Deliverable 7.1.

The rest of the document is based on the second iteration of the TEHDAS' Users'

Journey presented in Deliverable 7.1, which corresponds to the one depicted in Figure *2*. Where stated, the report may refer to the TEHDAS' Data Lifecycle, presented in Figure *3*, in particular to the publication phase, regarding the manipulation of the metadata catalogues, as these metadata publication services where detailed originally as the prerequisite for the data discovery phase.

2.3 Minimum services identified

Deliverable 7.1 included the list of the minimum services identified to guarantee a proper operation of the EHDS for secondary use. The services are the ones listed in Figure *2* boxes. Deliverable 7.1 provided a wide view of possible implementation and deployment options. In the Section 5 of this document there is an extension of such work, deepening in three key elements: the metadata publication systems, depicted as metadata publication services in the Users' Journey (Figure *2*) and the core of the Publication phase (in the data lifecycle, see Figure *3*); the data permit application systems that cover the Data permit application phase and the secure processing environments, that cover the Data use phase.

3 Architecture Scenarios

3.1 WP7 architecture evolution

As in the User's Journey, the architecture proposal has evolved during the JA and has influenced (and has been influenced) by the legislative proposal and the current HealthData@EU pilot.

3.1.1 First TEHDAS architecture

The original TEHDAS' architecture proposal, depicted in Figure *5*, was presented in Milestone 7.5 and introduced a pure peer-to-peer architecture (more details on this in section 0), where member states operate 'Nodes' (orange), that connect to each other, and serve as a frontend to 'Data consumers' (green, the actual users of the architecture) to the data search and data permit request related services, already identified the first TEHDAS' Users' Journey (Figure 1). 'Data providers' (black) and 'Secure Processing Environments' (blue) will intervene to make the data available for its use. Data providers will also support the search services. Finally, this initial architecture also included 'Data subjects', foreseeing the optional consent that was later removed in following TEHDAS' Users Journey (Figure 2).

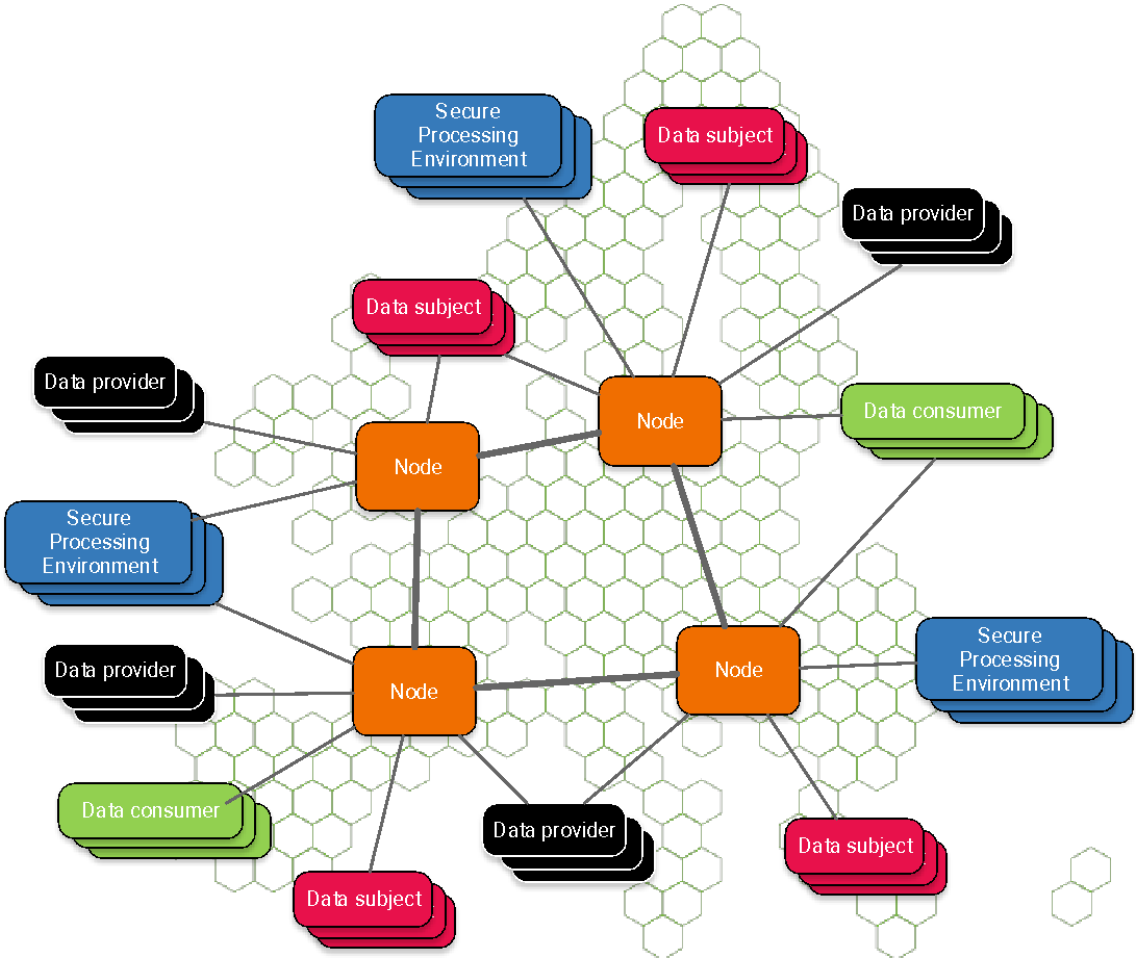

Figure 5: Original TEHDAS architecture proposal (Milestone 7.5)

3.1.2 Second version of the TEHDAS architecture

The second version of the TEHDAS' architecture proposal, depicted in Figure *6*, was introduced in Deliverable 7.1 and is an evolution of the first one. This architecture proposal maps the original 'Data providers' into roles of the GDPR (processors and controllers). In this case, the data subjects are not directly involved in the HealthData@EU operation, as the consent to use their data relies on their relationship with the data controller, in coherence with the second version of the TEHDAS' Users' Journey. In this new architecture, a new 'Centralised services' node is introduced moving towards a hybrid architecture for the services deployment. The discussion of the possible services deployment is the core discussion of Deliverable 7.1.

In this report, this discussion is extended, focusing mostly on the hybrid scenarios and extending specific services that result critical for the operation of the HealthData@EU infrastructure.

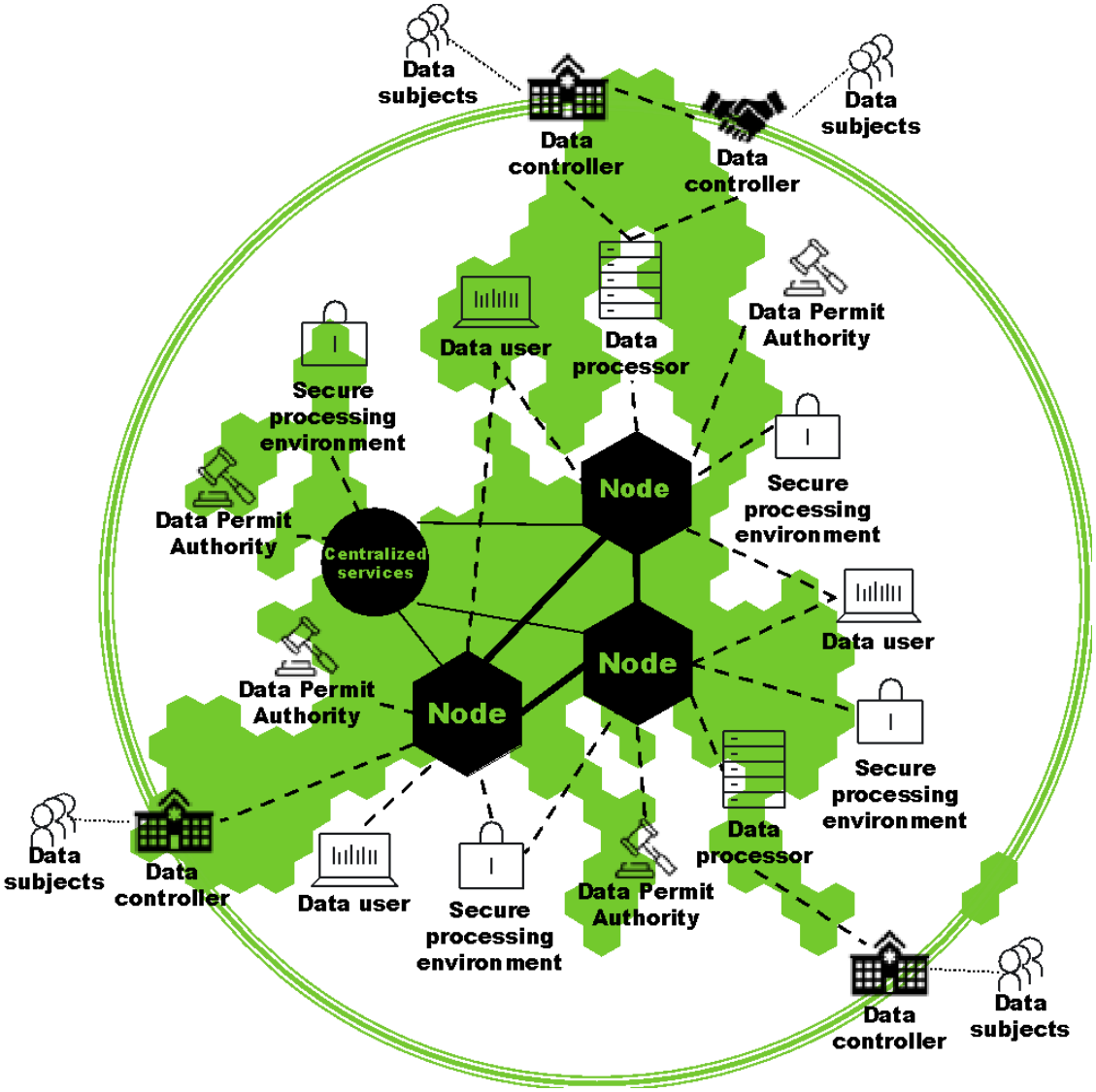

Figure 6: Second version of the TEHDAS architecture proposal (Deliverable 7.1)

3.1.3 HealthData@EU architecture proposal

Figure *7* presents the schema of the HealthData@EU architecture, defined in the Article 52 of the EHDS legislative proposal, but using the same drawing as the TEHDAS architecture proposal of Figure *6*. It is clear that the TEHDAS architecture has a direct mapping in the HealthData@EU one, being mostly a "renaming" of the actors participating on it. "Data processors" and "Data controllers" of the TEHDAS proposal (GDPR roles) are mapped as "Data holders" (EHDS proposal and Data Governance Act roles), "Data permit authorities" are mapped as "Health Data Access Bodies" (HDABs), please note here that there won't be a HDAB attached to the "Core Platform", the "Central Services Node" in the TEHDAS proposal. To conclude, it is important to clarify that the "Nodes" defined in the TEHDAS proposal are depicted as the "National Contact Points for Secondary Use" (Art.52(1-2)), but there has not been an explicit inclusion of other "Authorised participants" referred to in such an article. This has been due to the indefinion of its participation in the HealthData@EU infrastructure in the EHDS proposal.

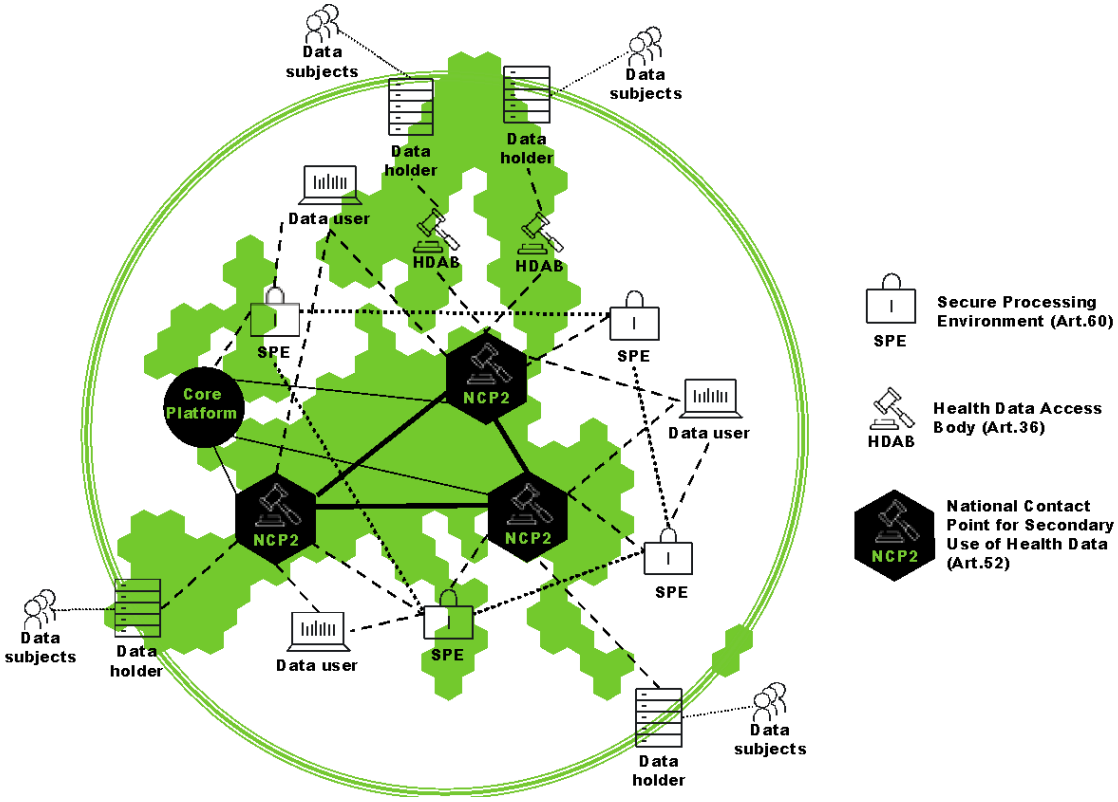

Figure 7: HealthData@EU proposed architecture (adaptation from EHDS legislative proposal)

Finally, Figure *8* includes a simplified schema of the one depicted in Figure *7*, for the sake of clarity, depicting how the actors interact within a single country and its cross-border connection. In this last architectural figure, there have been a couple adjustments. First, Secure Processing Environment ("SPEs") are renamed as "SPEs operators" to differentiate the technical solution (the SPE itself) to the actor (the SPE operator or provider itself). Second, there has been a direct connection between "Data subjects" and the "Health Data Access Bodies" as per the requirement to inform of possible incidental findings explicit in the Article 38 (3) of the EHDS regulation proposal.

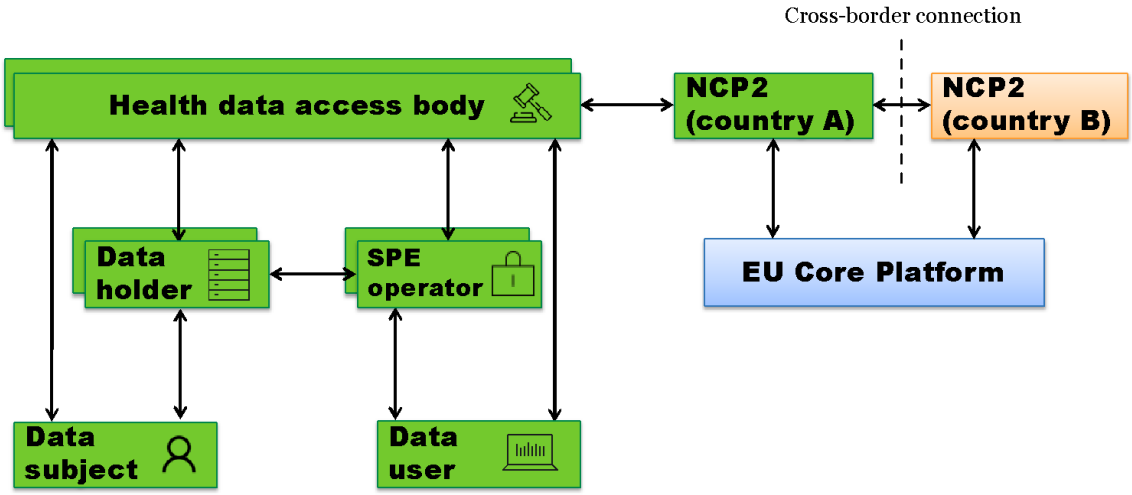

Figure 8: HealthData@EU simplified architecture

The rest of the document is based on the HealthData@EU architecture proposal both in terms of the actors and its interrelationships. It eases its communication only a single terminology is used and, with minor changes, it has a direct mapping to the architecture previously proposed in the TEHDAS Joint Action.
The main exception regarding the terminology is the regular use of the term “Node” along with the document to refer to National Contact Point for Secondary Use of health data.

3.2 Architectural options for services deployment

The architecture presented in the previous section is flexible enough to support different approaches for services deployment, i.e., how the different parts of the overall services (in general, software pieces) are distributed in the architecture to provide such service. Here there are briefly described their particularities. There is a distinction between a centralised approach (section 3.2.1) and a distributed approach (section 3.2.2). In general, from the different options presented here, the hybrid distributed approach is the first option, as it facilitates the interaction between the MSs, mediated by the EU Core Platform, foreseen in the legislative proposal, balancing the responsibilities of the different actors.

3.2.1 Centralised deployment

A single actor/component in the architecture has all the information and pieces to provide a given service. For example, a search service implemented using a central catalogue that resides in the EU Core Platform.

3.2.2 Distributed deployment

Multiple actors/components in the architecture have the information to provide a given service, namely the EU Core Platform and the rest of nodes (the national contact points for secondary use). There might be different distributed approaches depending on how the implied actors are organised.

Client-server deployment

In a client-server architecture deployment, there is a node that becomes the server, namely the EU Core Platform in the Figure *7*, and it oversees coordinating the rest of the nodes in the infrastructure to provide such service. It is the only node that data users should contact to access the service. For example, when searching for a particular data set, the data user should inquire about the Core Platform that will then consult the rest of the nodes to check the data availability.

In this architecture there is no interaction within “regular” nodes, but only between nodes and the core platform.

Peer-to-peer (p2p) deployment

In a peer-to-peer architecture deployment[9], the services are deployed in a way that all nodes communicate to each other to perform such services, this is due that all nodes have part of the information required to offer such service. For example, to implement a search service in a p2p deployment, every single node may launch a search to the rest of the nodes, acting as a server in the infrastructure.

Hybrid deployment

A hybrid approach is not a fixed pattern on where to place the different elements pieces of a service but a concept where some parts of parts of the service reside in the different nodes that are assisted by other parts available in the EU Core platform to provide such service to data users.

In the present document the main deployment foreseen is the hybrid deployment. So, in most of the service scenarios description different hybrid scenarios are discussed.

3.3 Data lifecycle and architecture actor’s involvement

Figure *9* contains a schema depicting the participation of the different actors described in the architecture proposal in the data lifecycle from Figure *3*. This Figure is very useful to clarify how the foreseen actors should provide inputs and interact in the different phases of this process. It is important to note that this mapping takes into consideration multiple deployment scenarios options, for example, the Core Platform is including in the ‘Permit Application’ phase, because a possible implementation of the permit application services (permit request and permit grant services, detailed in Figure *2*) includes the interaction between NCPs and the Core platform to ease in the coordination of the data permit request management between MSs (a hybrid approach). In p2p deployments, Core Platform might be removed for such Permit Application phase.

[9] 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.

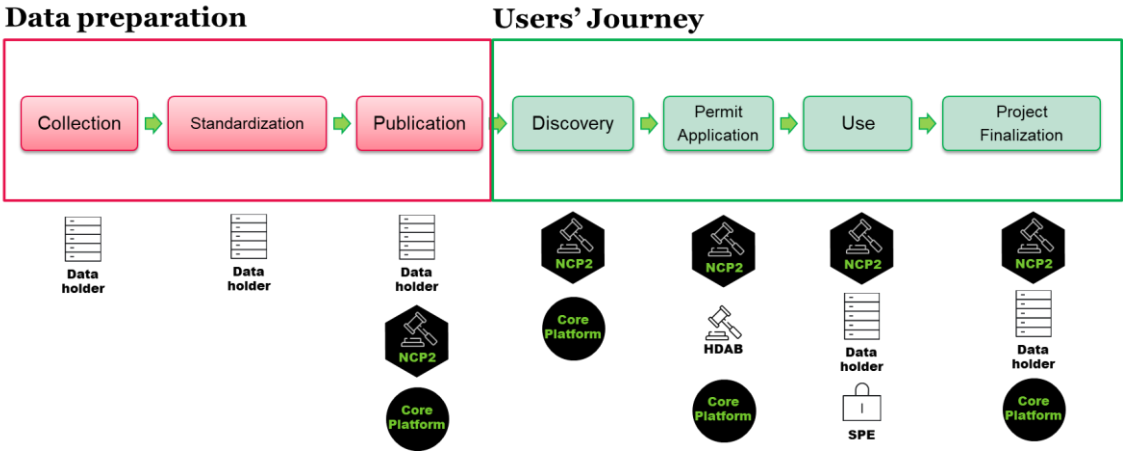

Figure 9: TEHDAS' Data lifecycle mapping HealthData@EU architecture actors

4 Options for services implementation

The aim of this section is to go deeper into the requirements of the services identified in the previous milestones and deliverables of this work package. Specifically, it focuses on adapting such requirements to the updates on the architectural specifications derived from the EHDS legislative proposal.

As introduced in previous sections, the architectural specifications seem to have a clear view towards a hybrid deployment of such services, in some cases tending to a client-server approach, while in others having more a peer-to-peer flavour. This section has the aim to present a discussion on this hybrid approach, presenting possible levels of "hybridness" that may be considered in the future services development.

In addition to the evaluation of the "hybridness" level, there has been a large effort on integrating the request made by the European Commision to produce guidelines for three components of the HealthData@EU:

- Guidelines for national dataset catalogues publicly available to register and facilitate the discovery of health datasets available for secondary use (Art. 37(1)(q)(i))
- Guidelines for management systems to record and process data access applications, data requests and the data permits issued, and data requests answered (Art. 37(1)(K))
- Guidelines for Secure Processing Environments (technical, information security and interoperability requirements). (Art.50(4))

These three guidelines will be provided in the final deliverable of this work package, but its analysis has a substantial impact on the work presented in this section. Discussion around national datasets catalogues is deeply introduced in section 4.1.1, on the Metadata publication services. Discussion related to the management of data access applications, data requests and data permits is present along section 4.2, as this

component covers both the data permit grant service and the data permit request service. Finally, the discussion related to the secure processing environments represents a big piece of the materials exposed in section 5.3, related to the data use phase.

4.1 Data discovery phase

The first phase of the users' journey to request access to health data for secondary use is the Data discovery phase. In this phase the data user should be able to search the data available and needed to perform their work.

To do so, it is that the information available in the different data holders is properly catalogued and such catalogues made available to be inquired. In general, the catalogues are expected to contain a set of metadata describing general features of the datasets. Then, the discovery will rely on search services built on top of this metadata catalogues.

4.1.1 Metadata publication services

As introduced in the Section 2.2, the TEHDAS' Users' Journey (Figure *2*) foresees a metadata publication service, i.e., the cataloguing service, that acts as a prerequisite of the actual data discovery phase. It is a prerequisite because it does not imply an explicit user interaction, and this is clearly depicted in the TEHDAS' Data Lifecycle (Figure *3*), as part of the data preparation process, to be taken at data holder or HDAB level.

According to the legislative proposal, each Member State (MS) must deploy a national dataset catalogue, settled in a HDAB (Art.37(1)(q)(i)). The internal coordination to generate the national datasets catalogue reflect the potential of having multiple HDABs in a MS connected to a coordinator HDAB (Art.36(1)). The decision to be made in each MS is to choose either a centralised catalogue service published by the coordinator HDAB, a distributed catalogue service published among local or regional HDABs (aligned to data holder services), or a hybrid approach where both scenarios are in place, depending on its technological infrastructure and deployments options.

The EHDS legislative proposal also introduces a central European Dataset Catalogue (Art. 57), where the data users also can perform searches to find common datasets in different Member States. It can enable multicentric research and health policies decision-making on a broader level. For this purpose, the coordinator HDAB in the MS shall coordinate the publication of a national dataset catalogue to interact with the EU Dataset Catalogue. The EU Dataset Catalogue also aims to publish health metadata available from other EU Agencies and Research Infrastructures (RIs) services, public Portals comprising aggregated data, either local at the MSs or European portals.

Although the distributed organisation scenarios between a coordinator HDAB and the local or regional HDABs can be feasible, it is foreseen that developing the centralisation of the national datasets catalogue at the coordinator HDAB will ease the semantic interoperability among the MSs nodes and the EU Datasets Catalogue. Other

advantages could rely upon the semantic harmonisation across the metadata publication services of each data Holder responsible for their metadata descriptions in a given MS. In addition, it would be recommendable that the metadata publication services of the coordinator HDAB should be built by design using the standard metadata standard adopted by the EU Datasets Catalogue.

This interoperability will allow a compatible technological environment that supports the communication between nodes and the deployment of the computational tasks, and the existence of common data models that enables semantic standardisation across data sources.

In this scenario where a centralised metadata publication service is in place represents a governance led by the coordinator HDAB who should manage the national datasets catalogue. In this case, the coordinator HDAB could promote the initial articulation among data holders of a MS, reinforce their cooperation, provide national and European legislations and guidelines to create a dataset fulfilling standards and metadata structures required to its publication. It could also support the clarification of the interfaces in use and the integration processes with such national metadata catalogue.

The preparation of a national dataset catalogue shall include, as minimum requirements, the metadata descriptions, such as the source and nature of electronic health data and the conditions for making electronic health data available (Art.37 (1) (q)(i)). Since the data holders own and grasp their health data, they will be responsible for the creation of a particular dataset metadata. It comprises gathering the description of the dataset, its characteristics and, where feasible, providing an exploratory analysis of the data. This analysis could present more information about the dataset, for example, its coverage, null data, average, standard deviation, percentiles. It is relevant to clarify and to ensure to the data holder that creating a particular dataset and submitting it into the national dataset catalogue does not involve sending any health data, nor personal data.

If a particular data holder does not have the technological infrastructure required to ensure the integration processes automatically, in a centralised schema, the coordinator HDAB, or the closer HDAB related to the data holder could maintain the metadata of the dataset and reduce the data holder burden by being responsible for a manual process of updating.

The onus of developing a centralised national datasets catalogue relies upon the coordinator HDAB of the MSs deploying and funding the technological infrastructure. This means that the coordinator HDAB could be responsible for automatically collecting the metadata, i.e., the *harvesting* of the metadata, and its updates from the data holders. It also could indicate that the decision about which dataset should be created and the data available for secondary use depends on the data holders. Once the dataset is structured and its metadata published in the national datasets catalogue, the updates could be initiated by the data holder or the HDAB, a decision to be made by the MS node. Nevertheless, to ensure a periodical update of datasets' catalogue, coordinator HDAB-led management can be seen as an advantage.

Finally, it is worth mentioning that it would be possible to plan the coordination of Open Data repositories with actual health data repositories. That might be useful in certain types of studies that combine, for example, contextual data with the patient's data, for example to evaluate the environmental effects in individual's health.

Table *1* contains the analysis of the possible scenarios proposed in the above text.

Table 1: Scenarios for metadata publication services

Scenarios	**Pros**	**Cons**
Multiple data holders that connect to a single HDAB	The HDAB manages the Catalogue, the organisational interoperability and its updates. It promotes the adoption of the same standard among Data Holders.	HDAB becomes the only responsible for deploying and funding the technological and organisational infrastructure.
Multiple HDABs connecting a certain number of data holders and one coordinator HDAB	Each HDAB deploys a metadata publication service. It will allow the control of the data accessed.	A second step is needed to send the datasets information to the national datasets catalogue. The central catalogue at coordinator HDAB needs to check the compliance of the standards and local/regional catalogue structure to promote the interaction with EU Dataset Catalogue.
Open Portals linked to HDABs HDABs	Possibility to combine more inputs, beyond personal data.	Open portals with aggregated data need to use the same standard as the National Dataset Catalogue to allow the publication of its metadata. Linkage issues between open data and individual level data may lead to ecological fallacies.
Coordinator HDAB contact EU Core Platform bodies publish their metadata to the EU Dataset Catalogue	Coordinator HDAB can finely tune the datasets catalogue synchronisation as it has a direct control of the national datasets updates	Extra burden on the coordinator HDAB technological solutions. Malicious attacks may pollute the EU Datasets Catalogue.
EU Core Platform harvests national datasets catalogue from coordinator HDAB to generate the EU Dataset Catalogue	The responsible to of keep the EU Datasets Catalogue is also in charge of gathering its pieces Leverages the technological burden of the coordinator HDAB.	Central EU Datasets Catalogue may be outdated at some points. EU Core Platform may incur in high capacity requirements on each EU-wide update.

4.1.2 Data search services

The data search is a service that will fully interact with the metadata publication service. Specifically, the search capabilities are directly influenced regarding where the metadata catalogues are placed, the information they contain and how is this information.

The first two points, regarding the metadata catalogues placement and the information they covered the present legislative proposal, as introduced in the previous section: the MS will need to provide a national datasets catalogue and there will be an EU Datasets Catalogue, a collection of all the national datasets catalogue. The existence of this hierarchy of catalogues implies that the data users may use the national catalogues to perform searches within the datasets stored in given MS, while the EU Datasets catalogue will be the entry point for a cross-border search. This situation presents a scenario where it is expected that the EU Datasets Catalogue will be the main system inquired to perform the data searches.

Table 2: Possible scenarios of the data search services

Scenarios	Pros	Cons
An EU datasets catalogue with metadata on "all levels"	Concept of "single-stop-shop" for discovering data in the infrastructure.	Single point of failure, with large computing capabilities.
An EU Central metadata catalogue with only metadata on data source level and URL to more detailed metadata catalogues at national datasets catalogue	Lighten the concept of "single-stop-shop" with closer involvement of the data holders. Less burden to EU Datasets Catalogue systems.	Extra coordination work between EU Datasets Catalogue system and coordinator HDAB in technical and semantical terms.
Search available on each coordinator HDAB, and/or other entry points, independently to the metadata capabilities of choice.	Multiple entry points to the search services that might be tailored to specific communities.	Same as scenario 2, but with extra replication of implementations per coordinator HDAB and/or other participants.
An EU Central metadata catalogue with or without metadata on data source level, but with open data of different kinds	Extra features focusing on open data searches. May offer a larger variety of data to analyse.	Extra burden to integrate the open data catalogues searches.

In any case, it remains undecided whether national datasets catalogue or EU Datasets Catalogue will be exposed through dedicated search applications, such as web portals, and, if so, what would be its interaction. As per the development of the HealthData@EU pilots project, it is expected to have an EU-wide web portal, where data users may inquire the EU Datasets catalogue, but not if there will be equivalent applications for national

level portals, and other dedicated portals, and, if so, if those portals will be able to inquire both the national datasets catalogues and the EU Datasets Catalogue. This situation leads to different scenarios described in the following table.

Table *2* contains the analysis of the possible scenarios presented for the data search services.

Regarding the third point mentioned at the beginning of the section, regarding how the information contained is, this is a discussion that resides in the semantic interoperability area, and thus it has been covered in WP6 activities. In deliverable 6.2[10] two recommendations on this topic were issued:

- RECOMMENDATION 1: In HealthData@EU, data discoverability may benefit from the combined use of generic standards and domain-specific standards.
- RECOMMENDATION 2: This combined use may on the side of data preparatory bodies require the implementation of a two-step process supporting the phase of data discovery; a) a first step focusing on gathering high-level knowledge on the data sets available that is agnostic to the domain or the type of data; and, b) a second step where the focus is the actual content of the data source, that would be domain-data type-specific.

In the HealthData@EU pilot project, it is expected to use the "DCAT Application Profile for data portals in Europe"[11] (DCAT-AP), promoted by the EC. This standard, based on the Data Catalogue Vocabulary[12] (DCAT) developed by the W3C, will be extended to cover the health particularities, following the WP6 recommendations. Depending on this extension, the second scenario introduced the Table make be more realistic, for example, if the national datasets catalogue retain a grade of metadata deeper than the one exposed in the EU Datasets catalogue, this will also open the possibility of perform more much sophisticated searches, for example those based on metadata summaries at variable level, as the ones offered in the Atlas[13] tool provided by OHDSI as part of the OMOP ecosystem,

4.1.3 Study feasibility analysis services

The study feasibility analysis services, in conjunction with the "Support and Training Services" (see Section 4.5.3), are purely consultancy services that will be provided by health data experts. Its purpose is to validate the data users' necessities to carry on their

[10] TEHDAS deliverable 6.2 "*Recommendations to enhance interoperability within HealthData@EU*" https://tehdas.eu/app/uploads/2022/12/tehdas-recommendations-to-enhance-interoperability-within-healthdata-at-eu.pdf

[11] "*DCAT Application Profile for data portals in Europe*" https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe

[12] Data Catalog Vocabulary (DCAT) - Version 2. W3C Recommendation (04 February 2020) https://www.w3.org/TR/vocab-dcat-2/

[13] OHDSI Atlas Wiki https://github.com/OHDSI/Atlas/wiki

projects considering the particularities of the data sets found in the data holders using the data search services.

The provision of this service is based on the availability of such experts on the data, for this reason, it is expected the provision of the service to be as close as possible to the data holders.

Table *3*: Possible scenarios of the study feasibility analysis services

Table 3: Possible scenarios of the study feasibility analysis services

Scenarios	Pros	Cons
Data experts reside at data holder level.	Highest knowledge of data available	Difficult on the consultancy operations management
Data experts reside at HDAB level	Aggregation of "national" or "thematic" data knowledge, depending on the HDAB deployment	Some decoupling with datasets knowledge
Data experts reside at EU level	Single point of contact for all datasets, easier management of petitions.	High decoupling with actual datasets' particularities.

It is important to note that, considering the varied typologies of health data that will be covered in the HealthData@EU infrastructure (see Art. 33 of the EHDS proposal), the data knowledge required to judge the feasibility of a given project may be distributed among different communities. This may imply two different situations: a minimum effort on evaluating the feasibility is provided and this responsibility will reside in the data user; or, alternatively, there will be a heavy coordination work among data experts that may become a stopper to serve data users.

4.2 Data permit application phase

According to the legislative proposal (Art. 45), any natural or legal person may submit a data access application. Data users seeking access to electronic health data from more than one Member State shall submit a single application to one of the concerned health data access bodies of their choice which shall be responsible for sharing the request with other health data access bodies and authorised participants in HealthData@EU.

The legislative proposal also differentiates the requests to have access to individual level data, called the data access requests, and the requests to have access to aggregated anonymized data, called the data requests. Even though the output provided by the HDAB on both processes are different, a data permit in the first case instead of aggregated data in the second case, the technical processes for requesting and approving the requests may be the same for both use cases. For simplicity reasons, the following analysis will focus on the data access request process but it is also applicable to the data request process.

For this purpose, the data permit application process is ensured by two subsystems. On one hand, the data permit request services focus on the data users need, by allowing him to apply and manage his applications. On the other hand, the data permit grant services target the approvers from HDABs by allowing them to review and validate or not the applications.

The following sections describe the scenarios for hosting each of those two subsystems and then describe the possible global solution, mixing all possible scenarios for request and grant services.

4.2.1 Data permit request services

The Data permit request services would allow the Data User to submit a data access request, check the status, review and complete it and check his history of submissions.

The first approach is to have a centralised portal where all European data users can request access to data located in any member state participating in the EHDS. This approach eases the implementation of the system by avoiding replication per member state and simplifying the integration of key components such as the authentication of data users. It also provides a single-entry point for the data users, where he can review all his applications. The main downside of this approach is that it may require a complex migration from current application portals existing today in some member states.

A distributed approach to build the data permit request services would consist in having a single instance of the portal per HDAB. This would allow each member state to retain control of the system and offer some variations regarding which HDABs the data user chooses to start its process from (for instance to allow a better control of local / national users). The main downside of this approach is the complexity to maintain this system over time in order to ensure consistency between the different instances of the system.

Table *4*: Possible scenarios for the data permit request services

Table 4: Possible scenarios for the data permit request services

Scenarios	Pros	Cons
Centralised	No replication of system per HDAB.	Complex migration from current application management systems.
Distributed	Each HDAB retains control of the system.	Complex maintenance to ensure consistency.

4.2.2 Data permit grant services

The Data permit grant services allows the HDAB to check pending data access requests for data in his scope of responsibility, accept or reject an application or data request, ask revision about the submission and check history of submissions.

The first approach is to have a centralised portal where approvers from all HDABs can have access to the applications on their data. This implementation eases the management of approvals for requests on data under different scope of responsibility because all HDABs manage their applications on the same system. The main downside of this approach is that the approval process should be the same for all HDABs and cannot be adapted for specific needs, such as for instance managing validation loops within the member state involving different actors (data holders for instance).

A distributed approach to build the data permit grant services would consist in having a single instance of application management system per HDAB. This would allow each member state to retain control of the system and offer some variations on the validation process. The main downside of this approach is the complexity in management of approvals for requests on data under different scope of responsibility.

Table *5*: Possible scenarios for the data permit grant services

Table 5: Possible scenarios for the data permit grant services

Scenarios	**Pros**	**Cons**
Centralised	Easier management of multi-country approvals/requests for revision	No customization based on specific needs for approval
Distributed	Possible customization in the approval process per HDAB	Complex management of multi-country approvals/requests for revision

Table 6: Possible scenarios for the interaction between data permit request systems and data permit grant systems

Scenarios	**Pros**	**Cons**
Centralised	No replication of system per HDAB. Easier management of multi-country approvals/requests for revision.	Complex migration from current application management systems. No customization based on specific needs for approval.
Distributed	Each HDAB retains control of the system. Possible customization in the approval process per HDAB	Complex maintenance to ensure consistency. Complex management of multi-country approvals/requests for revision
Hybrid with distributed requests and centralised grant services	Each HDAB retains control of the system. Easier management of multi-country	Complex maintenance to ensure consistency.

	approvals/requests for revision.	No customization based on specific needs for approval.
Hybrid with centralised requests and distributed grant services	No replication of system per HDAB. Possible customization in the approval process per HDAB	Complex migration from current application management systems. Complex maintenance to ensure consistency.

4.2.3 Interactions between Data permit request services and Data permit grant services

The different scenarios to build the full system for data permit application phase is a combination of all approaches for the subsystems for Data permit request services and Data permit grant services.

The first scenario is to have a single centralised system, where both Data Permit request services and Data permit grant services are deployed in a single place. This system will be hosted and operated by the European Commission.

The second scenario is to have a fully distributed system where both data permit request services and Data permit grant services are instantiated once per member state and communicate to each other through a peer-to-peer communication system.

The third scenario is to have a hybrid system where data permit requests services are instantiated once per member state, but the Data permit grant services are deployed in a single place.

The fourth scenario is to have a hybrid system where data Permit requests services are deployed in a single place, but the Data permit grant services are instantiated once per member state.

Table *6*: Possible scenarios for the interaction between data permit request systems and data permit grant systems.

4.3 Data use

The data use phase is the one where the data user will manipulate the data to perform the analyses he or she required, using the data he or she has been granted to. In this phase, there the data use phase finishes when the data user has finished its research project or has found the evidence to support new or existing policies or regulations. The finalisation of the data analysis phase may be also subject to contractual arrangements stated in the permit, for example, limiting the amount of time a data user has access to the data.

In this case, the work done around the guidelines for secure processing environments (SPEs), that will be part of the last deliverable, influenced the organisation of the different

services described in Deliverable 7.1 that conform the data use phase. This work consisted first in a survey circulated to a wide number of operators of infrastructures for sensitive data processing, equivalent to secure processing environments defined in the Data Governance Act. The second activity was a dedicated workshop with the work package advisory group (WPAG) focusing on different areas of the SPE operation.

4.3.1 Data integration services

The data integration refers to the process to transform the data to make it usable to the data user. The transformations are specifically the harmonisation of the datasets, in terms of the formats used to codify the contents, to have a common understanding of the information contained even when it comes from multiple data holders and are expected to be covered by the implementing acts of Art.58 or the EHDS regulation. It is not clear in the regulation the particularities of the dataset's linkability, i.e., how to univocally refer to information from the same citizen scattered across different datasets. To guarantee the data linkage across datasets scatter in multiple holders/MSs it might be necessary the inclusion of solutions to provide unique identifiers to subjects or directory services, that store the translation of the subjects' IDs used in different datasets. The linkability is an issue that will require a dedicated effort, as in the current context, data stored from different domains in different holders tend to use different solutions to identify subjects, in some cases being impossible to recover IDs (no reversible pseudonyms or anonymised data) to permit the linkage with other datasets.

The data integration have a main driver that operates at semantic level, and thus this is why it is being addressed in work package 6 activities, specifically those related to Data Quality Framework and the guidelines for "minimum specification for datasets exchange" (to inform the implementing acts described in the Art.58 of the EHDS regulation), and the guidelines for data quality and utility label (to inform the implementing acts covered in Art.56(5) of the EHDS regulation).

In any case, independently of the specific semantic contents to be integrated, the service deployment may be located at different locations in the HealthData@EU infrastructure, as detailed in the Table *7*.

Table 7: Possible scenarios for the data integration services

Scenarios	Pros	Cons
Integration of datasets at data holder level	Integration is done in the "primary container" of the data, closer to the expert of the data particularities.	Extra burden on the data holders, probably non-related to their day-to-day business. Extra technical solutions are required to provide external datasets linkability.
Integration of datasets at HDAB level	Leverage burden to data holders, but the expertise	May result in an unscalable approach.

	on data particularities is still closer.	Extra technical solutions are required to provide external datasets linkability.
Integration of datasets at Core Platform level	All transformation burden is delegated to a central point, with a unified view of all datasets. Potentially an easier linkability across datasets. May validate also possible reidentification situations where large amounts of data are provided.	May result in an unscalable approach. Data expertise is lost.
No integration of datasets, just minimisation of the variables provided	Data users may perform the harmonisation processes that fit the best for their analysis purposes. Potentially an easier linkability across datasets.	Huge burden to the data user. Possible re identification risk when providing large volumes of data. Note: that has been the traditional way of providing data to users.

The analysis presented in the table is similar to the one presented in the study feasibility analysis services, as it assumes that, the closer to the data holder, the better way to manipulate the data, at the cost of incurring an extra burden to their day-to-day operation.

Please note, that in the scenarios that there is an active integration/harmonisation process, it is done in the data holder / HDAB / Core Platform level, before its deposition in the secure processing environment placement. Only the fourth scenario considers an ad-hoc harmonisation done by data users, usually as part of their initial data cleanse work, that will be performed within the secure processing environment premises.

4.3.2 Data provision services

This section is limited to the scenario when data is deposited from the data holder to SPE. It is however important to remember that the SPE can also be the data holder's own environment. In the Deliverable 7.1 when describing the data provision, it was also foreseen a possible download of aggregated data from data holders to data users' premises. This direct download of personal data (usually pseudonymised) is a scenario still in place in some settings that should be deprecated.

The preferred approach for data transfer is harmonised data models and data retrieval via API from the data holder to SPE. Such a machine-to-machine approach will be able to increase both security and efficiency compared to a process that includes manual transfer and/or upload. There is an agreement that it is also important to focus on

achieving common platforms, technical requirements, and security features across member states.

In terms of data protection, it is possible to divide two steps to consider within the data provision: first, the use of commonly known standards for secure data transfer; second, the verification of the data once it is deposited in the SPE.

The data transport standards include using electronic signatures and strong, end-to-end encryption to protect transfer from both an integrity and confidentiality perspective. On a more detailed level encryption methods may be on transport or application layer, symmetric or asymmetric, with or without additional encryption of content. The responses from the survey mention a variety of these methods being used today. In the context of the secure data transfer, it will be relevant to consider the use of specific solutions for interoperable and secure data transport, such as eDelivery[14], the standard of choice for the HealthData@EU pilots project.

In terms of data verification, it is common to use electronic signatures and checksums to verify the integrity of the data that is transferred and has also been mentioned in relation to data protection. The verification should also include the validation of the deposited data against the uses detailed in the data permit issued in the previous step. Desirably, this last validation against data permits should be done in the most automated manner possible.

4.3.3 Data analysis services

Data analysis services refers especially to the secure processing environment services (SPE), the technological solutions where the EHDS legislative proposal obliges the data users to process the data they have been granted access to (Article 50). In this way, the SPE services are used after the data permit application has been approved.

Data Governance Act DGA gives the definition of secure processing environment. The EHDS regulation refers to the DGA definition in its Definitions Article, i.e. uses the same definition.

> *'secure processing environment' means the physical or virtual environment and organisational means to ensure compliance with Union law, such as Regulation (EU) 2016/679, in particular with regard to data subjects' rights, intellectual property rights, and commercial and statistical confidentiality, integrity and accessibility, as well as with applicable national law, and to allow the entity providing the secure processing environment to determine and supervise all data processing actions, including the display, storage, download and export of data and the calculation of derivative data through computational algorithms; (DGA, Article 2, EHDS, Article 2)*

[14] https://ec.europa.eu/digital-building-blocks/wikis/display/DIGITAL/eDelivery

Also, the term Trusted Research Environment (TRE) has been used when referring to the environment where personal data is processed for research purposes[15].

According to the policy option 2 described in the EHDS impact assessment, which implies a "Regulatory intervention with medium intensity", it will be possible to establish a decentralised model with several providers of commonly defined SPEs serving the HealthData@EU infrastructure. Common, European wide minimum requirements for SPEs will be highly important for successful EHDS implementation as it increases trust between actors to share data across borders. These requirements will be detailed in the implementing acts regulated under the Article 50(4).

In any case, several general requirements for SPEs have been defined in Article 50 in the EHDS proposal, serving as a basis and a minimum related to guidelines and further requirement specifications regulated under Art.50(4). The Table *8* present the comments and considerations to be made for each requirement in such work.

Table 8: Comments and considerations about the Article 50 of the EHDS legislative proposal

Text in EHDS proposal	**Comments and considerations**
1. The health data access bodies shall provide access to electronic health data only through a secure processing environment, with technical and organisational measures and security and interoperability requirements. In particular, they shall take the following security measures:	Further guidance will be needed. It is recommended to consider existing related frameworks, requirement sets and guidelines before determining if anything further needs to be developed. It is important to ensure requirements and guidance are on an appropriate level that will work in practice.
(a) restrict access to the secure processing environment to authorised persons listed in the respective data permit;	Detailed enough to work as a specific requirement related to access management. Such requirements may be implemented using both technical and organisational measures, although automation is often preferred.

[15] Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems. UK Health Data Research Alliance; NHSX; https://doi.org/10.5281/zenodo.5767586

(b) minimise the risk of the unauthorised reading, copying, modification or removal of electronic health data hosted in the secure processing environment through state-of-the-art technological means;	This requirement is very broad and needs further guidance. It is recommended to consider existing security related frameworks, requirement sets and guidelines before determining if anything further needs to be developed. The Guideline "State of the art" performed by TeleTrust in cooperation with ENISA may be of interest[16]. A summary of security related topics that have been discussed in workshops and the survey to SPEs can be found related to “Security” further down in this section.
(c) limit the input of electronic health data and the inspection, modification or deletion of electronic health data hosted in the secure processing environment to a limited number of authorised identifiable individuals;	Considerations related to this requirement are discussed related to “Upload of data user’s own content” further down in this section.
(d) ensure that data users have access only to the electronic health data covered by their data permit, by means of individual and unique user identities and confidential access modes only;	Detailed enough to work as a specific requirement related to access management. Such requirements may be implemented using both technical and organisational measures, although automation is often preferred. May consider providing some additional guidance on practical implementation.
(e) keep identifiable logs of access to the secure processing environment for the period of time necessary to verify and audit all processing operations in that environment;	Detailed enough to work as a specific requirement related to logging and monitoring. May be beneficial to provide some additional guidance on what to log and retention times.
(f) ensure compliance and monitor the security measures referred to in this Article to mitigate potential security threats.	This requirement is very broad and needs further guidance. There are several frameworks and standards when it comes to security governance and management. ISO27001 is one example that is mentioned related to “Security” further down in this section as a standard that is used by many. It may also be relevant to discuss the connection between this requirement and requirement 3.

[16] “State of the art on IT” – Guidelines by ENISA and TeleTrusT https://www.teletrust.de/en/publikationen/broschueren/state-of-the-art-in-it-security/

2. The health data access bodies shall ensure that electronic health data can be uploaded by data holders and can be accessed by the data user in a secure processing environment. The data users shall only be able to download non-personal electronic health data from the secure processing environment.	Requirements related to secure data transport from data holder to SPE will need further guidance and considerations are discussed in the previous section "4.3.2 Data provision services". Requirements related to restrictions in downloading personal data from the SPE will need further guidance and considerations are discussed related to "Privacy techniques" and "Data extract control" further down in this section.
3. The health data access bodies shall ensure regular audits of the secure processing environments.	Considerations related to this requirement are discussed related to "Verification and certification" further down in this section. Some components that may be worth considering is for instance: Development of European cybersecurity certification schemes that is mentioned for instance in Article 49 of Regulation (EU) 2019/881[17] (Cybersecurity Act) and Article 24 Directive (EU) 2022/2555[18] (NIS2) Cloud Infrastructure Service Providers Europe Code of Conduct for cloud infrastructure service providers[19], an effort approved by the CNIL, the French independent authority that veils for security and privacy of personal data.
4. The Commission shall, by means of implementing acts, provide for the technical, information security and interoperability requirements for the secure processing environments. Those implementing acts shall be adopted in accordance with the advisory procedure referred to in Article 68(2).	It will be very important to ensure that the development of SPE guidance is synchronised with the SPE requirements developed by the Commission.

The following subsections provide the TEHDAS WP7 views on the requirements for SPE's based on the work carried out in TEHDAS WP7 (advisory board workshops, SPE surveys to current SPE-like operators) and available public materials (especially the EHDS legislative proposal). There has been rather strong consensus on the general approach for SPE's and key requirements. For example, the approach of enabling several SPE's per country is largely supported. At the same time, there are still several

[17] Regulation (EU) 2019/881 (Cybersecurity Act) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R0881&from=EN

[18] Directive (EU) 2022/2555 (NIS2) https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2555&from=EN

[19] Data Protection Code of Conduct for Cloud Infrastructure Services Providers - CISPE https://www.codeofconduct.cloud/the-code/

details under discussion. In those cases, options or alternative requirements are presented to be further elaborated in future work.

General considerations

In terms of the GDPR roles, the EHDS legislative proposal defines the HDAB and the data user to be joint controllers of the data in the scope of the data permit application. The proposal also outlines that HDABs shall provide access to electronic health data only through an SPE (Art.50(1)). The SPE may be provided by the HDAB itself or the HDAB may use an external SPE service provider. With respect to the GDPR, the SPE service provider will be the data processor for the joint controllers. Figure *10* provides a schematic view of the interconnection within the actors that interact with a SPE.

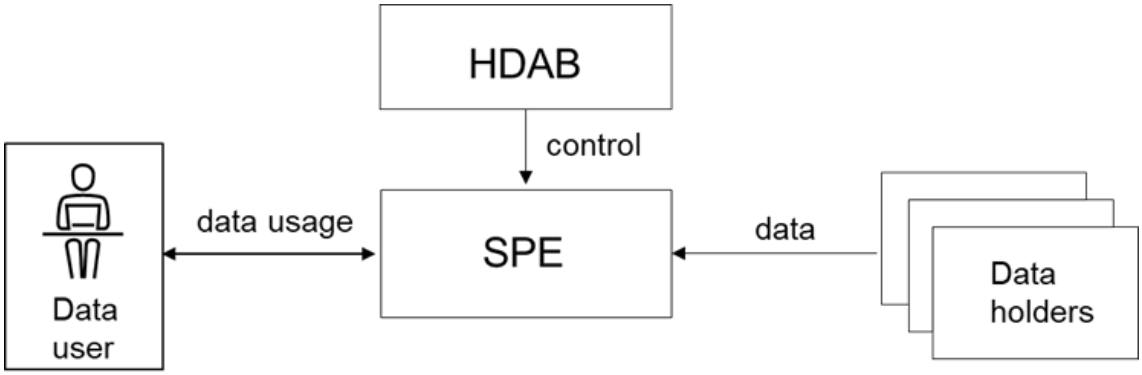

Figure 10: Access to data provided via secure processing environment (SPE)

The proposal does not specify, which kind of organisations can be SPE service providers and where the SPE service providers should be located. Most experts support the approach that there can be multiple SPE services per country and that both public organisations and private companies can provide SPE services. This approach is considered to be beneficial as it helps to maintain sufficient availability of computing services and to fulfil different types of needs of data users. Many data holders, such as university hospitals, are already providing SPE services or in the process to provide them. There are varying opinions on the need of a centralised SPE service provided by EC. We recommend keeping the centralised SPE service as an option, as regulated in Art.52(10). It may be an attractive option for those countries which do not want to set up their own SPE services.

Federated analysis has been frequently mentioned as an approach to follow the "bring questions to data instead of moving data" (also named "data-centric" computing) mentioned in the EHDS legal proposal recitals (recital 55). A following subsection analyses in detail the impact of federated analysis on SPEs.

Available analysis tools and materials

Various tools and support materials are needed in the SPE to support data processing. It is recommended that a standard set of available basic tools should be defined to be

available in the SPE by default. In general, such tools include statistical analysis software (R, Python, SPSS, …), basic office tools and data/software management tools (version control tools, database software)[20]. Additionally, it should be possible for the data user to order specific tools to be installed for a project as needed. In addition, to specific software packages, it is also desirable to permit the deployment of containerized software, to ease the management of the tools environments, a common issue when using scientific tools, that also eases the reproducibility of the results.

Support materials, such as basic terminologies, clinical codes (ICD10, SNOMED-CT, ATC, …) as well as genetic tables are needed. These needs vary considerably between projects, and therefore customisation for individual projects are expected to be needed.

As part of the data permit requested information, it is expected to define the data management plan within the SPE premises. A possible option would be to differentiate the input data location, one location for temporary files, and a third location for finalised results. This differentiation may facilitate the operation of the SPE, for example in terms of backup or the encryption of certain locations of the file systems.

A need for centralised maintenance of information about recommended tools and support materials were identified in the discussions. A centralised register would help the distributed SPE's to be aligned in terms of tools and support materials usage. The same register could also maintain information about security assessments, approvals and certifications of tools. This helps to avoid overlapping assessment and evaluation work in different countries and SPE providers.

Upload of data user's own content

In addition to the standard statistical software available in the environment, users might need other software applications or libraries, programs or pre-trained models to analyse their data. The users might also want to upload their own data, such as survey data or data from a different domain, if possible, linkable with the HealthData@EU provided data.

Most experts agree that users should indeed be able to upload their own content to the SPE. The trustworthiness of user-originated content can be ensured by using for example a quarantine/staging environment to scan for malware before the content is uploaded to the SPE. Other methods suggested by the experts include manual inspection or automatic (AI based) scanning. However, as the SPE is an isolated environment, the risk posed by insecure software or scripts is limited, and therefore it is important to carefully assess risk impact versus resources needed for a thorough inspection of all user-originated content.

It is important to highlight that, it is a well-defined security risk that when combining personal data from a high number of sources, or linking with semi-public registries it

[20] The list of software of Kapseli, Findata's SPE is available here https://findata.fi/en/kapseli/#software

might be possible to re-identify individuals present on de-identified data[21]. To avoid this situation, it is desirable to have a strong framework of well-known agreements, guidelines and legal penalties in place, as the one regulated by the Art.43 of the EHDS legislative proposal.

The majority of the examined SPEs avoid allowing users to import their own data or software. For those cases where the import is made possible, prior approval and audit by the service provider is usually required.

Federated analysis

Federated analysis refers to approaches where data is processed in multiple distributed locations and final results are obtained by combining these partial results of the distributed computations. The federated analysis approach would enable it to keep data in the original country and even in the original organisation (or data holder), and thereby it would be aligned with the recommendation to "bring questions to data instead of moving data whenever possible" expressed in the EHDS legal proposal recital 55. Despite this recommendation, it is widely understood that the federated approach is not feasible in all cross-border use cases (in particular those related to rare diseases) and the HealthData@EU infrastructure will support cross-border data transfers and pooling data to designated SPEs.

Following from the legal proposal (Art.50) the data shall always be processed in an SPE. This applies also to federated analysis so that the distributed computations shall be executed in an SPE. The following section elaborates on the impact of federated analysis approach on the required SPE characteristics.

Figure *11* shows a simplified architecture for federated analysis with data sources in two SPEs. If the SPEs are in different countries and their data comes from national sources, this setting enables operation without cross-border data transfers. Note that, as later discussed, the orchestration of the SPEs may be also executed in one of the SPEs assigned to data user to perform the federated analysis.

[21] Dankar, F.K., El Emam, K., Neisa, A. et al. Estimating the re-identification risk of clinical data sets. *BMC Med Inform Decis Mak* 12, 66 (2012). https://doi.org/10.1186/1472-6947-12-66

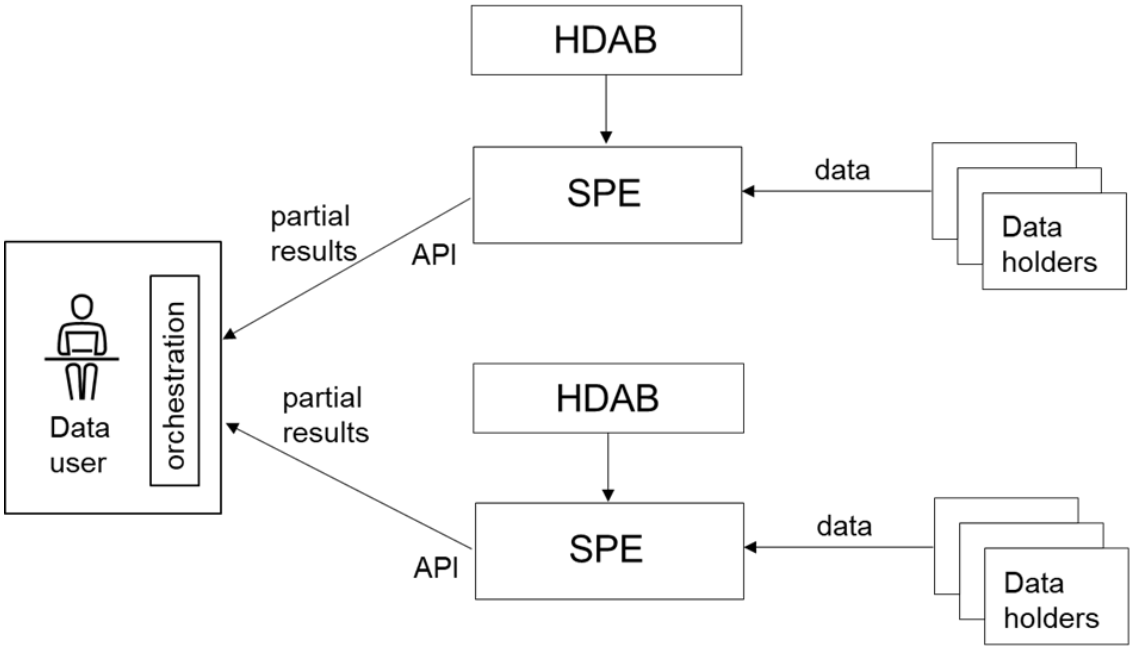

Figure 11: Simplified federated analysis architecture

The following specific characteristics are required by an SPE to support federated analysis as outlined by Figure *11*:

- **API access support**. The execution of the federated analysis and combination of the partial results is performed by orchestration software running at the data user. This approach is most feasible if the SPE exposes an open API interface which enables these tasks to be done automatically. Manual execution of the tasks would be possible, but laborious.
- **Privacy of partial results.** The results retrieved from the SPEs shall be anonymous in the same way as in the case of conventional processing. The federated algorithms shall ensure that all outputs from the SPE to the data user are anonymous and do not leak personal information. Most preferably, the anonymity of results would be ensured automatically. One approach is to allow only approved scripts with known output types to be executed.
- **Use of a common data model**. It is highly desirable that the distributed computations can be executed with identical software agents. This implies that the same data and same data model is present in all involved SPEs. For example, the OMOP model is already widely used and is, therefore, a strong candidate for the HealthData@EU infrastructure. A common data model is also highly beneficial in the case of conventional processing as it allows analysis tools to be reused across SPEs.
- **Orchestration layer (option)**. In Figure *11*, it is assumed that orchestration of the federated analysis is carried out by a software component executed in the data user's environment. Also, a separate orchestration layer between data users and the SPEs has been proposed in order to simplify the processing from the data user's perspective. Such an orchestration layer could be set up and

maintained by a trusted partner, such as an HDABs or the EC. Further investigation concerning the feasibility of such an approach would be needed.

There will also be other considerations to be made related to privacy and security in a federated model. Such have for instance been addressed by the Norwegian Data Protection Authority through a project in their Artificial intelligence sandbox environment[22]. The project concerned data in the finance industry, but the considerations and conclusions can be transferred to the use of federated analysis on health data. The main conclusions are related to:

- **Processing responsibility**: In the project case the conclusion is that the owner of the data repository will most likely be the data controller. The provider of the algorithm will likely be the data processor and responsible for ensuring that vulnerabilities in the AI model does not lead to that the model contains personal data.
- **Data minimization**: It may be difficult to determine how much data is needed for the AI model to be efficient. The recommendation is to wait to collect data until it is certain that it will be useful for the model.
- **Security challenges**: It is considered positive that federated learning reduces the need to share data. It is however mentioned that this is a relatively new model which means it may have some unknown vulnerabilities. Model inversion attacks, with the intent to reconstruct personal data based on access to trained models, is mentioned as a potential threat. The risk for such attacks is considered low, but is also difficult to assess.

Security

Several security related requirements have been defined in Article 50 of the EHDS proposal. These should be a basis and a minimum related to guidelines and further requirement specifications. There are also several existing security frameworks, requirement sets and guidelines that should be considered before determining if anything further needs to be developed.

Below is a summary of security related topics that have been discussed in workshops and the survey and should be considered when it comes to requirements and guidelines for data analysis services. It does not include security topics that are covered in other sections in this document, such as secure transfer of data, control of digital material uploaded by users and privacy, including data extract control.

- **Security frameworks**: The survey reveals that many respondents have institutional security policies and operational documentation in place. It also highlights several respondents being certified or looking to become certified. The most common certification is ISO27001.

[22] Finterai, final report: Machine learning without data sharing (NO) https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/finterai-sluttrapport

- **Access**: The survey sent to existing SPEs examined how they identify and authenticate the users and how they manage different users' permission to access the data. The majority use strong authentication to safely and reliably confirm user identity, and support multi-factor authentication for federated IDs. Only users identified in the data permit should be granted access to the SPE. Many respondents automatically lock access rights after the data permit has expired. Regular checks that the access is still valid and appropriate are either done by the respondents or the accountable for the project.
- **Isolation of environments between projects**: Since researchers may work in several parallel projects, and may have been granted access to sensitive data from different cohorts, it needs to be ensured that the researchers have no permission (and even no possibility?) to merge the data from different projects, unless that has been presented in an approved data access application. To enforce this, data access rights should not be linked to a person and their affiliation, but to a project. If the researcher has multiple data permits, they need to decide what data permit they are going to access at that time. According to the survey results, most infrastructures report that each permit corresponds to a single research project, and that each project has a dedicated environment, which is technically and logically isolated from other environments. Moving or sharing any data between the environments is not possible.
- **Logging and monitoring**: The survey respondents typically monitor data usage and user actions and store logs in a secure and separated IT-environment with limited access.
- **Vulnerability management and security testing**: Among the respondents of the survey there are generally routines for regular vulnerability scans and also independent penetration tests by professional third parties.
- **Data retention**: In relation to termination of use of the environment there are some variations on how long the data is stored in the environment. However, the storage period is often related to what is stated in the data permit and 6 months after. Some refer to that this is the responsibility of the data controller.
- **Disaster recovery**: Most of the survey respondents confirm that they have a disaster recovery plan.
- **Employee obligations and security training**: The common practice among survey respondents is that employees are bound by confidentiality agreements or similar. The respondents also generally provide regular training of staff.

Privacy techniques

In Article 44 of the EHDS regulation proposal it is laid down that the health data must be provided in an anonymised format "[...] where the purpose of processing by the data user can be achieved with such data [...]". Whilst the term "pseudonymisation" is clearly defined in Article 4 of the GDPR, a legal definition of "anonymisation" at EU level is lacking to date. Therefore, it will be necessary to discuss, evaluate and harmonise different privacy preserving methods among the member states. For this purpose, the current state of scientific knowledge should be taken into consideration in order to provide for the best possible reduction of re-identification risks while maintaining the

usability of the data for the respective research purposes. It is important to keep in mind that a "complete" anonymisation that entirely prevents any re-identification can frequently not be achieved. In addition to minimizing the exposure of personal information to be processed in the SPE, privacy techniques are also relevant for ensuring the privacy of the analysis results output from the SPE. Protecting privacy of the analysis results is elaborated in section 4.4.

A well-known concept for enhancing data privacy is k-anonymity[23]. This concept accounts for the fact that even after removing identifiers such as names, addresses etc. an identification of individuals is still possible by combining other distinctive variables called "quasi-identifiers" to unique patterns that, in particular in combination with other sources of information, make a person identifiable. K-anonymity has been described as follows: "A table provides k-anonymity if attempts to link explicitly identifying information to its contents ambiguously map the information to at least k entities" Generalisation and suppression are possibilities to enforce k-anonymity. Another approach for protecting privacy is the generation of artificial data based on an original dataset. The so-called synthetic data ideally maintains the relevant statistical characteristics of the original. In Denmark, a study on the use of synthetic data ("'Vision for better use of Danish Health Data'") is being performed[24]. The German Health Data Lab is currently conducting a study aiming to compare classical anonymisation methods such as k-anonymity with synthetic data in terms of utility and the remaining risk of disclosure[25]. As both institutions participate in TEHDAS and the HealthData@EU pilot project, the two studies may provide valuable contributions to finding suitable privacy preserving methods for the HealthData@EU.

With regard to the data preparation and provision workflow, it needs to be recalled that, as introduced in the Data integration services section, a patient-level record linkage cannot be performed after anonymisation of the data. In cases where a patient-level record linkage is required, this needs to be completed before data anonymisation.

As a final mention, indicate that work package 6 of TEHDAS oversees addressing the issue of de-identification by developing data minimization and data de-identification guidelines, a more in-depth evaluation of this topic will be elaborated for their final deliverable 6.3.

Verification and certification

Security requirements aim to ensure that the SPE service provider has sufficient security arrangements to prevent unlawful disclosure of personal information. Due to the complexities of data management, many of the existing SPE service providers have decided to pursue an ISO accreditation. For instance, ISO/IEC 27001 demonstrates that

[23] Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression"

[24] TEHDAS country visits factsheets (Denmark) https://tehdas.eu/packages/package-4-outreach-engagement-and-sustainability/tehdas-country-visits

[25] "Research meets data protection: Analysing synthetic health data using artificial intelligence" https://www.bfarm.de/EN/News/Blog/_docs/2022-03-10-forschungsdatenzentrum.html

the organisation has implemented an effective information security management system, and taken steps to protect data in the event of a breach. This certification is known to primarily verify the design of controls, but it does not verify the effectiveness of controls, i.e. that the information security management system actually works as described.

Harmonisation of SPE security requirements will be extremely important for the EHDS. Existing models can be followed. For example, Findata, has published a regulation for SPE requirements. Each SPE where personal health data is processed for secondary use is required to be certified against these requirements[26]. The certification needs to be carried out by an accredited information security inspection body. A register of certified SPEs is maintained by Valvira (National Supervisory Authority for Welfare and Health).

A full certification process with such defined audit procedures provides a high level of verification on compliance to defined SPE requirements. Such a certification process may however demand a lot of resources to operate and may not be sustainable for all member states. With such a certification process it is also important that the responsibility of the certifying body and the data controller responsibilities according to GDPR are clarified.

Therefore, some guidance on the minimum requirements of a verification process is needed, and preferably aligned between member states. In addition to a full certification process according to the Finnish model, the following could for example be considered, either stand-alone or in combination:

- Self-assessment
- Voluntary compliance testing of SPEs performed by a certification body
- Random compliance testing of SPEs performed by a certification body
- Audit procedures, the same or similar to that of Data Protection Authorities according to GDPR

It should however be considered that the level of protection required for processing of health data may also require a high level of verification of compliance. A full certification process will also be able to provide a high level of trust to the data users. So even though it may require a lot of resources centrally, it can decrease the level of resources that needs to be used by data users to verify compliance as part of their responsibility as data controller.

Regardless of method to be used for verification the health data access body should be responsible for ensuring that there is an overview on the verification status for SPEs and on the method used.

[26] Regulation by the Health and Social Data Permit Authority: Requirements for other service providers' secure operating environments (REGULATION 1/2022 - Diary number THL/214/14.00.07/2022) https://findata.fi/wp-content/uploads/sites/13/2022/03/Regulation-Requirements-for-other-service-providers-secure-operating-environments.pdf

4.4 Project finalisation phase

The project finalisation phase gathers the services related to disclose the findings obtained while analysing the data (*use* the data). In this phase, part of the services are expected to be provided also in a secure processing environment, as per the guarantee that the possible data transfers *outside* the environment are just for those authorised datasets or variables. These services include the assistance elements on how to prepare these results. The document also analyses the services the requirement on accessing the original datasets, partially or anonymised, so as to guarantee the reproducibility of the results in research context (or others).

4.4.1 Results validation and archival services

The EHDS proposal states that data users shall only be able to download non-personal electronic health data from the secure processing environment. The SPE survey and workshop discussions show that there are some technical measures that are currently used to prevent this, but they will need to be complemented with organisational measures to provide sufficient protection.

The technical measures include the operation in a virtual desktop, disabling of cut and paste, monitoring and control criteria such as size, type of exported data or minimum count within a cell. Complementary organisational measures mainly include manual check/verification of exports to prevent possible disclosure of personal data, either through quarantine prior to release or by retrospective follow up. The responsibility of such controls varies and can be performed by the project manager or the SPE provider, sometimes using a 4-eyes principle. The level of detail also varies and can be done for all exports or only for random samples or for samples that fall within defined criteria such as size, type or minimum count.

Complementary to validation or auditing relates to disclosing the analysis results, there is another requirement of providing access to the original data for its possible validation and reproducibility of scientific publications. In that case, it will be necessary to provide access to the original data, or a subset of the data, in some manner. The access to this data is partially related to the data retention.

In all cases, this process of validation and archival to exemplar data for reproducibility is related to the SPE where the analyses were performed. In the case of the federated analysis, the analysis design should guarantee that the validation of the partial results, if the orchestration occurs outside the SPE.

In the following tables there is an overview of the foreseen scenarios of these two elements.

Table 9: Possible scenarios for the services for results validation for disclose

Scenarios	Pros	Cons
Results disclose manually operated	No false positive of the results disclosure	Non-scalable approach. Not applicable to a federated analysis.

Results disclose AI operated	Higher scalability. Possible false positives. Can be used to guarantee the validation of federated analysis	AI solutions yet to be widely tested

Table 10: Possible scenarios for accessing data for reproducibility in scientific publication context

Scenarios	Pros	Cons
Access to all the original data	Easy setting to reproduce the results	Currently not considered in the actual EHDS data access models.
Access to a subset of the original data	Subset of data can be easily controlled	
Disclose anonymised version of the original input datasets	Easy setting to disclose the original data	May prevent to actual reproducibility
Generation of synthetic data similar with same patterns as original data	Equivalent to anonymised data, with higher level of security.	Technologies to generate fully equiparable synthetic data yet to be widely tested.

Last but not least, there will also be a need to provide the clear channels between data users and HDABs to notify possible incidental findings that might affect the health of data subjects of the analysed datasets, as detailed in Art.46(12). Similarly, but technically more challenging, there should also be clear channels to provide feedback both when providing possible enrichment to original datasets quality, as presented in recital 39.

4.4.2 Results output preparation services

The preparation of the results for its output consists of a series of resources to transform the results in the format required for a possible further cataloguing and archival in external repositories, such as Zenodo, EU open data portal, EOSC or the European Health Information Portal. To aid in this publication, materials regarding the FAIRification process, e.g., metadata standards for cataloguing, appropriate ontologies to codify the data, should be made available. This requirement is partially aligned with the requirements for the available analysis tools and materials to be included in the SPEs. As per the interaction of the archival services described above, it would be possible to provide data users with anonymisation toolkits to prepare their output data (not only the input data) and disclose their results.

This process is also subject to the support and training services

4.5 Transversal services

The transversal services were identified in Deliverable 7.1 as a set of services that do not provide a specific feature associated with the effective data management, but are necessary for the proper functioning of the HealthData@EU infrastructure.

4.5.1 Node Management Services

The node management services cover the services required to evaluate the proper functioning of the nodes that participate in the HealthData@EU, i.e., the National Contact Points for Secondary Use[27], and, up to some extent, with the coordinator HDABs that might expose some of the services, if specific services deployments are selected, if coordinator HDABs are harvested by the EU Core Platform to consolidate the EU Dataset Catalogue joining every single national datasets catalogue.

Node management services will comprise a set of auditing elements to guarantee the availability, integrity and security of such nodes. For this auditing purposes, it will be necessary to define the acceptance criteria, as part of the technical description of the architecture.

The possible scenarios foreseen for these services, listed in the following table, are related to who is responsible to carry out the auditing processes.

Table 11: Possible scenarios for the node management systems

Scenarios	Pros	Cons
Auto reported node auditing	No extra burden on Core Platform	High trust requirements to the NCPs. Extra burden on NCPs.
External auditing by Core Platform	No burden on NCPs to perform the auditing.	Only external auditing expected, e.g., only intrusion tests.
Internal auditing led by nodes combined with external auditing led by Core Platform.	Balanced responsibilities between elements Core Platform	Higher coordination required to perform the audits.

4.5.2 Authentication and Authorisation Infrastructure (AAI) services

The authentication and authorisation infrastructure services play a crucial role to ease the user experience across the overall infrastructure. As detailed along the present document and in previous ones, and from other work packages produced in TEHDAS and other projects, the operation of the HealthData@EU will suppose a complex interaction between multiple actors and technological systems that will interoperate to provide a series of services with the aim of easing the access to health data for its secondary use.

For these reasons, minimising the complexity of the user management across all the possible systems is key both for the seamless integration of the Users' Journey processes, and to guarantee security of the processes themselves. Having a robust AAI system will serve to orchestrate all the Users' Journey phases, giving a sense of

[27]At this point, it is not clear the involvement of other Authorised Participants defined in Art.52 of the EHDS legislative proposal.

continuity and uniformity of all the services, without requiring the users to have multiple credentials on each component.

To provide such service, two possible scenarios are foreseen, described in the Table *12*. No other scenarios have been put in place as it might be a security issue the inclusion of AAI systems operated by third parties. In that scenario, it would be preferable that these AAI systems are first coordinated by MSs and then with the EU Core Platform.

Table 12: Possible scenarios for AAI services

Scenarios	Pros	Cons
Central AAI system maintained at the EU Core Platform	Unique identification by design, that will ease the implementation of the AAI solutions in the rest of the systems.	Single point of failure, that may have an extra burden on computational capacity and security
Federated AAI coordinated by the EU Core Platform, joining AAI systems operated at MS level.	Share responsibility between actors, easing the user management for example by using national IDs / eIDAS[28].	Extra complexity of the AAI system to guarantee the interoperability between MS systems.

4.5.3 Support & Training Services

The support services are a collection of services that cover both the technical substrate and the consultancy side, i.e., the manpower. Technically they involve the information systems for: 1) manage the inquiries about the operation of the HealthData@EU as well as the incidences derived from it actual use; and 2) teach the data users of the infrastructure to make the proper use of it, maximising the sources they are offered.

In Deliverable 7.1 there was a short list about the possible software solutions that might be put in place (ticketing systems, conferencing software or remote desktop services). Regarding the possible deployments, Table *13* lists the two scenarios foreseen for the deployment of the support system, which has high resemblance with the prestudy services.

Table 13: Possible scenarios for support and training services

Scenarios	Pros	Cons
Support services are provided at EU Core Platform	“Single-stop-shop” to access support of the infrastructure.	High burden for EU Core Platform. Possible scalability issues.
HDABs offer support to its users, which is	Higher availability of support services, closer to the users.	High coordination burden when support implies

[28] Regulation (EU) No 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2014.257.01.0073.01.ENG

coordinated support from EU Core Platform	A central knowledge hub may help to centralise common issues.	actors from more than one MS.

The equivalent analysis for the training services is not presented, as it can be easily deducted from the one referring to the support. Please note the introduction of the central knowledge hub where support issues may be shared among MSs. This concept could be also applicable to share training materials for the training services.

4.5.4 Financial Services

As defined in the Article 42 of the EHDS legislative proposal, HDABs and "single data holders" (those referred in Article 49 that can also provide access to health data) "*may charge fees for making electronic health data available for secondary use*". That implies a set of services to guarantee the fee collection, ensuring both the scenario where the data is provided in a single MS or by multiple MS (involving multiple HDABs).

This cross-border exchange of fees shall be managed with high security standards, ensuring that the invoicing system is reliable among the different actors involved. Depending on the fee system and the organisation of the infrastructure, different scenarios are possible, as the ones listed in Table *14*.

Table 14: Possible scenarios for the financial services

Scenarios	**Pros**	**Cons**
A central payment system operated at the EU Core Platform	"Single-stop-shop" for data users. Less burden to HDABs.	High burden on the Central Platform to redistribute the payments to the different HDABs
Payment is done in the HDAB where the data user accessed and the redistributed to rest of HDABs	"Single-stop-shop" at HDAB level.	Possible duplication of features among HDABs. Less burden to EU Centra Platform.
Data users should access the invoice system of each HDAB involved in their petition.	No cross-border fee exchange system required.	Excessive burden to data users.

5 Infrastructure options

This section contains the initial analysis of the possible infrastructural support to be used to perform the actual deployment of the services analysed in the document.

The section is structured in two parts, first refers to the computation infrastructure itself, i.e., the technological infrastructure to deploy the services that imply processing and

storing data, and the second is the communication infrastructure, the technological infrastructure to facilitate the data exchange between the computation infrastructure

In the Deliverable 7.2, this section will be extended by deepening in the contents of the current requirement analysis and adding an extra section to provide possible mappings and foreseen interactions with existing infrastructure providers, e.g., GEANT, EUDAT, EOSC, as well as other community-specific technological infrastructures, e.g., GDI or EUCAIM.

5.1 Computation infrastructure

In the computation infrastructure presents the analysis of the foreseen hardware to the different services analysed, structured along the three systems widely studied, i.e., the systems to manage national datasets catalogues; the systems to manage the data access requests; and the secure processing environments.

A discussion around data lakes as possible solutions to store data is also included as the last sub

5.1.1 Infrastructure for national datasets catalogues

The infrastructure requirements to deploy national datasets catalogues does imply a highly specific hardware: a regular server with medium capacity (16 computation cores, 32GiB of RAM, 1TB of disk space with backup) should be enough to guarantee the proper operation of such systems. These requirements might be extended if the queries traffic increases, especially in the scenario of where data users search the national datasets catalogue, and not only inquire the EU Datasets Catalogue. The possible requirement implies that it might be recommended to use a cloud environment to deploy them, always considering the security of the information already discussed in the Section 4.1.1.

5.1.2 Infrastructure for data access requests management systems

The foreseen infrastructure required for the data permit management systems is similar to the required for the national datasets catalogue. In this case, there should be a superior requirement for storing the data access requests information, specially to guarantee the privacy of the information provided on them, as they may contain confidential information regarding project proposals or regulatory studies.

5.1.3 Secure Processing Environments

The case of the secure processing environments poses an extra challenge in terms of the infrastructure provision, due to two main reasons: the high levels of security necessary to run such systems, and the specific particularities of the different types of analysis may have in terms of computing resources to be committed. The security of SPEs is its primary reason to exist, so the infrastructure provision will need to consider all these particularities. Regarding variability of the computing resources necessary to support the different workloads foreseen, implies that there should be a representation of different infrastructures willing to deploy such systems. For example, high-performance computing systems (HPC) are foreseen for *omics* related analysis or drug

discovery; GPU-based solutions are expected for AI-based deep learning modelling; and more basic analysis servers are in use for regular statistical inference.

It is interesting to learn from the Finnish experience to understand how the infrastructure provision has evolved in a real setting. In their initial steps, the only SPE certified to operate was Kapseli©. Kapseli© is the provided by Findata and technically operated by CSC[29], the Finnish IT centre for Science. This SPE has offered a set of tools to data users in remote desktop fashion, as covered in the section 4.3.3 “Available analysis tools and materials”. After the requirements of the research community, new SPEs have been certified to operate under the Finnish legislation, for example, SPEsior[30], a privately operated SPE. In addition, some new features are being added to Kapseli© to provide a Linux environment with access to GPUs, that will be used primarily for deep learning modelling purposes. Further scenarios, such as HPC facilities are not yet available in the Finnish SPE ecosystem.

5.1.4 Data lakes

In general terms, a data lake is a massive and centralised repository of raw data (both in structured unstructured and binary forms) for secondary use. The purpose of a Data Lake is to store, give access and process data from multiple sources, allowing entities to analyse their data to produce information. Data lakes are also used to store data for long-term archiving and backup. As such, it is a larger and more complex progression from typical data warehouse solutions – where less amounts of data are stored for operations that are typically more routine and predictable in nature.

In the context of health data, data lakes can be used to create reusable dimensions of fact tables, attributes, modelled on business semantics, that allows the analysis of data from electronic health records, medical images/imaging reports, laboratory reports or wearables. A data lake can be used to gain insights into public health administrative and clinical decision-making and health-related research (e.g., trends in health outcomes, disease risk predictions, safety/efficacy of new treatments). In function, it supports the conception and implementation of a platform which supports a self-service/on demand use: data lake users should be able to find the data sets they are looking for without direct guidance from support staff. The self-service is critical for successful data lakes, since data must not be undocumented or unusable, and should have usability for all intended users (avoid becoming a data swamp).

Data Lakes enable organisations to use computational power to process large amounts of data quickly and more efficiently. They can be used to run complex analytics (such as machine learning and data mining algorithms) and perform predictive analytics to anticipate future trends and produce better health-related insights. Additionally, Data Lakes may also be used to run real-time analytics, which enables health authorities to respond more quickly to relevant changes.

[29] Kapseli © Standard Terms of use https://findata.fi/en/kapseli/standard-terms-of-use/
[30] https://esior.fi/en/spesior/

The Data Lake can be deployed on a federated architecture, where it is feasible the non-replication of raw data, keeping it at their origin, available in a logical way for their processing within the framework of the data lake. If the ingestion or integration of the raw electronic health data from data holders and HDABs occurs in real-time in the coordinator HDAB linked to the NCP of a MS, the main components of the HealthData@EU can be run into the Data Lake.

The key components of the HealthData@EU deeply analysed in this document, namely the national dataset catalogue, the data access requests management system and the secure process environment, can be designed to be fully functional and integrated securely with a data lake infrastructure. Once the data request is authorised and managed by the data access requests management system, one option is to allow access to the data in a sandbox of the data lake, where the data user could process the data into the SPE, also installed in the infrastructure of the data lake. In theory, if all components of HealthData@EU are planned with full integration in its environment, the automatic processes could be more efficient.

In order to perform a linkage and to combine data from different sources, a data harmonisation plan, using the same standard adopted by the HealthData@EU project is required. This harmonisation becomes a prerequisite to build a data lake repository that aims to serve for both the HealthData@EU project and other national purposes, which could require data mining and analytics for decision-making and policy formulation. The data harmonisation provides data users with an option to compare data from different sources, either from different databases, health information systems or portals containing aggregated data.

Deploying a data lake relies on the guarantee of not using the data from their original sources, since it is a replication of the data from its sources. There are several advantages, namely, the ability to provision the use of many applications and users simultaneously, the versioning of datasets worked in the environment, where it could be possible to versioning the various iterations of the data processing, and the feasibility to deploy the encryption of data either on its storage or on the transport.

5.2 Communication infrastructure

The communications infrastructure refers to the hardware and software pieces devoted to the exchange of data between the computation infrastructure.

As can be seen in the Figure *7* and Figure *8*, there are several interconnected actors, and thus, technological infrastructures, that will participate in the overall HealthData@EU infrastructure, which may have different communication requirements. In general, we can simplify the communication requirements according to the Table *15*, where two dimensions are exposed: the volume of data transfers expected and the level of security in the communications.

Table 15: Characterization of the communication requirements between HealthData@EU actors

Security level		**Volume of data**		
High	**Highest**	**Small**	**High**	**Connections**
	X		X	• Data holder to SPE - Data deposition (Pseudonymised data) • SPE to Data holder - Enriched data return (Pseudonymised data)
X			X	• Data holder to SPE - Data deposition (Anonymised/Aggregated data)
	X	X		• Data user to HDAB / HDAB to HDAB / HDAB to Central Platform (Data access requests) • SPE to data user (Analysis/ Analysis results) • SPE to SPE (Federated learning) • SPE to HDAB (Incidental results)
X		X		• Data holder to HDAB / HDAB to Central Platform (Catalogues)

Going through the table, it is possible to synthesise the requirements mostly focusing on the data volumes to transfer. In this way, the communication channels between SPE and data holders should be the one that will need the highest bandwidths for transferring the requested datasets (especially when dealing for example with imaging datasets), for example using non-TCP transfers such the one offered by Aspera[31] (that relies on private protocol, similar to UDP transfers). The rest of the communication links may rely on regular TCP interconnections. In all cases, the payload messages should be signed and encrypted to guarantee the integrity and security of the data exchange, using network (Internet) layer encryption (IPSec for Virtual Private Networks[32]), or transport level encryption (SSL/TLS[33]). It is worth to remind that, as previously introduced in the report, in the context of the HealthData@EU pilot project, the solution designed for the links with

[31] https://www.ibm.com/aspera/connect/

[32] Frankel, Sheila and Suresh Krishnan. "IP Security (IPsec) and Internet Key Exchange (IKE) Document Roadmap." *IETF RFC 6071* (2011): 1-63. https://www.rfc-editor.org/rfc/rfc6071

[33] Rescorla, Eric. "The Transport Layer Security (TLS) Protocol Version 1.3". *IETF RFC 8446* (2018): 1-159. *https://www.rfc-editor.org/rfc/rfc8446*

small volume of data transfers is eDelivery, which acts as a secure documental exchange at application level, based on the AS4 protocol[34].

6 Glossary

To facilitate the understanding of the present document, as well as its transposition with existing regulations, this is the list of the terms used in this report, its definition, and the document from where it was taken.

Table 16: Glossary

Term / Acronym	Definition	Source
Anonymisation	Processing of personal data in a manner that makes it impossible to identify individuals from them.	Office of the Data Protection Ombudsman, Finland. tietosuoja.fi
Data source	Data collection or a set of linked data collections sustained by a specified organisation, which is the data holder.	Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources, EMA, 2022.
Data user		
De-identification	Process of removing the association between a set of identifying data and the data subject.	NIST Glossary
EHDS	European Health Data Space	EC
EHDS2 pilot	EI pilot project; European Health Data Space - EHDS HealthDat@EU Pilot	ehds2.eu
European Health Data Access Body (EHDAB) / Health data Access Body (HDAB)	Orchestrator intermediating the communications between all participants in the infrastructure (in the policy option 3, centralised architecture).	Impact assessment report of the EHDS reg. SWD(2022) 131 final PART 1/4

[34] *AS4 Profile of ebMS 3.0 Version 1.0*. 23 January 2013. OASIS Standard. http://docs.oasis-open.org/ebxml-msg/ebms/v3.0/profiles/AS4-profile/v1.0/os/AS4-profile-v1.0-os.html.

Term / Acronym	Definition	Source
European Interoperability Framework (EIF)	EIF gives specific guidance on how to set up interoperable digital public services.	Part of the Communication (COM(2017)134) from the European Commission.
Metadata	Set of data that describes and gives information about a dataset.	Good Practice Guide for the use of the Metadata Catalogue of Real-World Data Sources, EMA, 2022.
Metadata catalogue	Key component in a service-oriented architecture, managing shared resources. Contains metadata, and the standards make sure the information is described in a unified way.	INSPIRE, ISO
MS	Member State of the European Union	EC
National contact point for secondary use (NCP/NCP2)	“An organisational and technical gateway enabling the cross-border secondary use of electronic health data, under the responsibility of the Member States;”	EHDS proposal regulation. COM(2022) 197
Node	Synonym of National contact point for secondary use	
Pseudonymisation	Processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;	GDPR

Term / Acronym	Definition	Source
Re-use	The use by persons or legal entities of documents held by public sector bodies or public undertakings, for commercial or non-commercial purposes other than the initial purpose.	Directive on open data and the re-use of public sector information. PE/28/2019/REV/1
Secondary use	The secondary use of health and social data means that the customer and register data created during health and social service sector activities will be used for purposes other than the primary reason for which they were originally saved.	Secondary use of health and social data. Ministry of Social Affairs and Health, Finland stm.fi
Secure Processing Environment (SPE)	Physical or virtual environment and organisational means to ensure compliance with Union law, in particular with regard to data subjects' rights, intellectual property rights, and commercial and statistical confidentiality, integrity and accessibility, as well as with applicable national law, and to allow the entity providing the secure processing environment to determine and supervise all data processing actions, including the display, storage, download and export of data and the calculation of derivative data through computational algorithms;	DGA /EU) 2022/868, Article 2.
Trusted Research Environment (TRE)	Equivalent to Secure Processing Environment but with a wider governance framework defined by the Health Data	Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems, NHS, 2021.

Term / Acronym	Definition	Source
	Research (HDR) UK. TRE is based on the Five Safes framework enabling data services to provide safe research access to data.: safe people, safe projects, safe settings, safe data and safe outputs.	

References and further reading

EUR-Lex documents with the term “Secure processing environment” (13.2.2022)

1. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act). COM/2020/767 final. Proposal for a regulation. 25.11.2020. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&qid=1676270990522
2. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act). ST 13351 2020 INIT. Cover note. 25.11.2020.
3. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance (Data Governance Act) - Outcome of the European Parliament's first reading (Strasbourg, 4-7 April 2022). ST 7853 2022 INIT. Information note. 11.4.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_7853_2022_INIT&qid=1676270990522
4. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space. COM/2022/197 final. Proposal for a regulation. 3.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0197
5. COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT REPORT Accompanying the document PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space. SWD/2022/131 final. Impact assessment. 3.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022SC0131&qid=1676270990522
6. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). PE 85 2021 INIT. Legislative Act. 4.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:PE_85_2021_INIT&qid=1676270990522
7. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL A European Health Data Space: harnessing the power of health data for people, patients and innovation. ST 8828 2022 INIT. Cover note. 6.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CONSIL%3AST_8828_2022_INIT&qid=1676270990522
8. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL A European Health Data Space: harnessing the power of health data for people, patients and innovation. COM/2022/196 final. Communication. 3.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022DC0196&qid=1676270990522
9. COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT REPORT Accompanying the document PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space. ST 8751 2022 ADD 3. Cover note. 6.5.2022. https://eur-

lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_8751_2022_ADD_3&qid=1676270990522

10. COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT REPORT Accompanying the document PROPOSAL FOR A REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space. ST 8751 2022 ADD 4. Cover note. 6.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_8751_2022_ADD_4&qid=1676270990522
11. REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ON EUROPEAN DATA GOVERNANCE AND AMENDING REGULATION (EU) 2018/1724 (DATA GOVERNANCE ACT). PE 85 2021 REV 1. Legislative Act. 30.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:PE_85_2021_REV_1&qid=1676270990522
12. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act, DGA) PE/85/2021/REV/1. Regulation, in force. 30.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_8751_2022_INIT&qid=1676270990522
13. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space. ST 8751 2022 INIT. Cover note. 6.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R0868
14. COMMISSION STAFF WORKING DOCUMENT Final evaluation of the European Interoperability Framework (EIF) Accompanying the document Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act). SWD/2022/720 final. Staff working document. 18.11.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022SC0720
15. COMMISSION STAFF WORKING DOCUMENT Final evaluation of the European Interoperability Framework (EIF) Accompanying the document Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act). ST 14973 2022 ADD 1. Cover note. 18.11.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_14973_2022_ADD_1&qid=1676270990522