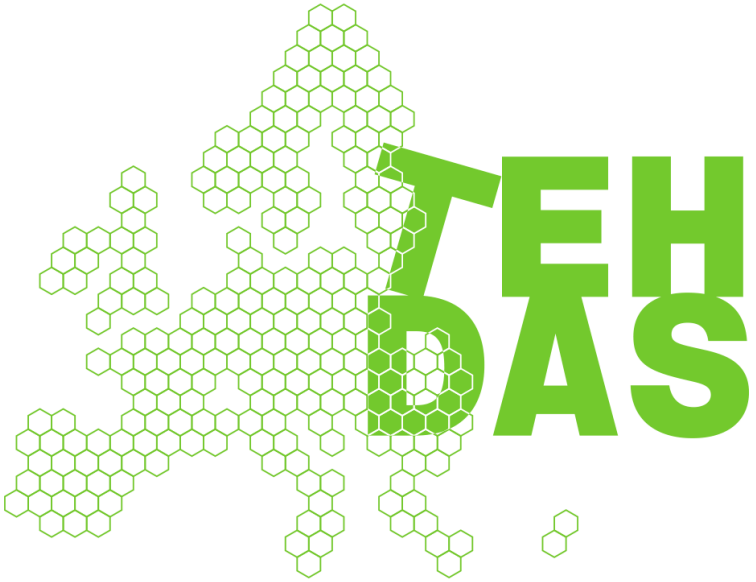


Towards European Health Data Space

Deliverable 7.2

Options for the services and services architecture and infrastructure for secondary use of data in the EHDS

4 July 2023

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

0 Document info

0.1 Authors

Author	Partner
Inês Antunes	Shared Services for Ministry of Health, Portugal
Enrique Bernal-Delgado	Aragón Institute of Health Sciences, Spain
Antal Bódi	Semmelweis University, Hungary
Fidelia Cascini	Università Cattolica del S. Cuore, Italy
Pascal Derycke	Sciensano, Belgium
Sérgio Dinis	Shared Services for Ministry of Health, Portugal
Francisco Estupiñán-Romero	Aragón Institute of Health Sciences, Spain
Yasmin Fonseca	Shared Services for Ministry of Health, Portugal
Juan González-García (JGG)	Aragón Institute of Health Sciences, Spain
Lionel Grondin	Health Data Hub, France
Irene Kesisoglou	Sciensano, Belgium
Truls Korsgaard	Directorate of e-health, Norway
Vanessa Lima	Shared Services for Ministry of Health, 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
Vanessa Mendes	Shared Services for Ministry of Health, Portugal
Juha Pajula	VTT, Technical Research Centre of Finland, Finland
Cátia Pinto	Shared Services for Ministry of Health, Portugal
Marja Pirttivaara	Sitra, Finnish Innovation Fund
Philipp Schardax	Federal Ministry of Social Affairs, Health, Care and Consumer Protection, Austria
Katharina Schneider	Health Data Lab (Federal Institute of Drugs and Medical Devices), Germany
Anne Heidi Skogholt	Directorate of e-health, Norway
Dylan Spalding	CSC, IT Centre for Science, Finland
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 30 May 2023. The European Commission gives final approval to all joint action's deliverables.

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

Copyright Notice
Copyright © 2023 TEHDAS Consortium Partners. All rights reserved. For more information on the project, please see www.tehdas.eu.

Contents

1 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, and its continuation released as Milestone 7.6. 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 key technical components of the HealthData@EU infrastructure: the information systems to manage the national metadata catalogues, the information systems to manage the cross-border data access applications and data requests, and the secure processing environments. The dissertation about these three elements is presented in the guidelines included in Annexes A to C and synthesize a considerable amount of discussion that took place as part of the WP7 activities, in the form of work package internal meetings, workshops with the Work Package Advisory Group, meetings with external stakeholders and surveys to external services providers.

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

[2] TEHDAS Deliverable 7.1 "*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/

2 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 second and last deliverable produced within WP7, which addresses O7.4, using as inputs the results described in the previous milestones where objectives O7.1 to O7.3 were addressed. This deliverable builds on top of the Milestone 7.6[5], that provided the initial analysis of the architectural options for services identified Deliverable 7.1 "*Options for the minimum set of services for secondary use of health data in the EHDS*"[6] as well as the deployment options. This deliverable also provides a further view on the evolution of the Users' Journey and architecture proposals made during the Joint Action, a further dissertation regarding the possible

[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] TEHDAS Milestone M7.6 "*Report on architecture and infrastructure options to support EHDS services for secondary use of data*" https://tehdas.eu/results/tehdas-analysis-on-ehds-technical-infrastructure/

[6] TEHDAS Deliverable 7.1 "*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/

implementations for the described services, putting some stress 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 (SPEs). The analysis of requirements and specific solutions for these three components were discussed in dedicated workshops with internal and external stakeholders, especially in dedicated workshops with the Work Package Advisory Group, whose results are covered in the Milestone 7.4 "*Validation report on the proposed services and architecture and infrastructure solutions*"[7]. For the specific case of the SPEs, a survey to existing institutions operating such systems was already in place. The results of such exploration and further discussions are covered in the main body of the deliverable and a set of three annexes (A to C) in the form of specific guidelines to drive the design of the aforementioned systems. These three guidelines are part of a direct request given by the European Commission to this work package.

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.

[7] TEHDAS Milestone M7.4 "Validation report on the proposed services and architecture and infrastructure solutions" https://tehdas.eu/app/uploads/2023/04/tehdas-validation-report-on-the-proposed-services-and-architecture-and-infrastructure-solutions.pdf

3 TEHDAS Users' Journey

3.1 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.

3.2 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.

3.2.1 *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 must 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).

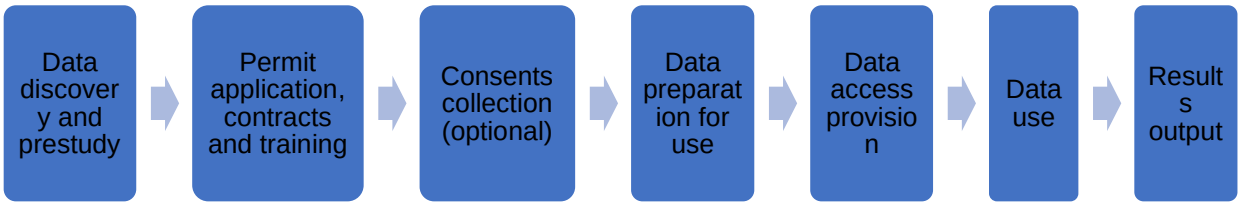


Figure 1: Original TEHDAS' Users' Journey

3.2.2 The revised User Journey

The revisited User Journey, depicted in Figure *1*, is richer in terms of separation of concerns than the one presented in the Milestone 7.5[8]. 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.

[8] 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/

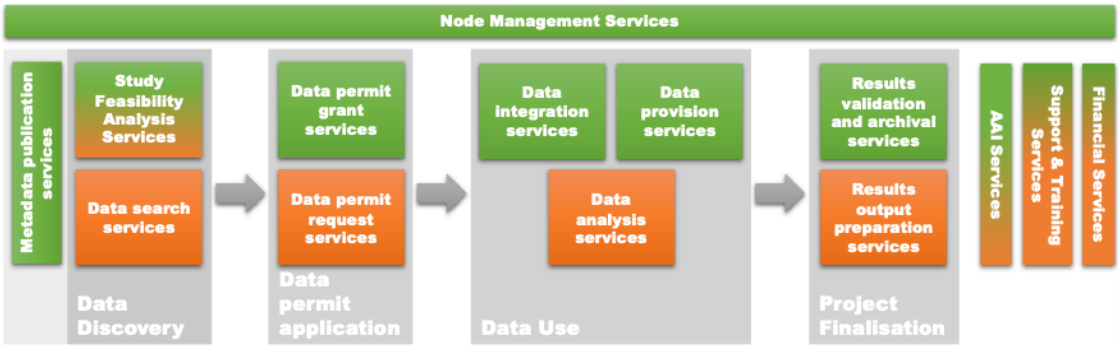


Figure 1: Second version of TEHDAS Users' Journey

In the schema (Figure *1*), 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 *1* 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 EHDS.
3. **The data use phase**: the data use phase is the phase where the data user is provided by individual level data. Then, he or she finally performs the data analyses he or she needs to perform the work, to answerthe 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[9] for the results. It may imply a notification of the incidental findings to the data controllers.

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

3.2.3 The TEHDAS' data lifecycle

The TEHDAS' data lifecycle is depicted in Figure *2*. 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.

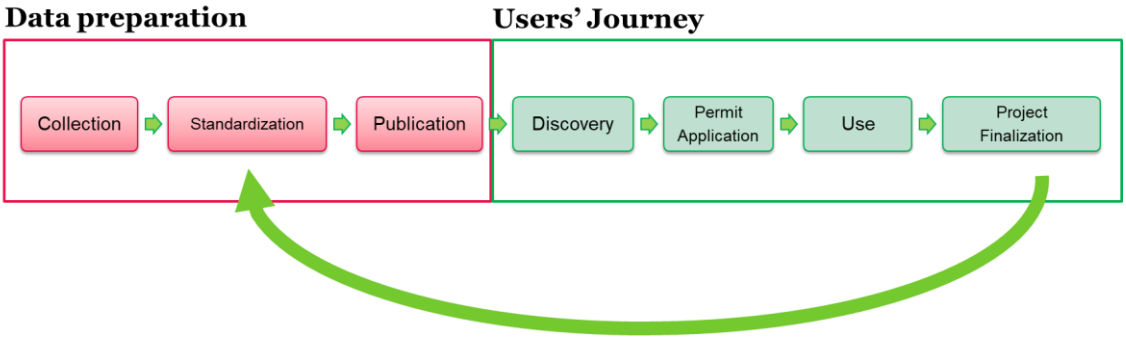


Figure 2: TEHDAS' proposed data lifecycle

3.2.4 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 *3*.

Figure 3: 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 *1* '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.

3.3 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 *1*) and the core of the Publication phase (in the data lifecycle, see Figure *2*); the data permit application systems that cover the Data permit application phase and the secure processing environments, that cover the Data use phase.

4 Architecture Scenarios

4.1 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.

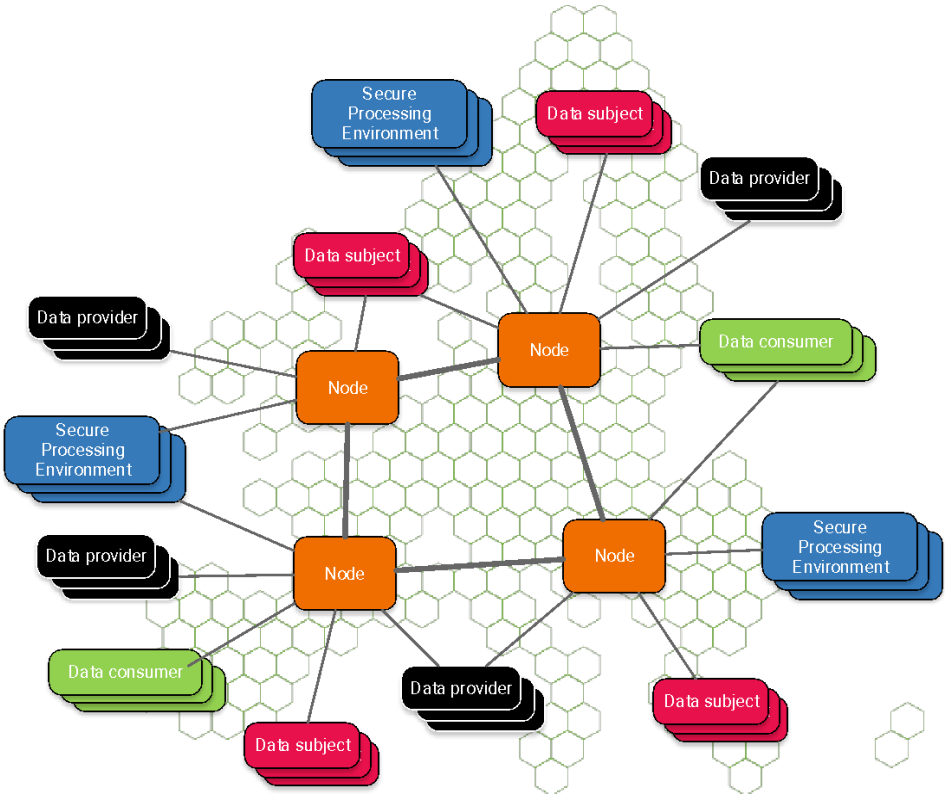


Figure 4: Original TEHDAS architecture proposal (Milestone 7.5)

4.1.1 First TEHDAS architecture

The original TEHDAS' architecture proposal, depicted in Figure *4*, was presented in Milestone 7.5 and introduced a pure peer-to-peer architecture (more details on this in section 4.2.3), 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 *1*).

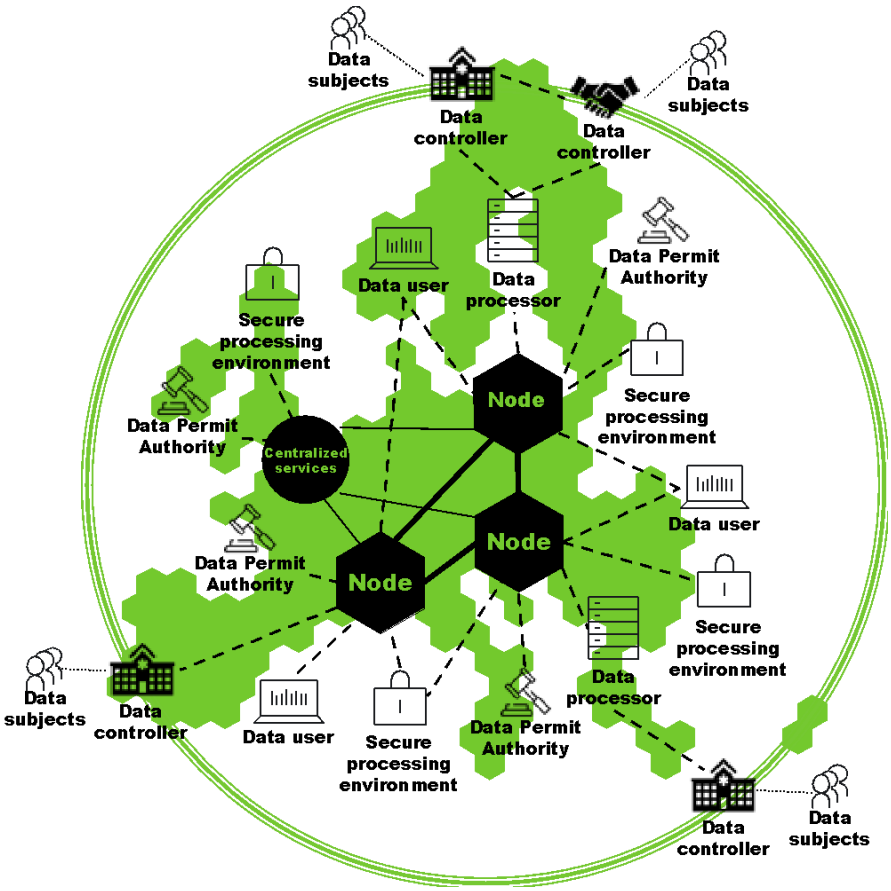


Figure 5: Second version of the TEHDAS architecture proposal (Deliverable 7.1)

4.1.2 Second version of the TEHDAS architecture

The second version of the TEHDAS' architecture proposal, depicted in Figure *5*, 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.

4.1.3 HealthData@EU architecture proposal

Figure *6*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 *5*. 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.

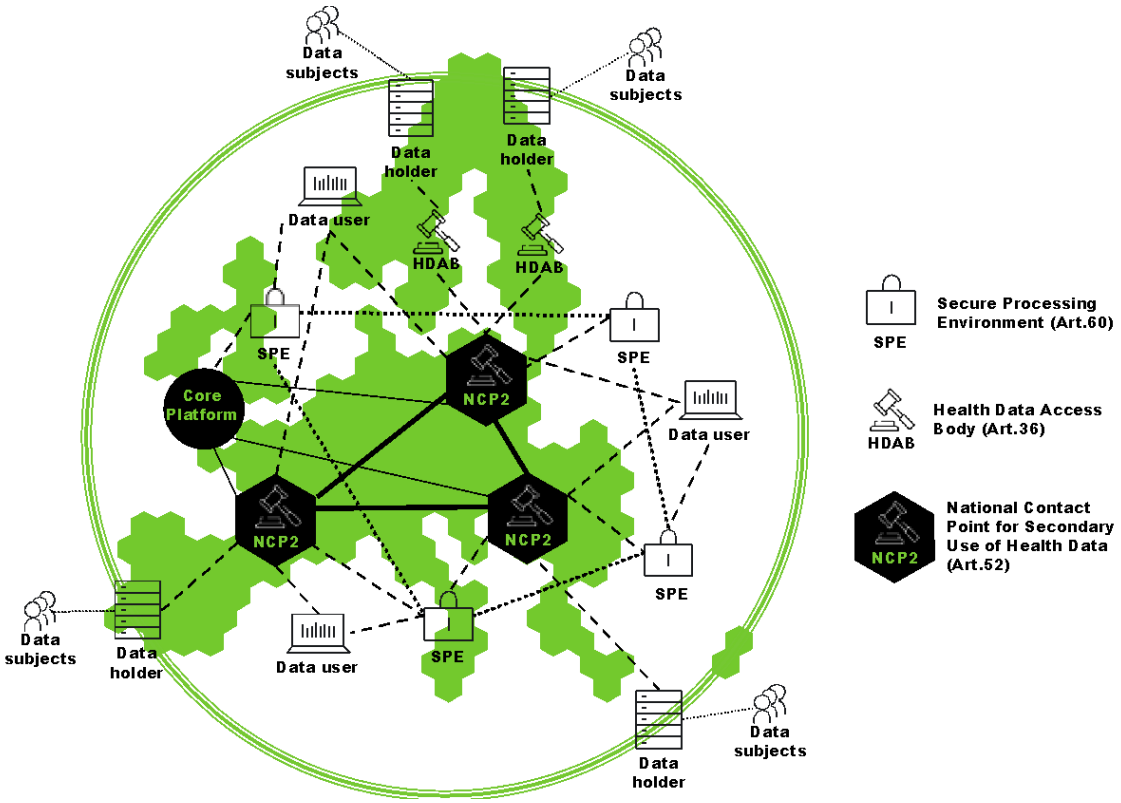


Figure 6: HealthData@EU proposed architecture (adaptation from EHDS legislative proposal)

Finally, Figure *7* includes a simplified schema of the one depicted in Figure *6*, 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.

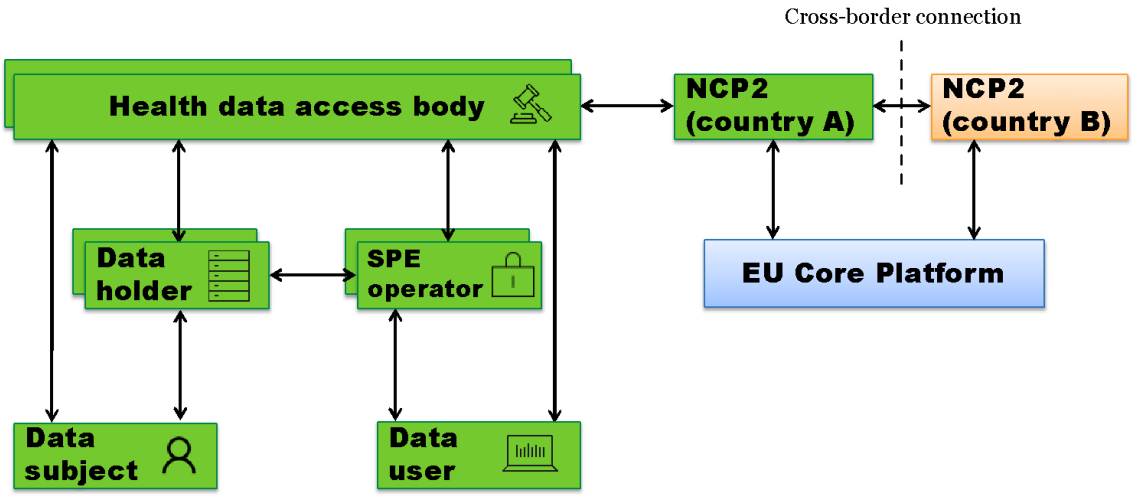


Figure 7: HealthData@EU simplified architecture

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.

4.2 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 they briefly described their particularities. There is a distinction between a centralised approach (section 4.2.1) and a distributed approach (section 4.2.2). 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.

4.2.1 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.

4.2.2 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.

4.2.3 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 *6*, 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.

4.2.4 Peer-to-peer (p2p) deployment

In a peer-to-peer architecture deployment[10], 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 query to search to inquire the rest of the nodes, thus each node may act as a search server in the infrastructure, depending on where the data users are accessing.

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.

4.3 Data lifecycle and architecture actor's involvement

Figure *8* contains a schema depicting the participation of the different actors described in the architecture proposal in the data lifecycle from Figure *2*. 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 *1*) 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.

[10] 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.

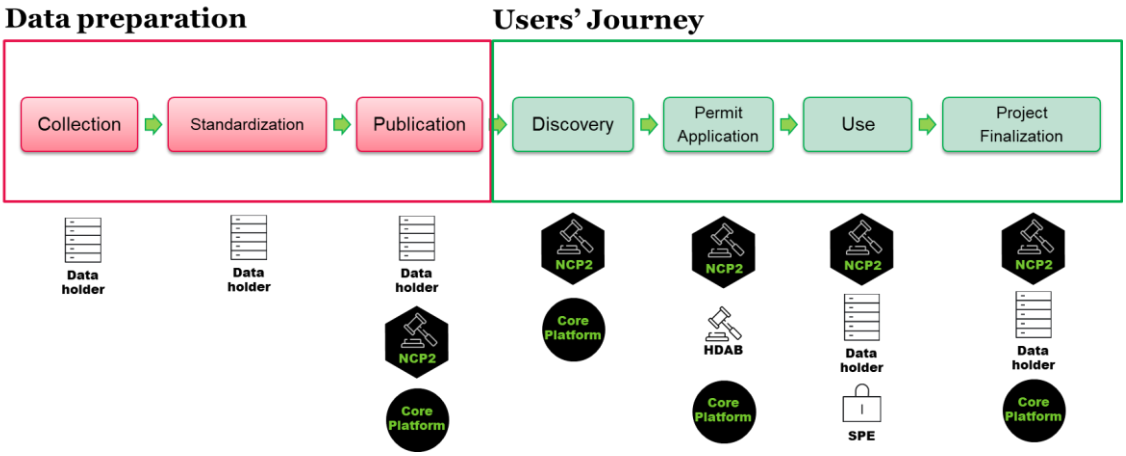


Figure 8: TEHDAS' Data lifecycle mapping HealthData@EU architecture actors

5 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 5.1.1, on the Metadata publication services. Discussion related to the management of data access applications, data requests and data permits is present along section 0, 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 5.2, related to the data use phase.

5.1 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.

5.1.1 Metadata publication services

As introduced in the Section 3.2, the TEHDAS' Users' Journey (Figure *1*) 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 *2*), 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 datasets catalogue, settled in a HDAB (Art.37(1)(q)(i)). The internal coordination to generate the national datasets catalogue reflects 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 and ensure a seamless interconnectivity among the MSs nodes and the EU Datasets Catalogue. Interconnectivity means that technical and semantic interoperability is achieved by the adoption of a common metadata standard, one and several common exchange protocols, common serialisations, the same security rules, the same data quality framework, etc. It would be recommendable, for instance, 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. Even if a gateway mechanism and a mapping of metadata can be envisaged.

Other advantages could rely upon the harmonisation across the metadata publication services of each data holder responsible for their metadata descriptions in a given MS node. 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 node, 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)), and also the data quality and utility label (Art.56) 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 and the application of the data quality and utility evaluation tools. In any case, it would be possibleto present more information about the dataset, for example, its coverage, missing rates, 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 bonus 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.. 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 on an individual's health.

Table *1* contains the analysis of the possible scenarios proposed in the above text regarding the metadata publication services. Table *2*, contains the scenarios for the metadata synchronisation between the national datasets catalogue and the EU Datasets Catalogue.

Table 1: Scenarios for metadata publication services

Priority	Scenarios	Pros	Cons
1	Multiple data holders that connect to a single HDAB per country	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.
2	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 catalogue maintained by each HDAB to the national datasets catalogue, maintained by the coordinator HDAB. 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.
3	Open Portals linked to 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.

Table 2: Scenarios for metadata synchronisation

Priority	Scenarios	Pros	Cons
1	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 in some periods of time. EU Core Platform may incur in high-capacity requirements on each EU-wide update.
2	Coordinator HDAB interact with the EU Core Platform bodies to 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. National datasets catalogue updates may be transferred to the EU Datasets Catalogue as they occur.	Extra burden on the coordinator HDAB technological solutions. Malicious attacks may pollute the EU Datasets Catalogue.
3	Coordinator HDAB directly stores national datasets catalogue in a dedicated space of the EU Core Platform	A single information system provides the overall cataloguing features. Leverages the technological burden of the coordinator HDAB	Single point of failure for both national and EU cataloguing systems

5.1.2 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 is codified.

The first two points, regarding the metadata catalogues placement and the information they covered the present legislative proposal, have beenintroduced 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 foresees a scenario where it is expected that the EU Datasets Catalogue will be the main system inquired to perform the data searches.

Table 3: Possible scenarios of the data search services

Priority	**Scenarios**	**Pros**	**Cons**
1	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.
2	An EU Datasets 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.
3	EU Datasets Catalogue to also include metadata of open data sets (in addition to the metadata of the national register datasets).	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.
4	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.

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 *3* 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[11] 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"[12] (DCAT-AP), promoted by the EC. This standard, based on the Data Catalogue Vocabulary[13] (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 becomes more realistic, for example, if the national datasets catalogue retains a grade of metadata deeper than the one exposed in the EU Datasets catalogue, this will also open the possibility of performing more much sophisticated searches, for example those based on metadata summaries at variable level, as the ones offered in the Atlas[14] tool provided by OHDSI as part of the OMOP ecosystem. It is worth noting that this possible "advanced search" (sometime named as "federated querying"), would be possible in the first scenario as the EU Dataset Catalogue described in this scenario contain all the detailed data.

In addition, there will be also key for an even successful search that, in addition to the discovery of the datasets themselves, it is also exposed the quality attributed to such datasets. This is regulated by the Article 56 of the legislative proposal, and will also

[11] 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

[12] "*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

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

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

suppose an extra burden within data holders so as to apply the procedural elements defined in implementing act. It will be recommended, that the application of the labeling tools,yet to be defined in the implementing acts regulated in Art.56(5) are assisted by Health Data Access Bodies.

5.1.3 Study feasibility analysis services

The study feasibility analysis services, in conjunction with the "Support and Training Services" (see Section 5.4.3), 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 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 *4* contains the possible scenarios of the study feasibility analysis services.

Table 4: Possible scenarios of the study feasibility analysis services

Priority	Scenarios	Pros	Cons
1	Data experts reside at data holder level.	Highest knowledge of data available	Difficult on the consultancy operations management
2	Data experts reside at HDAB level	Aggregation of "national" or "thematic" data knowledge, depending on the HDAB deployment	Some decoupling with datasets knowledge
3	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.

Considering that the current legislative proposal does not foresee a direct interaction between data holders and data users, it would be desireable that the health data access bodies facilitate and encourage this interaction, as part of the regular exchange between data holders for example, for preparing and transferring the datasets catalogues/descriptions or to access to data. This would be accepted by MS as the scenario's prioritisation for this service, see Table *4*, thee TEHDAS WP7 recommeded to to have this expertise allocated at data holder level. In this situation, as this attribution won't be mandatory per law, this specific exchange would be organised on per country basis, but it would be recommended to be included as an extension of the search services. A compensation scheme between HDABs and data hodlers might be in place for example, by charging to data users this secific services provided by data holders or including this cost as part of the HDAB operational fees. Data permit application phase

The data application phase starts when a data user found the data of interest for its purpose. The legislative proposal distinguishes to types of interaction to access to the data available through the HealthData@EU. First, a data access application (Art.45), when data user petition asks for individual level data, which in case of being accepted by the competent HDAB will generate a data permit (Art. 46). Second, a data request (Art.47), when data user petition asks for aggregated data, e.g., health indicators, which in case of being accepted by the competent HDAB, the HDAB will provide the aggregated data itself.

As introduced when presenting the TEHDAS Users' Journey and Data Lifecycle, in the TEHDAS conception of the EHDS process, the current phase is followed by a data use phase where the individual level data is analysed, so the current phase focuses on the obtention of a data permit and so the phase name. In any case, even though the outcome of a data access application or a data request is different, the technical processes for requesting and approving such petitions are expected to have a high technical overlapping in both cases, so the services included in this phase here can be applicable when processing a data request.

The data permit application process is ensured by two services. On one hand, the data permit request services focus on the data users' interaction, i.e., the requester side, by providing he or she the mechanisms to apply and manage his or her petitions. On the other hand, the data permit grant services target the approvers from HDABs, i.e., the granter side, by allowing them to review and validate or not such petitions.

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

5.1.4 Data permit request services

The Data permit request services would allow the data user to submit a data access application petition, 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 to ensure consistency between the different instances of the system.

Table *5* contains the possible scenarios for the data permit request services.

Table 5: Possible scenarios for the data permit request services

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

5.1.5 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. The main downside of this approach is that the approval process organisation will tend to have a fixed structure, limited by the capabilities offered by the solution selected in the central server. The selection of the solution may limit for example the particularities different member states (HDABs) may have, e.g., a given MS hay have many HDABs scattered that may need to be consulted and re-consulted in a particular order to emit its decision. On the other hand, recurring to a highly flexible solution may suppose an extra configuration burden for MS with simple approval chains.

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 approval

process. The main downside of this approach is the complexity in management of approvals for requests on data under different scope of responsibility.

Table *6* contains the possible scenarios for the data permit grant services.

Table 6: Possible scenarios for the data permit grant services

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

Table 7: Possible scenarios for the interaction between data permit request systems and data permit grant systems

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

		Possible customisation in the approval process per HDAB	Complex management of multi-country approvals/requests for revision

5.1.6 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 *7* contains the possible scenarios for the interaction between data permit request systems and data permit grant systems.

5.2 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 access to. In this phase, 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.

5.2.1 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 scattered in multiple data holders/MSs nodes 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. In addition, it would be in this step of the overall process where, once the datasets requested have been integrated, it would be necessary to apply the pseudonymisation, if not already provided, or the anonymisation.

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 *8*.

Table 8: Possible scenarios for the data integration services.

Priority	Scenarios	Pros	Cons
1	Integration of datasets at HDAB level	Leverage burden to data holders, but the expertise on data particularities is still closer.	May result in an unscalable approach. Extra technical solutions are required to provide external datasets linkability.
2	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.

3	Integration of datasets at EU 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.
4	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 like 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.

5.2.2 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 integrity, and possibly the anon 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[15], 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.

5.2.3 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)

In the literature, it is possible to find other terms for such environments, for example Trusted Research Environment (TRE) has been used by the Health Data for Research (HDR) UK to refer to the environments where personal data is processed for research

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

purposes[16], which also include organisational elements. In the Scotland context, the HDR UK TRE implementation was already deployed under the name of Safe Havens[17].

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 *9*present the comments and considerations to be made for each requirement in such work.

Table 9: 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.

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

[17] Gao C, McGilchrist M, Mumtaz S, Hall C, Anderson LA, Zurowski J, Gordon S, Lumsden J, Munro V, Wozniak A, Sibley M, Banks C, Duncan C, Linksted P, Hume A, Stables CL, Mayor C, Caldwell J, Wilde K, Cole C, Jefferson E. A National Network of Safe Havens: Scottish Perspective. *J Med Internet Res.* 2022 Mar 9;24(3):e31684. doi: 10.2196/31684. PMID: 35262495; PMCID: PMC8943560.

Text in EHDS proposal	Comments and considerations
(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[18]. 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"

[18] "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/

Text in EHDS proposal	Comments and considerations
	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.
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 "5.2.2 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[19] (Cybersecurity Act) and Article 24 Directive (EU) 2022/2555[20] (NIS2) Cloud Infrastructure Service Providers Europe Code of Conduct for cloud infrastructure service providers[21], an effort approved by the CNIL, the French independent authority that veils for security and privacy of personal data.

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

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

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

Text in EHDS proposal	Comments and considerations
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 SPEs 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 SPEs and key requirements. For example, the approach of enabling several SPEs per country is largely supported. At the same time, there are still several 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 *9* provides a schematic view of the interconnection within the actors that interact with a SPE.

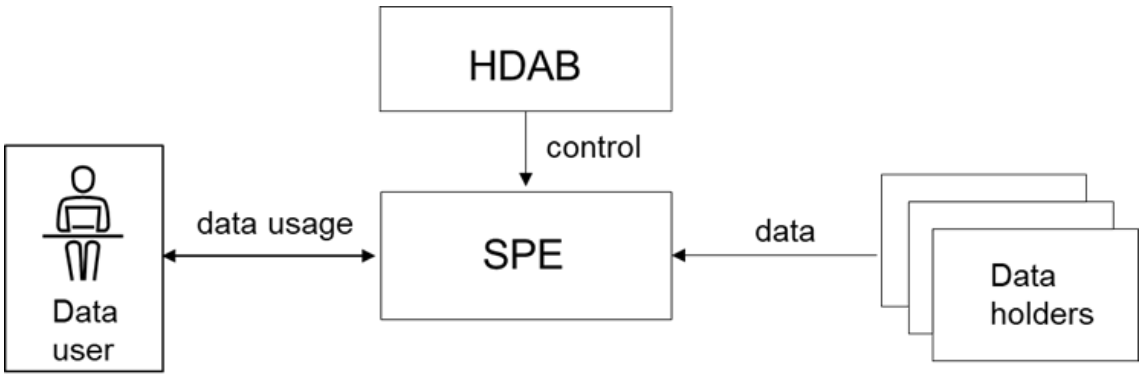


Figure 9: Access to data provided via secure processing environment (SPE)

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)[22]. Additionally, the data permit the requirement to facilitate the description of the specific tools available in the SPE for 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 SPEs 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.

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

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 might be possible to re-identify individuals present on de-identified data[23]. 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.

[23] 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

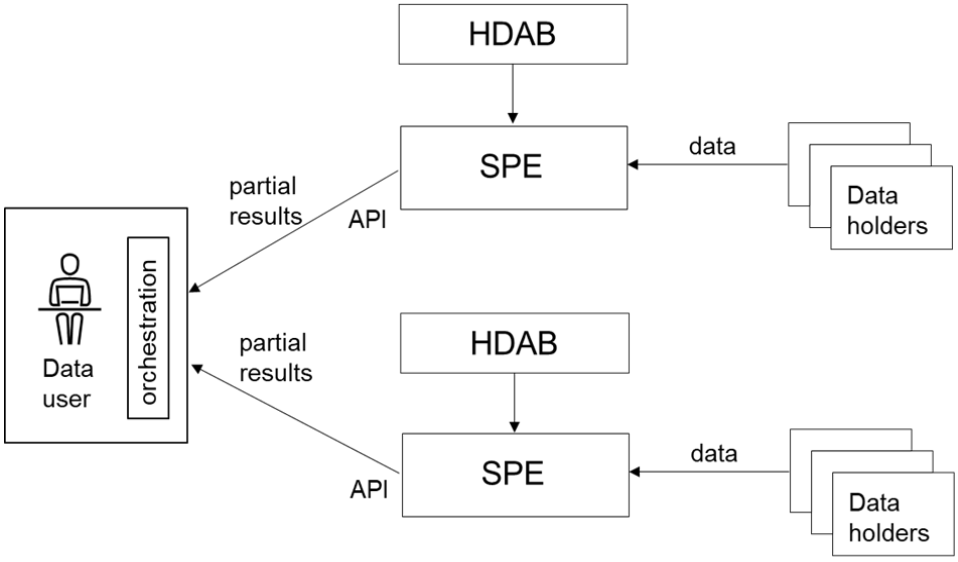


Figure 10: Simplified federated analysis architecture

Figure *10* 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.

The following specific characteristics are required by an SPE to support federated analysis as outlined by Figure *10*:

- **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 *10*, 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 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[24]. 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 minimisation**: 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.

[24] 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 confirm user identity safely and reliably, 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 enhancing 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. Recital 26 of the GDPR only states that:

"*The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*" and "*To determine whether a natural person is identifiable,*

account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."

The given definition of whether a natural person is identifiable leaves much room for interpretation. 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 minimising the exposure of personal information to be processed in the SPE, privacy techniques are also relevant for ensuring the privacy of the analysis results export from the SPE. Protecting privacy of the analysis results exports is partially elaborated in section 5.3.

A well-known concept for enhancing data privacy is k-anonymity[25]. This concept accounts for the fact that even after removing identifiers such as names, addresses etc. an identification of individuals can still be 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. There is a wide range of synthesis methods available, and more are being developed every year. It is still an open question which methods perform best for which use cases and more standardised benchmarks are needed in this regard. Also, which metrics are most suitable to quantify the utility and privacy of the synthetic data is still a subject of ongoing research. In addition, specific questions regarding the typical structure of health data need to be answered. Most synthesis methods currently take a single table as input, while large health datasets are often stored in relational databases. This poses additional challenges for synthesis methods and raises the question of whether synthesis of the entire database is possible or whether synthesis of smaller analysis-specific datasets is more feasible. The study is still ongoing at the time of drafting this deliverable, but it can be expected that the outputs can be harnessed in the subsequent HealthData@EU projects.

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

In Denmark, a study on the use of synthetic data ("'*Vision for better use of Danish Health Data*") is being performed[26]. The German Health Data Lab is currently conducting a study ("*Artificial Intelligence at the Health Data Lab - Investigation of anonymisation methods and AI-readiness (KI-FDZ)*") aiming to compare classical anonymisation methods such as k-anonymity with synthetic data in terms of utility and the remaining risk of disclosure[27]. 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. The KI-FDZ study could already identify relevant questions that need to be addressed when applying data synthesis methods.

Two relatively new privacy preserving methods are differential privacy and homomorphic encryption. The concept of differential privacy has been presented for the first time by the group of Dwork et al.[28] and involves adding random noise to the data. A review by Ficek et al.[29] evaluated the use of differential privacy in health research and concluded that, while being "one of the strongest methods of controlling disclosure risk in recent years" it is "at an early stage of development for applications in health research, and accounts of real-world implementations are scant". Regarding the privacy-utility trade off, they stated: "Significant gaps exist, however, for applications involving explanatory modelling and statistical inference, which are particularly important in epidemiology and clinical research." Therefore, it appears that even though being effective in protecting sensitive information, the current status does not provide enough evidence for the feasibility of a general application in health data provision. The same problem seems to apply to homomorphic encryption. This approach enables performing analyses on encrypted data without the necessity of decryption. It can be considered as an emerging technology for controlling disclosure risk, but there are still obstacles that hinder a routine application in health data provision. A recent review scrutinising homomorphic encryption for privacy-preserving biometrics finds that "*it still faces unsolved issues, such as high computational complexity, low efficiency, and inadequate deployment in the real world. Further research is needed to make Homomorphyc-related encryption, decryption, and matching processes more efficient and practically implementable.*"[30]

In summary, it can be said that there is no current standard for privacy preserving techniques within the framework of health data provision. Some interesting approaches

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

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

[28] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T, eds. *Theory of Cryptography TCC 2006*. Berlin, Heidelberg: Springer; 2006: 265–284.

[29] Joseph Ficek, Wei Wang, Henian Chen, Getachew Dagne, Ellen Daley, Differential privacy in health research: A scoping review, *Journal of the American Medical Informatics Associatio*n, Volume 28, Issue 10, October 2021, Pages 2269–2276, https://doi.org/10.1093/jamia/ocab135

[30] Yang W, Wang S, Cui H, Tang Z, Li Y. A Review of Homomorphic Encryption for Privacy-Preserving Biometrics. *Sensors* (Basel). 2023 Mar 29;23(7):3566. doi: 10.3390/s23073566. PMID: 37050626; PMCID: PMC10098691.

for controlling disclosure risk exist, but to date there is not enough evidence verifying their suitability for routine use. It would be crucial to account for this during the HeathData@EU preparations to promote finding appropriate anonymisation solutions.

Apart from the question of "how" to anonymise the data, the question of "when" needs to be addressed. Regarding 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 further step towards a better accuracy of the results, it could be considered to include an automated process that applies the analysis scripts to the original data after their development on the anonymised data. This way, aggregated results based on the original data could be achieved without having to expose sensitive details to the user.

As a final mention, indicate that work package 6 of TEHDAS oversees addressing the issue of de-identification by developing data minimisation 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 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[31]. 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.

[31] 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

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 like 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.

5.3 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, to guarantee the reproducibility of the results in research context (or others).

5.3.1 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 10: Possible scenarios for the services for results validation for disclose (data export verification)

Priority	Scenarios	Pros	Cons
1	Results export audit manually operated	Higher precision in the audit of the results to be exported	Non-scalable approach. Not applicable to the export of the federated analysis partial results.
2	Computer-based results export audit (e.g., AI assisted)	Higher scalability. Possible false positives (results that should be approved marked as non-exportable) Can be used to guarantee the validation of federated analysis	AI solutions yet to be widely tested

Table 11: Possible scenarios for accessing data for reproducibility in scientific publication context.

Priority	Scenarios	Pros	Cons
1	Generation of synthetic data similar with same patterns as original data	Equivalent to anonymised data, with higher level of security.	Technologies to generate fully comparable synthetic data yet to be widely tested.

2	Access to a subset of the original data (after permit granted to the HDAB)	Subset of data can be easily controlled	May be difficult to provide a representative sample
3	Disclose anonymised version of the original input datasets	Easy setting to disclose the original data	May prevent to actual reproducibility
4	Access to all the original data (after permit granted to the HDAB)	Easy setting to reproduce the results	Currently not considered in the actual EHDS data access models.

Finally, 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.

5.3.2 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.

5.4 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.

5.4.1 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[32], and, up to some extent, with the coordinator HDABs that might expose some of the services, if specific services deployments are selected, if

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

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 12: Possible scenarios for the node management systems

Priority	Scenarios	Pros	Cons
1	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.
2	External auditing by Core Platform	No burden on NCPs to perform the auditing.	Only external auditing expected, e.g., only intrusion tests.
3	Self-reported node auditing	No extra burden on Core Platform	High trust requirements to the NCPs. Extra burden on NCPs.

5.4.2 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 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 *13*. 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 13: Possible scenarios for AAI services

Priority	Scenarios	Pros	Cons
1	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[33].	Extra complexity of the AAI system to guarantee the interoperability between MS systems.
2	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

5.4.3 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 its 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 *14* lists the two scenarios foreseen for the deployment of the support system, which has high resemblance with the prestudy services.

Table 14: Possible scenarios for support and training services

Priority	Scenarios	Pros	Cons
1	HDABs offer support to its users, which is coordinated with the EU Core Platform	Higher availability of support services, closer to the users. A central knowledge hub may help to centralise common issues.	High coordination burden when support implies actors from more than one MS.
2	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.

[33] 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

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.

5.4.4 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 *15*.

Table 15: Possible scenarios for the financial services

Priority	Scenarios	Pros	Cons
1	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.
2	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
3	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.

6 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 to other community-specific technological infrastructures.

6.1 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 applications and data access requests; and the secure processing environments.

6.1.1 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 5.2.3.

6.1.2 Infrastructure for data access requests management systems

The foreseen infrastructure required for the data permit management systems is like 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.

6.1.3 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.

Machine learning, and especially its deep learning subdomain, typically rely on the optimisation of large-scale arrangements of relatively simple (activation) functions. This makes it possible to break complex training tasks down into smaller and simpler chunks, and parallelize the calculations needed for optimisation. In many cases, the training of machine learning models can be sped many folds if computations are shifted to designated units performing vector, matrix, or tensor (a cube of values with 3 or more dimensions) operations in parallel. This is especially important in use cases where data easily reaches the dimensions of Big Data - for example when working with image data, genetic data, or complex molecular structures such as proteins.

While CPUs rely on sequential operations distributed between their few (physical and virtual) cores, GPUs can handle tens of thousands of parallel operations per cycle. Even more specialised, designated Tensor Processing Units (TPUs) currently can handle up to more than 100.000 operations per cycle. They are highly optimised for large batch sizes and to optimise models with large numbers of parameters, but support for them by software libraries is comparably limited. GPUs in many regards perform similarly to TPUs, but as of 2023 are still significantly cheaper and more versatile. As such, even for smaller projects in these domains it becomes crucial to at least have access to GPU compute, as otherwise waiting times for researchers can become a great burden. A hardware sizing form may be employed to determine data users' requirements especially for larger analysis projects.

Another consideration are memory requirements, which put physical limits to the amount of trainable model parameters and data points in a batch that can be held while training. Considering that many data users may have a background other than IT, and catering for their convenience, it would be beneficial to avoid or at least reduce the need for manual load distribution between hardware by providing single processing units with sufficient resources. Here, a widely accepted recommendation for smaller deep learning projects is to at least provide 12 GB of GPU memory and 12 of CPU RAM (orienting on popular implementations for convolutional neural networks). This is also very much in line with what popular cloud-based machine learning platforms like Google Collab and Kaggle offer per virtual machine: 12 to 16 GB of RAM plus one GPU like Nvidia's K80, T4 or P100 with 12 to 16 GB of VRAM. In conclusion, these specifications found in popular existing platforms may serve as a starting point per VM for future Secure Processing Environments computational substrate.

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[34], 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 5.2.3 "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[35], a privately

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

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.

In many cases, data collected in the health sector is stored in single (structured) documents. For example, a single hospital can produce hundreds of thousands of hospital discharge letters per year, structured for example in HL7 CDA format. If a data user requests such data from a single or several data users, it may be easily possible that data from millions of documents must be extracted for further analysis. Either at a pre-processing stage at the HDAB, or within the secure processing environment, this consideration must be considered, and optimised computing resources for such information extraction has to be provided, e.g., by allowing parallelisation of corresponding processes.

More conveniently for data users, structured data should be provided in the secure processing environments as .csv format, or by providing a complete database for the data in question. Where the latter is the case, data users also must be provided the corresponding data models, to make sense of it and provide the foundation for valuable analysis. Every processing step within the SPE should be clearly logged.

6.2 Storage infrastructure

In the computation storage it is presented the possible technical solutions that might be applied by data holders and / or health data access bodies, beyond the traditional plain fails or relational databases.

6.2.1 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 its data to be used 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, by nature, usually occur in a routinely and predictable manner.

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 an ecosystem which supports 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 aspect of the interface is critical for successful data lakes, since it should have a good usability for all actors interacting with the systems (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.

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 advanced 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 version the various iterations of the data processing. It also enables data integration, scalability, training artificial intelligence models, and developing security measures to protect the data being feasible to deploy the encryption of data either on its storage, use, process or on the transport.

6.2.2 Data Warehouses

Where data is generated routinely and predictably, data warehouse architectures can offer important advantages. Especially where data is collected over longer time periods, the establishment of designated data warehouse infrastructures at HDABs will provide the foundation for improved data integration, process automation, automated reporting,

and improvements in data pre-processing. For example, national public health agencies may be permanently granted access to data to regularly provide statistics and analyses for public health decision making. This may especially be the case where permanent access to specific data is legally granted to authorities or agencies on a national level. Or third party data users may simply need longer access to structured data in order to perform longitudinal studies. In these cases, data from different registries will have to be extracted in regular intervals from various registries, transformed and integrated, and loaded into designated data warehouses. This process will have to follow a typical data warehouse ETL pipeline. Data users may be granted to data marts or specific data views only, where data keeps updated in an automated fashion, but access remains restricted to what is necessary to each analysis purpose.

Even though storing data in such structured (tabular) databases requires typically much less storage than the same number of entries stored in document-formats, due to the sheer amount of data, storage requirements for data warehouses can be both substantial and hard to anticipate. For instance, a single data warehouse storing national billing data from a small European country's health system holds 4 terabytes of data, growing every year. Where such longer-term storage shall be achieved, it is hence not only important to provide sufficient storage hardware, but also to ensure that the hardware employed is extensible and scalable. Over time, many data warehouse projects reach stifling limits if their storage is not sufficiently extensible. Storage requirements become even more important, as data will have to be mirrored to provide for enhanced failure safety and reliability.

6.2.3 Data Marts

In the context of health data, data marts may provide for individual data requests to the Health Data Access Body, where data shall be accessed over longer periods. Depending on the requirements and approval process, the provision of the data alone can take up to several weeks. Therefore, it is essential to provide sufficient documentation about previously performed analyses to replicate analyses that have already been performed. Whether the data is stored in virtual or dedicated servers must be decided depending on the infrastructure, scalability, cost, and necessary performance parameters. The scalability of virtual databases is particularly interesting for the already growing problem of storing large quantities of information (e.g., image data, pathology data, genomic data).

6.2.4 Option for storage organisation

The following describes how a possible computational backend infrastructure for an HDAB may be implemented.

At the first level are the data holders/data sources, which transmit their respective data to the HDAB on request. This can be done on a (non) regular basis in unstructured form (Data Lake), or in regular and structured form (DWH). In a second step, the structured data sources are loaded into the data warehouse. Particularly data that is regularly delivered and transformed and data that serves local reporting solutions is of importance in this context. The schema-on-write is intended for this purpose, which is a complex

ETL task, but efficient for recurring analyses (e.g., annual surveys of the WHO, national public health, reports to policy makers…). During the ETL process, the data is pseudonymized and subsequently integrated into the DWH's unified master dataset. Similarly, unstructured data is loaded into the data storage node, when only submitted once or without regularity by means of the One-Read scheme. The second level has a link to the preparation node and the metadata node which, as described above, is displayed using a User Interface. This layer is equipped with automated data content management means and automatically displays various parameters in connection with the National Health Dataset Catalogue.

On the third level, namely the preparation layer, the unstructured data from the data lake sources can be pseudonymized, if needed. Furthermore, project-based data set preparation can also be carried out on this level. Additionally, unstructured data can be combined and enriched with structured data. Again, the concept of unified master data is applied, which creates a framework to unify information from several data sources. For instance, the integration of data from HDABs located in different member states may require EU-wide common codes, which may not be available in the unstructured source. At the fourth and final level consist of the upload to the Secure Processing Environments for its further analysis.

6.3 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, it is possible to simplify the communication requirements according to Table *16*, where two dimensions are exposed: the volume of data transfers expected and the level of security in the communications.

Table 16: Characterisation 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[36] (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[37]), or transport level encryption (SSL/TLS[38]). 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

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

[37] 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

[38] 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[39].

6.4 Mapping to existing infrastructures

6.4.1 eHealth Digital Service Infrastructure (eHDSI)

The eHealth Digital Service Infrastructure (eHDSI) is a platform that guarantees European citizens' access to continuity of care while travelling within the EU. This makes it possible for EU Member States to exchange health data in a secure, efficient, and interoperable manner. The services are easily identifiable by the availability of the "MyHealth@EU" brand.[40]

MyHealth@EU is an operational infrastructure as defined in the Commission Implementing Decision 2019/1765 of 22 October 2019, providing the rules for the establishment, the management, and the functioning of the network of national authorities responsible for eHealth, under the Cross-Border Healthcare Directive 2011/24/EU, and embedded where relevant in the Member States national law.[41]

This infrastructure currently supports the exchange of Patient Summary and ePrescription / eDispensation services, both as country of affiliation and country of treatment, and in the short term it will also gradually expand to support other categories of electronic health data, such as laboratory results, discharge reports, and medical images and reports. MyHealth@EU connects healthcare providers in 11 Member States, and by 2025 it is expected that most European countries will have the MyHealth@EU services implemented.

In the proposed EHDS regulation, it is envisaged that MyHealth@EU services will become mandatory for all Member States, so that natural persons can exchange their personal electronic health data cross-border in a foreign language. Therefore, MyHealth@EU is intended to serve as a main building block of the EHDS for the primary use of health data. Therefore, considering the maturity of these services in the Union, several building blocks could be leveraged for HealthData@EU, not only in terms of the primary data and corresponding standards, but also regarding the existing infrastructure and requirements already in routine operation. Different lessons can be uptake for the secondary use and exchange of data among Member States.

6.4.1 Other EC-funded infrastructures of interest

At the moment of writing the present report, there are two projects in progress, funded by the European Commission, whose development may have synergies with the developments of the HealthData@EU infrastructure. The first one is the Genomics Data

[39] *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.

[40] https://health.ec.europa.eu/ehealth-digital-health-and-care/electronic-cross-border-health-services_en

[41] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32011L0024

Infrastructure (GDI), devoted to providing access to genomic datasets, and the second one is the EUropean Federation for CAncer IMages (EUCAIM), whose aim is to provide federated access to European Cancer Imaging Repositories.

At this stage in the development of both projects, it is possible to foresee a certain degree of commonality in the technical components and solutions required for the operation of the infrastructure. In the case of the Genomic Data Infrastructure, similarities with the HealthData@EU services point to a deeper analysis of the proposed solutions and their potential applicability to the latter. In the case of EUCAIM, the maturity of the project is currently not sufficient to identify commonalities.

Genomic Data Infrastructure (GDI)

The European Genomic Data Infrastructure project (GDI) is a Digital Europe project to deploy a federated, secure, and sustainable infrastructure to make genomic and associated phenotypic and clinical data as FAIR (Findable, Accessible, Interoperable, Reusable) as possible subject to the Ethical, Legal and Societal Impact (ELSI) relating to these data. Nodes, hosted in each participating country, will host their own data, with discovery services accessible via a common User Portal. Each node must support the five functionalities required for performing data analysis on genomic or phenotypic data - data discovery, Data access management, data reception, Storage and Interfaces, and Processing (Data Analysis). An example of the proposed architecture so a node can provide these functionalities is given below, including the supporting Global Alliance for Genomics and Health (GA4GH) standards, and an example research use case for access to the data.

The GDI will support data visitation, i.e., sending the analysis to the data as opposed to data distribution where the data is sent for analysis, with the use of Secure Processing Environments (SPEs) allowing analysis of the data. A federated Authentication and Authorisation Infrastructure (AAI) will manage the identities of the users and transmit the list of resources the user has access to within the GDI, with the LifeScience AAI (LS AAI) being proposed as it supports the GA4GH AAI and Passport standards, and utilises the GA4GH recommended OpenID Connect / OAuth2 protocols. Two levels of data discovery will be supported - aggregated and open access data, as well as custom discovery queries over pseudonymised data. In the former, the data will be searchable via a User Portal which will "harvest" these data from the nodes, ensuring the node has control over the data that is sent to the User Portal. For custom queries, the Beacon standard will be used, which allows queries over individual level data, but the data returned to the user can be restricted – for example to Boolean (true/false) or count responses. The search queries will be performed at registered level access, whereby the identity of the user performing the query is known and recorded by the system.

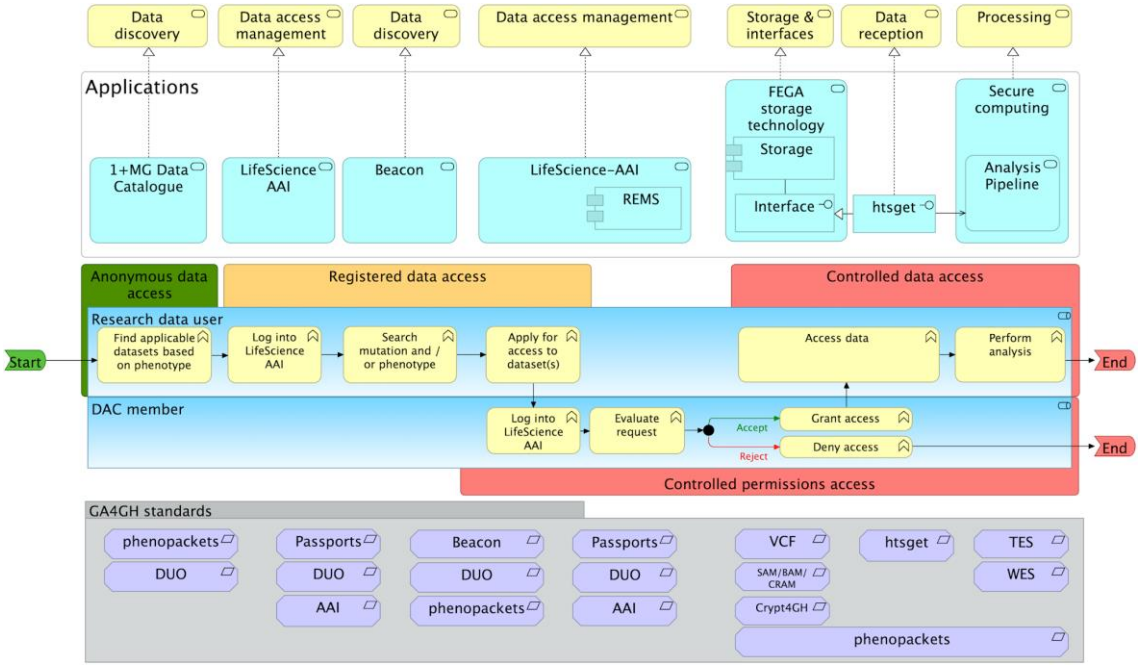


Figure 11 Proposed architecture for the Genomics Data Infrastructure (GDI) - the five functionalities are at the top, with an example secondary use for research use case in the middle and supporting GA4GH standards below.

If data of interest is discovered within the GDI by either or both methods, a research user may apply for access to these data. The research user can use a tool such as the Resource Entitlement Management System (REMS) to make such an application, which can include the research question, any ethics review, agreement to any data access policies etc. The application is reviewed by a Data Access Committee (DAC), who can approve or deny the access request using REMS or similar tool. Assuming that access to the data is granted, REMS or a similar tool will, on request, identify the resources that individual has access to via a GA4GH ControlledAccessGrant visa. This visa, or claim, can be used to access a controlled access resource, such as a SPE or controlled access data, or both. The LS AAI is effectively the transport mechanism for these visas.

At this point the user may access an SPE and perform a data analysis, or subject to the specific ELSI and policies applying to the data, may be able to securely download the data to a location for analysis.

The aggregate or open data for the User Portal is expected, at least in part, to conform to DCAT AP to help ensure interoperability with EHDS. Additionally, GDI is working with SIMPL to try and ensure that relevant data within GDI is accessible via this method.

Developments in eID and eWallet are being monitored, as well as requirements for role-based access control (RBAC) such as those used by different healthcare systems.

GDI will not host personally identifiable data, such as Electronic Health Records (EHR) etc. However, by utilising technologies such as Privacy Preserving Record Linkage (PPRL) it would be possible to link the genomic data within GDI to the health data within EHDS, for those users who have access to the data within the EHDS. It could be that the SPE within GDI are co-located with the SPE for EHDS, with the possibly different

security and data protection requirements and policies, which would facilitate the development of personalised medicine and secondary use for research of suitably consented genomic data.

EUropean Federation for CAncer IMages (EUCAIM)

The objective of the EUropean Federation for CAncer IMages (EUCAIM), is to build a pan-European digital federated infrastructure of cancer-related images, which will be used for the development of AI tools towards Precision Medicine. This infrastructure will provide the means to develop AI tools that will be able to enhance the (cancer) diagnosis procedure, treatment, and the identification of the need for predictive medicine benefiting patients across Europe. The legal grounds for the operation of the federated data repository on the European scale will be defined by adapting to the particularities of the data management regulations of the different European countries.

Organisation of the project[42]

This project has various objectives that include addressing legal and ethical challenges in areas such as medical ethics, data protection, cybersecurity, AI ethics, and social aspects. It also aims to comply with GDPR when dealing with data-intensive research that crosses national borders. To ensure credibility and trust, the project aims to ensure fairness, transparency, accountability, and confidentiality in the processing of personal data.

Another objective of the project is to operate a repository for fair medical imaging data sharing by defining the framework and governing bodies, creating a central data access portal, and developing infrastructure to support authentication, authorisation, traceability, and anonymized data.

The project also aims to specify requirements and implement tools for data pre-processing and interoperability, define a common data model and hyper-ontology, and develop interoperability plugins to enable data and technical interoperability.

A federated analysis infrastructure will be delivered to provide seamless adaptation and integration of existing software solutions into the project infrastructure, with procedures defined for future integration and extension of the infrastructure.

End-users such as hospitals, academic medical centres, image data providers, and AI innovators will work together to identify use-cases that will drive the consensus-based definition and implementation of procedures, protocols, and services that integrate different data sources and processing services to the central hub.

The project aims to ensure long-term sustainability by defining a business model, financial plan, long-term sustainability plan, and implementing an IPR office to follow up on data usage, data ownership, and recognition of providers.

[42] https://cancerimage.eu/what-we-do/

7 Recommendations summary

This section summarises the recommendations about the different scenarios proposed along the deliverable. The prioritisation was based on a survey responded by member states' representatives participating in the Work Package during the final drafting rounds, to have a major knowledge of the scenarios.

The actual results of the voting, as well as the comments received on the different questions are available in the 1.1.1.1.1Annex D. Please note that in the Annex there are also the questions related to specific SPE elements, whose answers are used for its specific guideline, available in the 1.1.1.1.1Annex D.

7.1 Recommendations for metadata publication services

Priority	Scenarios	Pros	Cons
1	Multiple data holders that connect to a single HDAB per country	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.
2	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 catalogue maintained by each HDAB to the national datasets catalogue, maintained by the coordinator HDAB. 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.
3	Open Portals linked to HDABs	Possibility to combine more inputs, beyond personal data.	Open portals with aggregated data need to use the same

Priority	Scenarios	Pros	Cons
			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.

7.2 Recommendations for metadata synchronisation alternatives

Priority	Scenarios	Pros	Cons
1	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 in some periods of time. EU Core Platform may incur in high capacity requirements on each EU-wide update.
2	Coordinator HDAB interact with the EU Core Platform bodies to 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. National datasets catalogue updates may be transferred to the EU Datasets Catalogue as they occur.	Extra burden on the coordinator HDAB technological solutions. Malicious attacks may pollute the EU Datasets Catalogue.
3	Coordinator HDAB directly stores national datasets catalogue in a dedicated space of the EU Core Platform	A single information system provides the overall cataloguing features. Leverages the technological burden	Single point of failure for both national and EU cataloguing systems

Priority	Scenarios	Pros	Cons
		of the coordinator HDAB	

7.3 Recommendations for search services architecture

Priority	Scenarios	Pros	Cons
1	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.
2	An EU Datasets 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.
3	EU Datasets Catalogue to also include metadata of open data sets (in addition to the metadata of the national register datasets).	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.
4	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.

7.4 Recommendations for feasibility study services organisation

Priority	Scenarios	Pros	Cons

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

7.5 Recommendations for data permit request-side services architecture

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

7.6 Recommendations for data permit grant-side services architecture

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

7.7 Recommendations for request and grant side services interaction architecture

Priority	Scenarios	Pros	Cons
1	Hybrid with centralised requests and	No replication of system per HDAB.	Complex migration from current

	distributed grant services	Possible customisation in the approval process per HDAB	application management systems. Complex maintenance to ensure consistency.
2	Fully centralised	No replication of system per HDAB. Easier management of multi-country approvals/requests for revision.	Complex migration from current application management systems. No customisation based on specific needs for approval.
3	Hybrid with distributed requests and centralised grant services	Each HDAB retains control of the system. Easier management of multi-country approvals/requests for revision.	Complex maintenance to ensure consistency. No customisation based on specific needs for approval.
4	Fully distributed	Each HDAB retains control of the system. Possible customisation in the approval process per HDAB	Complex maintenance to ensure consistency. Complex management of multi-country approvals/requests for revision

7.8 Recommendations for data integration services location

Priority	Scenarios	Pros	Cons
1	Integration of datasets at HDAB level	Leverage burden to data holders, but the expertise on data particularities is still closer.	May result in an unscalable approach. Extra technical solutions are required to provide external datasets linkability.
2	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.

Priority	Scenarios	Pros	Cons
			Extra technical solutions are required to provide external datasets linkability.
3	Integration of datasets at EU 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.
4	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.

7.9 Comments and considerations on Article 50 of EHDS proposal, on "Secure Processing Environments"

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. 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.
(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[43]. 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.

[43] "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/

Text in EHDS proposal	Comments and considerations
(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.
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 "5.2.2 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.

Text in EHDS proposal	Comments and considerations
	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[44] (Cybersecurity Act) and Article 24 Directive (EU) 2022/2555[45] (NIS2) Cloud Infrastructure Service Providers Europe Code of Conduct for cloud infrastructure service providers[46], 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.

7.10 Recommendations on results extraction for secure processing environment organisation

Priority	Scenarios	Pros	Cons
1	Results export audit manually operated	Higher precision in the audit of the results to be exported	Non-scalable approach. Not applicable to the export of the federated analysis partial results.
2	Computer-based results export audit (e.g., AI assisted)	Higher scalability. Can be used to guarantee the	AI solutions yet to be widely tested Possible false positives/negatives

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

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

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

		validation of federated analysis	(results that should be approved marked as non-exportable or the other way around)

7.11 Recommendations for data access for reproducibility

Priority	Scenarios	Pros	Cons
1	Generation of synthetic data similar with same patterns as original data	Equivalent to anonymised data, with higher level of security.	Technologies to generate fully comparable synthetic data yet to be widely tested.
2	Access to a subset of the original data (after permit granted to the HDAB)	Subset of data can be easily controlled	May be difficult to provide a representative sample.
3	Disclose anonymised version of the original input datasets	Easy setting to disclose the original data	May prevent to actual reproducibility
4	Access to all the original data (after permit granted to the HDAB)	Easy setting to reproduce the results	Currently not considered in the actual EHDS data access models.

7.12 Recommendations for node management services organisation

Priority	Scenarios	Pros	Cons
1	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.
2	External auditing by Core Platform	No burden on NCPs to perform the auditing.	Only external auditing expected, e.g., only intrusion tests.
3	Self-reported node auditing	No extra burden on Core Platform	High trust requirements to the NCPs. Extra burden on NCPs.

7.13 Recommendations for Authentication and Authorisation Infrastructure (AAI) architecture

Priority	Scenarios	Pros	Cons
1	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[47].	Extra complexity of the AAI system to guarantee the interoperability between MS systems.
2	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

7.14 Recommendations for support and training services organisation

Priority	Scenarios	Pros	Cons
1	HDABs offer support to its users, which is coordinated with the EU Core Platform	Higher availability of support services, closer to the users. A central knowledge hub may help to centralise common issues.	High coordination burden when support implies actors from more than one MS.
2	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.

7.15 Recommendations for financial services architecture

Priority	Scenarios	Pros	Cons
1	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.

[47] 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

Priority	Scenarios	Pros	Cons
2	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
3	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.

8 Closing remarks

The EHDS regulation proposal aims to create a secure and protected digital space for sharing health data in the EU. One of the main objectives of the EHDS is to facilitate the secondary use of health data for research, innovation, and public health purposes while ensuring the protection of individuals' privacy and personal data. The European Commission, via this regulatory proposal, aims to establish a legal framework for the secondary use of health data that is transparent, trustworthy, and respects individuals' rights and freedoms. The EHDS proposal represents a significant step towards the secure and responsible use of health data for the benefit of society while ensuring individuals' privacy and data protection rights are respected.

Establishing a circle of trust is crucial for a successful secondary use of health data, and technology plays a significant role in its implementation. Trust is vital for citizens to willingly share their health data for the public good, for decision-makers to make informed choices, and for data users to access credible and useful information. Building a network of trust, cooperation, and transparency among all parties involved requires clear rules and guidelines for data collection, processing, storage, use, and sharing, with a focus on respecting privacy and confidentiality. Technology implications include informing citizens about data security measures, implementing social participation mechanisms for discussing consent and data ownership, and utilizing technology to safeguard data and ensure secure collection, storage, and sharing. For public health authorities, technology facilitates access to accurate and reliable health data for policymaking and decision-making, ensuring transparency and trustworthy methods for data collection and analysis. Data users rely on technology to access accurate and high-quality health data, with metadata providing details on data quality variables. Technology enables secure data handling and communication, supporting transparency, collaboration, and solutions oriented toward the common good across all stakeholders involved in the use of health data.

This TEHDAS deliverable analyses different architecture possibilities for EHDS for secondary use technological substrate, considering implications for data subjects, users, holders, and health data access bodies. In the annexes, several guidelines can be found that are meant to offer Member States more tools and resources when making choices that suit their goals and meet the expectations of their citizens. Member States across the EU have different needs, challenges, and priorities – which means that "one size fits all" solutions for EHDS for secondary use would be of little value, and a negation of European plurality. Cooperation, dialogue, flexibility, and compromise will be needed, and should not be considered a hindrance to effective implementation. While many elements of EHDS may be up for negotiation, there are others that are firmly non-negotiable – such as transparency, security, and integrity. According to the European Commission's digital targets for 2030, "*digital technologies should protect people's rights, support democracy, and ensure that all digital players act responsibly and safely*". The public good (and European values of personal rights and the rule of law) should be at the heart of EHDS.

9 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 17: 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	Natural or legal person who has lawful access to personal or non-personal electronic health data for secondary use;	EHDS proposal regulation. COM(2022) 197
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	Impact assessment report of the EHDS reg. SWD(2022) 131 final PART 1/4

Term / Acronym	Definition	Source
	infrastructure (in the policy option 3, centralised architecture).	
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 used in this document	EHDS proposal regulation. COM(2022) 197
Pseudonymisation	Processing of personal data in such a manner that	GDPR

Term / Acronym	Definition	Source
	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;	
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	DGA /EU) 2022/868, Article 2.

Term / Acronym	Definition	Source
	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;	
Trusted Research Environment (TRE)	Equivalent to Secure Processing Environment but with a wider governance framework defined by the Health Data 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.	Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems, NHS, 2021.

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

Annex A Guidelines for national dataset catalogues publicly available to register and facilitate the discovery of health datasets available for secondary use.

A.1 Use Case Description: publication of national datasets metadata catalogues and search systems.

The use of the national dataset catalogues publication systems foresees the creation, gathering and organisation of the metadata descriptors of the existing datasets available in the member states, covered in the Article 33 datatypes to be included in the European Health Data Space for secondary use, to create a single national datasets catalogue, that will the synchronise with the EU Datasets Catalogue, Article 57.

A.2 General considerations

The setting up of any operational network of metadata catalogues that can exchange information requires careful planning, implementation, and maintenance to ensure that it remains effective and efficient over time. Several essential technical, process and governance requirements need to be considered.

A.3 Legal and Regulatory Considerations

The Article 37 of the European Health Data Space (EHDS) proposal on *Tasks of health data access bodies*, at paragraph (1)(Q)(i) includes the obligation of making public a national dataset catalogue. This is intended as a collection of dataset descriptions, accessible through an online portal, arranged in a systematic manner to be user oriented. The dataset catalogue shall include details about the source and nature of electronic health data, in accordance with Articles 56 and 58 of the EHDS proposal, and the conditions for making electronic health data available.

The Health Data Access Body (HDAB) should provide information about the available datasets and their characteristics so that data users can be informed of elementary facts about the dataset and assess their possible relevance to them. For this reason, each dataset should include, at least, information concerning the source, nature of data and conditions for making data available. The national dataset catalogue shall also be made available to single information points under Article 8 of the Data Governance Act[1] (Regulation (EU) 2022/868).

An EU Datasets Catalogue (Article 57) should be further established: to facilitate the discoverability of datasets available in the EHDS; to help data holders to publish their datasets; to provide all stakeholders and the public with information about datasets placed on the EHDS (such as quality and utility labels, dataset information sheets); to provide the data users with up-to-date data quality and utility information about datasets.

The health data access bodies tasks concerning the dataset catalogue are also related to the duties of data holders, which pursuant to Article 41 of the EHDS proposal are obliged to make electronic health data available under Article 33 or under other Union law or national legislation implementing Union law, cooperating in good faith with the health data access bodies, where relevant. Specifically, about the catalogue (Article 41, paragraph 2), the data holder shall communicate to the health data access body a

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52020PC0767

general description of the dataset it holds in accordance with Article 55 of the same regulation, which provides rules for the dataset description.

Pursuant to Article 55, in fact, the HDABs shall inform through a metadata catalogue the data users about the available datasets and their characteristics (including source, scope, nature of electronic data) and the conditions for making electronic health data available. The minimum information elements that data holders will provide for datasets and their characteristics are going to be defined by means of implementing acts.

A.3.1 Data protection

The information contained in the national dataset catalogue is a compendium of metadata descriptors that do not contain personal data. In any case, there should be an enough information security level to guarantee the integrity of the catalogue contents, to avoid possible supplantation of catalogues, i.e., misleading catalogue entries that redirect to malicious end points, that might try to steal data users information when pointing data users to the location of specific dataset(Data Governance Act, Art. 11(5)).

A.3.2 Authorisation, authentication, and identification

The communication between data holders and health data access bodies, or between health data access bodies required to create and maintain the national datasets catalogue will be identified at institutional level, so their authentication will be based on the end points of the communication and not the individuals that initiate such communication.

A.4 Organisational and Policy Considerations

Setting up a network of metadata catalogues requires collaboration and communication between participating organisations. Multiple data holders connecting to a HDAB to create and manage the metadata records of their datasets and an EU central data catalogue harvesting the national datasets metadata from the HDABs involves several technical, organisational, and governance challenges. It is essential to establish effective communication channels and to organise regular meetings to discuss any issues that arise. Governance and policies should be established to manage the network, including decision-making processes and policies around data quality and security.

The example of the implementation of the INSPIRE Directive and the establishment of the EU Open Data Portal and the European Data Portal could be cited. One can argue that health datasets should be also discoverable in cross domain data portals (https://dataeuropa.gitlab.io/data-provider-manual/).

In this regard, DCAT-AP[2] (Data Catalogue Application Profile) is a metadata standard developed by the European Commission for describing public sector datasets in a machine-readable format. It is an extension of the W3C's DCAT[3] (Data Catalogue Vocabulary) standard, which provides a set of guidelines for publishing structured

[2] 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/release/11

[3] https://www.w3.org/TR/vocab-dcat-2/

metadata about datasets in a standard format that can be easily discovered and accessed.

The DCAT-AP standard was designed to support interoperability between different data portals and catalogues in the public sector, making it easier for users to discover and access public sector datasets. It provides a common vocabulary for describing datasets, including information such as the dataset's title, description, keywords, distribution formats, licensing information, and more.

The European Commission played a central role in the development and promotion of the DCAT-AP standard. It has been actively involved in the development of the standard since its inception, working with stakeholders from across the public and private sectors to ensure that it meets the needs of the wider data community. The Commission also provided funding for projects aimed at implementing the standard in different contexts, such as the EU Open Data Portal and the European Data Portal. DCAT-AP standard is an important tool for improving the discoverability and accessibility of public sector datasets, and the European Commission has played a key role in its development and adoption.

Some EU countries have created their own DCAT profiles to adapt the standard to their specific needs and requirements. These country-specific profiles are based on the DCAT-AP standard but may include additional elements or modifications to meet local needs.

One can consider that facilitating the discovery of health datasets would require creating a Health DCAT profile to address the specificity of health data. For example, the health DCAT profile could include additional elements such as the "DQV" (Data Quality Vocabulary) for describing data quality or additional elements for describing specific health data related services (e.g., the secure processing environments capabilities to process such data).

A.4.1 Enablers for implementation

If a Health DCAT profile should be designed, a governance structure would be necessary to maintain it. Engaging stakeholders from the outset, including data holders and researchers, will be important for ensuring that the metadata standard in use in the catalogues meets their needs and priorities. Sufficient funding and resources will be necessary to ensure that the network of metadata catalogues, i.e., national datasets catalogues at the HDABs and the EU Datasets Catalogue, can be implemented effectively and sustained over the long term. A governance structure would be necessary to ensure that the network is managed effectively and transparently, with clear roles and responsibilities for each organisation involved, enabling more effective sharing and discoverability of data.

A.4.2 Quality standards and validation

The governance structure would be responsible that the network achieves its data-driven goals. Quality standards could be defined and serve as operational frameworks. Validation tools could be implemented and support the entire community in improving and reinforcing the reuse of data.

A.4.3 Education, training, and awareness

Undoubtedly, it will key to build capacity within national dataset catalogues to ensure that data experts have the necessary skills and resources to manage metadata and participate in the network. Education, training, and awareness are crucial for the discovery and effective use of health datasets available for secondary use, which can provide valuable insights into the health of populations and inform research, policy, and clinical decision-making. Education should inform individuals about the availability of health datasets and potential benefits of using them, while training should provide practical experience and guidance on how to effectively use and interpret these datasets. Awareness raising is also important to ensure the uptake and sustainability of these national dataset catalogues, as well as to inform data users about the ethical use and proper governance of these datasets.

Capacity building on the discovery of health datasets available for secondary use can be achieved through a variety of methods, including online tutorials and webinars, in-person training, collaborative workshops, networking opportunities, and access to support and resources. A comprehensive approach that combines multiple methods and resources is necessary to ensure individuals have the necessary knowledge and skills to effectively use and interpret these datasets.

A.5 Semantic Considerations

One of the key challenges of data discovery and metadata exchange is the lack of standardised vocabularies for describing datasets. Without a common vocabulary, it is difficult for different data portals and catalogues to exchange metadata and for users to discover relevant datasets. DCAT-AP addresses this challenge by providing a standardised vocabulary for describing datasets, including metadata such as titles, descriptions, keywords, distribution formats, licensing information, and more. DCAT-AP addresses the interoperability challenge by providing a common metadata format that can be easily exchanged between different portals and catalogues. This makes it easier for users (i.e.: humans and machines) to discover and access datasets across different portals and catalogues, and it also facilitates the integration of datasets from different sources. Moreover, DCAT provides a standardised vocabulary for describing relationships between datasets, such as those based on common themes or geographical regions, it is easier for users to discover relevant datasets and to understand how different datasets relate to each other. The DCAT metadata standard is typically published in RDF format, which is a core component of the linked data stack meaning that metadata is not only readable but also actionable by machines as its content comes with meaning. If implemented in a linked-data platform, DCAT offers a semantic layer to discover data, especially important when deepening in the datasets contents, a still open problem, for example, when aiming to query such datasets at variable level.

In this regard, one can cite the extensive list of semantic structures, such as controlled vocabularies, i.e., authority tables: dataset types, access right, etc., maintained by the EU Publications office and in use in the EU open data portal. It will be a governance requirement for the community to define and maintain the semantic structures necessary for facilitating the discoverability and reuse of health data. The ability to perform a

semantic search of all data in the network of catalogues should provide enhanced findability, access, interoperability, and re-use of health data.

A.5.1 Metadata standards

One of the first technical requirements relates to the use within the network of a common metadata standard: the mapping of metadata elements between the different catalogues should be established to ensure that the metadata is consistent across the network. In other words, it is essential to establish and rely on a common metadata standard across all participating catalogues to ensure seamless exchange of information. It is also a prerequisite condition to achieve semantic interoperability by making metadata machine actionable: same metadata standard, same shared common vocabularies, and its associated links to ontologies give the capabilities to the machines for interpretation, inference, and logic. DCAT-AP, as it is widely used in the EU and considering it was designed to support interoperability between data catalogues, should be the go-to standard. Furthermore, the design of a Health DCAT AP would address specific needs.

A.6 Technical Considerations

A.6.1 Communication protocols

The catalogues should be designed to be interoperable, so that they can communicate and exchange data with each other using standard protocols. Data can be metadata records or search queries. The supported formats by the network must be defined and accepted by all participating organisations. This could include protocols such as OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), or SRU (Search and Retrieval via URL).

A.6.2 Metadata as a service

Moreover, the network will require a reliable and robust infrastructure that includes servers, network hardware, and software applications that can support communication and the exchange of metadata between catalogues. The technical infrastructure should be established and maintained with service level objectives (i.e.: SLAs, SLOs, and SLIs.).

A.6.3 Quality and Security

Concerning some of the process requirements, procedures for the control of metadata ownership and quality will help ensure that the metadata shared between catalogues is accurate, complete, and up to date. Authentication and authorisation mechanisms should also be implemented to ensure that only authorised users can access and modify the metadata, as well as to maintain the integrity of the multiple catalogues involved in such a system.

Annex B Guidelines for management systems to record and process data access applications, data requests and the data permits issued, and data requests answered.

B.1 Use Case Description: a system to manage data access applications.

The use case starts once the users of the HealthData@EU infrastructure have been able to discover and locate the data (individual level data sets containing personal data, aggregated statistics) required to perform their analyses (discovery step). This use case consists of requesting the permit to access such data (data access applications for individual level data, data requests for statistics or aggregated data) to one through the mechanisms, e.g., an EU HealthData@EU data access portal.

In addition to the request for the data access permits, data users and other HealthData@EU actors may have access to the repository of the permits issued and rejected applications or requests processed in the HDABs, including both approved and rejected.

B.2 General considerations

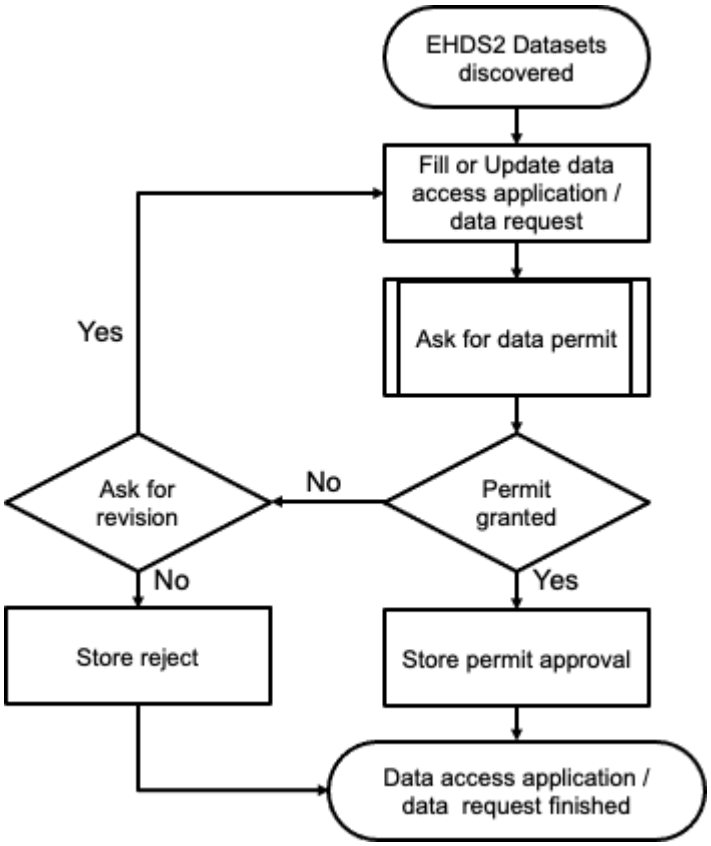


Figure 1: Data access application / Data request process

For the sake of simplicity, in this guideline there will be no explicit distinction between the process to manage a data access application and a data request, beyond the

contents and format of the digital object that represent such applications/requests to be provided to the HDAB to decide on them.

Figure 1 contains a basic data flow depicting the process of a data access application or a data request. The different steps define the necessary functions of the management system that will take care of it:

- Fill/Update the data access application / data request: capture the required information the HDAB designated actors will be required to decide on such application/request (scientific committees, ethical committees, others)
- Ask for data permit: this is the key process that implies the distribution of the information included in the applications/requests to the data access committees or equivalent bodies (the *authorisers*), within a country or cross-border if necessary. The authorisers will review the provided information and decide accordingly, asking in some cases for clarification.
- If a data permit is not granted, it might be required to review it for further review, which will imply a new loop in the request process. If no revisions are allowed, the rejection is stored for public information.
- If data permit is granted, it is stored for public information and for the continuation of the data provision.

B.3 Legal and Regulatory Considerations

According to the EHDS proposal, the common framework for secondary use has the aim to reduce the fragmentation and barriers for access to data for secondary use, including for cross-border accesses. Member States will have to set up a health data access body for secondary use of electronic health data and ensure that electronic data are made available by data holders for data users. The articles of the EHDS proposal which are relevant as legal bases for the contents of this document are mainly the Art. 37, which at paragraph 1 (k) concerns the management system to record and process data access applications, data requests and data permits, and the articles 45 and 54, on data access applications and the mutual recognition of data permits, respectively.

Article 37 on Tasks of health data access bodies, provides the tasks that health data access bodies shall carry out the maintenance of a management system to record and process data access applications, data requests and the data permits issued and data requests answered, providing at least information on the name of the data applicant, the purpose of access the date of issuance, duration of the data permit and a description of the data application or the data request.

The data access applications are based on Article 45 as follows:

1) Any natural or legal person may submit a data access application for the purposes referred to in Article 34.
2) The data access application shall include:
   a) a detailed explanation of the intended use of the electronic health data, including for which of the purposes referred to in Article 34(1) access is sought;
   b) a description of the requested electronic health data, their format and data sources, where possible, including geographical coverage where data is requested from several Member States;

c) an indication whether electronic health data should be made available in an anonymised format;
d) where applicable, an explanation of the reasons for seeking access to electronic health data in a pseudonymised format;
e) a description of the safeguards planned to prevent any other use of the electronic health data;
f) a description of the safeguards planned to protect the rights and interests of the data holder and of the natural persons concerned;
g) an estimation of the period during which the electronic health data is needed for processing;
h) a description of the tools and computing resources needed for a secure environment.

3) 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 referred to in Article 52, which have been identified in the data access application. For requests to access electronic health data from more than one Member States, the health data access body shall notify the other relevant health data access bodies of the receipt of an application relevant to them within 15 days from the date of receipt of the data access application.

The data permits issued by the health data access bodies are an administrative decision defining the conditions for the access to the data. Related to data permit is also the concept of mutual recognitionThe Article 54 of the EHDS proposal on *Mutual recognition* recites:

1) When handling an access application for cross-border access to electronic health data for secondary use, health data access bodies and relevant authorised participants shall remain responsible for taking decisions to grant or refuse access to electronic health data within their remit in accordance with the requirements for access laid down in this Chapter.

2) A data permit issued by one concerned health data access body may benefit from mutual recognition by the other concerned health data access bodies.

B.3.1 Data protection

No personal **health** data is transmitted as part of the data permit application or the data requests processing.

Patients' personal health data is not managed in these systems. It will be necessary to manage the personal data of the requesters and the personal data of the data access committees (*authorisers*) that will grant or reject the permits.

It is important to consider the protection of the protocols, data management plans, or any other companion documentation provided in the data access request for its further consultation.

B.3.2 Authorisation, authentication, and identification

The access to the systems in charge of processing data access applications or data requests, data users will be required to be individually identified.

B.4 Organisational and Policy Considerations

B.4.1 Process transparency

To raise trust between the participant organisations in the HealthData@EU, transparency in the data access applications processing and permit provision will be key, being the ultimate goal the mutual recognition of health data access bodies data permit concessions covered in Art.54. It is worth noting that at the country level, ethics committees usually operate under mutual recognition, due to the harmonised interpretation of the regulatory frameworks. In the cross-border scenario, this is not expected to be the case in the short term.

B.4.2 Quality standards and validation

The process's transparency will benefit from using quality standards on its implementation. For this transparency purpose, the following elements will be necessary:

- The definition of health data access bodies authorisation processes on each country, including well documented permit chains.
- The Identification of actors that participate in the permit chain.
- Public access of the permits issued and rejected, including the detailed information about the deliberations.

It will be important to define a clear procedure for these countries that do not have a health data access body operational, but existing data holders may be ready to provide data or to process statistics.

B.4.3 Cross-border data access applications / data requests operation

The cross-border operation of the data access applications / data requests will require a distribution mechanism of the digital objects required to decide on them.

It is expected that the cross-border operation is assisted by the EU core platform. In this way the EU core platform will directly receive the digital objects with the necessary information, generated at a EU or national level portal, and will distribute such objects to the HDABs containing the datasets or data statistics requested. The EU core platform will then receive the decisions taken by the HDABs and distribute them back to the requesters.

B.4.4 Education, training, and awareness

Effective management systems for health data access require a comprehensive approach that involves education, training, and awareness. This includes providing HDAB staff and data users with the necessary knowledge and skills to understand the importance of data management and the protocols in place to safeguard patient privacy and confidentiality. Ongoing training should be provided to keep staff up to date with changes in regulations and existing best practices. Creating a culture of accountability and responsibility among actors can help ensure compliance and reduce the risk of data breaches.

In addition to education and training, awareness is crucial for all parties involved in the data access applications / data request process (data users, HDAB personnel, authorisers). Data users should be aware of the steps they need to follow, the information they need to provide, and the conditions for accessing the data once their request is approved. HDABs should also create awareness among data users of the importance of protecting patient privacy and confidentiality. The data access applications / data requests management systems interface should be designed for ease of use, and procedures should be made transparent and understandable for data users. By implementing these measures, organisations can create a robust management system that ensures the accuracy and integrity of health data while safeguarding patient privacy and confidentiality.

B.5 Semantic Considerations

B.5.1 Data access applications

To facilitate the management of the data access application decision process, especially in a cross-border scenario, it will be necessary that the digital objects that encapsulate data access applications follow a standardised schema, in well-known format such as XML, JSON or turtle if using RDF vocabulary.

Expected contents (in parentheses those that are explicitly covered in the EHDS Art.45):

1. Applicant ID: digital ID of the data user responsible of the data access application, should be provided by an AAI system.
2. Persistent IDs of the dataset(s) (Art.45(2)(b)): unique and persistent identifiers of the datasets applied, gathered from the metadata catalogue.
3. Project protocols (Art.45(2)(a)): documentation format standard to be defined. Translation services will be required for this content in cross-country data access applications. The project protocol should include the justification of the need to access pseudonymised data if the case (Art.52(2)(d)).
4. Data Management Plan (Art.45(2) (c, e-h)): documentation format to be defined, ideally in a Machine Actionable DMPs format under study (Argos).
5. Authorised data users: ID's list of the authorised data users associated to such application (e.g., the members of the research team), provided by an AAI system.
6. Computation requirements: the list of the expected computational requirements necessary to analyse the datasets applied, ideally in an Infrastructure as a Code (IaC) format.
7. Tools requirements: the list of the foreseen analysis tools and libraries required to perform the analyses. Ideally it should be the persistent ID's list of the tools / libraries certified repositories.
8. Approval status: list of HDABs identifiers, its decision regarding the data access application (None, Approved, Rejected, Revision) and the reasons for the decision.

9. Other documentation: any other information that might be required for the data access application, e.g., information for specific requirements of some data sets or access to open data sets.
10. Dates: listing of dates reflecting the events on the application (submission, decisions, updates, etc.)

These contents should be summarised (for example, removing details of the project protocols or the data management plants) unless stated when providing historical data of the data access applications processed in the EU Core platform / HDAB.

B.5.2 Data requests

To facilitate the management of the data requests decision process, especially in a cross-border scenario, it will be necessary that the digital objects that encapsulate data access applications follow a standardised schema, in well-known format such as XML, JSON or turtle if using RDF vocabulary.

Expected contents (in parentheses those that are explicitly covered in the EHDS Art.47):

1. Applicant ID: digital ID of the data user responsible for the data request, should be provided by an AAI system.
2. Description of the result expected from the health data access body (Art.47(2)(a)): narrative description of the metric required. Translation services will be required for this content in cross-country data access applications.
3. Description of the statistic's content (Art.47(2)(b): mathematical definition of the expected results, according to the data available in the HealthData@EU datasets.
4. Justification of the results: description of the motivation to obtain the requested data. It might include a project protocol. Translation services will be required for this content in cross-country data access applications.
5. Approval status: list of HDABs identifiers and its decision regarding the data request (None, Approved, Rejected, Revision)
6. Dates: listing of dates reflecting the events on the application (submission, decisions, updates, etc.)

These contents should be summarised (for example, removing details of the project protocols or the data management plants) unless stated when providing historical data of the data requests processed in the EU Core platform / HDAB.

B.5.3 Cross-border APIs

Cross border APIs will be necessary to facilitate the interoperability between the HDABs and the EU Core platform when managing cross-border data access applications or data requests. The verbs shown represent the data access application management, equivalent verbs will be used for data requests (changing the references to `data_access_application` to `data_request`).

Note that no explicit mention to the ID tokens of the requesters is provided. This is expected to be managed at session level.

EU core platform verbs

submit_data_access_application(data_access_application_info):data_access_application_id
Description Starts a new cross-border data access application in the EU Core platform.
Inputs: • data_access_application_info Digital object containing the required information of the data access application
Outputs: • data_access_application_id Identifier of the created data access applications, to be used in the following operations. NULL otherwise

update_data_access_application(data_access_application_id, new_info):update_status
Description Modifies the data application digital object on an existing data access application in course
Inputs: • data_access_application_id Identifier of a previously initiated data access application • data_access_application_info: digital object representing the data access application with the updated information
Outputs: • status Results of the operation (Data access application not found, Update OK, Update NOK, others)

check_data_access_application_status(data_access_application_id):data_access_application_status
Description Returns the status of an existing data access application
Inputs: • data_access_application_id Identifier of a previously initiated data

access application
Outputs: • `data_access_application_status` Approval status list of the data access application, NULL if the data access application is not found

`update_data_access_application_decision``(data_access_application_id, hdab_id, decision, decision_info):application_decision_status`
Description Updates a data access application providing the decision of a given HDAB
Inputs: • `data_access_application_id` Identifier of a previously initiated data access application • `hdab_id` identifier of the HDAB providing the decision • `decision` codification of the decision • `decision_info` codification of the reasons of the decision
Outputs: • `application_decision_status` Status of the data access application decision update (Decision update OK, Decision update NOK), NULL if the data access application is not found

`get_data_access_applications_list``(filters):data_access_applicaitons_list`
Description Provides the list of the existing data access applications based on a set of filters. To be used to scrutinise the historical database of data access applications treated in the EU Core platform
Inputs: • `filters` Expression indicating the filtering of the data access applications based on dates, datasets requested or other information
Outputs: • `data_access_applicaitons_list` List of summarised data access applications processed in the EU Core platform that passed the filter.

HDAB interface

submit_local_data_access_application(data_access_application_info):data_access_application_id
Description Starts a new single country data access application in the HDAB
Inputs: • data_access_application_info Digital object containing the required information of the data access application
Outputs: • data_access_application_id Identifier of the created data access applications, to be used in the following operations. NULL otherwise

submit_eu_data_access_application(data_access_application_id, data_access_application_info):data_acces_application_status
Description Starts a new EU level data access application in the HDAB. To be call by the EU **submit_data_access_application** of the EU core platform
Inputs: • data_access_application_id Data access application ID generated in the EU core platform • data_access_application_info Digital object containing the required information of the data access application
Outputs: • data_acces_application_status OK on correct creation of the data application, NOK otherwise

update_data_access_application(data_access_application_id, new_info):update_status
Description Modifies the data application digital object on an existing data access application in course. Can be used for local data access applications initiated in the HDAB or by the EU core platform
Inputs:

• `data_access_application_id` Identifier of a previously initiated data access application • `data_access_application_info`: digital object representing the data access application with the updated information
Outputs: • `update_status` Results of the operation (Data access application not found, Update OK, Update NOK, others)

`check_data_access_application_status``(data_access_application_id):data_access_application_status`
Description Returns the status of an existing data access application. Can be used for local data access applications initiated in the HDAB or by the EU Core platform
Inputs: • `data_access_application_id` Identifier of a previously initiated data access application
Outputs: • `data_access_application_status` Approval status list of the data access application, NULL if the data access application is not found

`get_data_access_applications_list``(filters):data_access_applications_list`
Description Provides the list of the existing data access applications based on a set of filters. To be used to scrutinise the historical database of data access applications treated in the HDAB.
Inputs: • `filters` Expression indicating the filtering of the data access applications based on dates, datasets requested or other information
Outputs: • `data_access_applications_list` List of summarised data access applications processed in the HDAB that passed the filter.

B.6 Technical Considerations

B.6.1 Secure Access

Secure access implies the use of reliable authentication and authorisation infrastructure (AAI) solutions, complemented with secure communication channels. Desirably, it should support federated AAI assisted by the EU core platform, to facilitate the reuse of existing AAI solutions of the different authorised participants. AAI solutions should be compatible with the secured API-based interfaces.

The AAI should be necessary to guarantee the identity of the data users submitting the data access application or data requests as well as the individuals in charge of taking the decisions at the HDAB (or HDABs).

Multiple factor authentication should be mandatory to increase the security levels of the access.

Regular penetration tests should be included as part of the security protocols.

B.6.2 Request side elements

The request side typically corresponds to interfaces at the EU core platform, for regular cross-border applications and requests, and eventually the HDAB, for single country applications and requests.

The request side elements will be composed by a web application front-end capable of interacting with the AAI, providing the necessary forms to gather all the information present in a data access application or a data request.

The backend of the request side will contain the database of the applications/requests and the logic to forward them to other actors involved in the processing through the APIs.

B.6.3 Granting side elements

The grant side corresponds to the interfaces and control logic that will be deployed at HDAB level to coordinate the distribution of the applications or requests to the individuals or committees in charge of the authorisation (the "authorisers") and gathering the decisions.

An interactive web application will be provided to the authorisers to manage the documents provided in the applications/requests as well as to submit the documents supporting the decisions.

B.6.4 Interaction of both request side and grant side

The interaction of both sides will be operated by two business processes management systems one operating at EU core platform level, and another on the HDAB level (one per country), that will connect the database of applications/requests (local or remote), its notification to the granting side, the notification of the decisions (local or remote) and the management of the reviews and updates.

The business process management system will oversee archiving the final results of the application / request process, storing the data permits or the rejection digital objects at the EU core platform and/or the HDABs accordingly, for its further processing.

B.6.5 Security

Business process management system to facilitate the authorisation chain management including:

- Notification system for the data permit actors
- Document management for the companion documentation (to deliberate and to store for further consultation).

Annex C Guidelines for Secure Processing Environments (technical, information security and interoperability requirements)

C.1 Use Case Description: secure processing of health data for secondary exploitation

The use case starts once the users of the HealthData@EU infrastructure have obtained a data permit to access the data sets required to perform their analyses (discovery step) and have been assigned one or multiple computation resources in the form of a secure processing environment or where the datasets will be deposited and the analysed.

C.2 General considerations

The term Secure Processing Environment (SPE) is defined in the Data Governance Act, Art.2(14) as "(...) *the physical or virtual environment and organisational means to provide the opportunity to re-use data in a manner that allows for the operator of the secure processing environment to determine and supervise all data processing actions, including to display, storage, download, export of the data and calculation of derivative data through computational algorithms.*"

For the specific purpose of providing the services listed in the Art.50 of the EHDS legislative proposal, the overarching description of the functional capabilities proposed of such a computer system are the following:

- Analysis features to process sensitive data (statistics tools, AI libraries, code versioning systems, others)
- Interactive access[1] (remote desktop, secure shells, others). Optionally, it is expected that some of the SPEs provide an API-based access for federated analysis.
- Strong access control (data holders for data deposition, data users for data analysis, system administrators for SPE management)
- Communications control (data imports, data exports, outbound communications)
- High security requirements
- Clearly defined operational protocols

In terms of the overall use of the HealthData@EU infrastructure, the use of SPE is located once the data user has been granted access to a given dataset or datasets and a data permit has been issued by the HDAB used for that purpose. A general SPE lifecycle is depicted in Figure 1, and have the following steps:

- **Environment creation**: once the health data access body generated the permit, the assigned SPE operator should create an isolated environment instance, according to the computing resources and tools requirements explicit by the data user in the data access application. In the current IT context, this provision is

[1] Interactive access refers to those computer systems' interfaces which accept input from the user as it runs, e.g., a window system where users point and click or a command-line terminal where users write the commands to be executed.

embodied in a virtual machine, deployed indistinctly in a dedicated cluster or a computing cloud.

- **Data upload**: depending on the organisation of the SPE/data holder interfaces, the SPE operator[2] pulls data from the data holder(s) or the data holder(s) pushes the data to the assigned SPE storage location.
- **Data analysis**: using the tools in the environment creation, the data user processes the deposited data to find the insights he or she is looking for.
- **Results extraction**: once a data user obtains results (partial or final), he or she should request to download such results outside the SPE premises. This process implies the control mechanism of what data that leaves the SPE.
- **Environment decommissioning/archival**: once the data user finishes his or her project or the duration of the data permit finalises, the environment may be destroyed (including all its contents) or archived upon request for further use under new conditions (e.g., sample data access for scientific reproducibility) or new data permits (e.g., new derived projects)

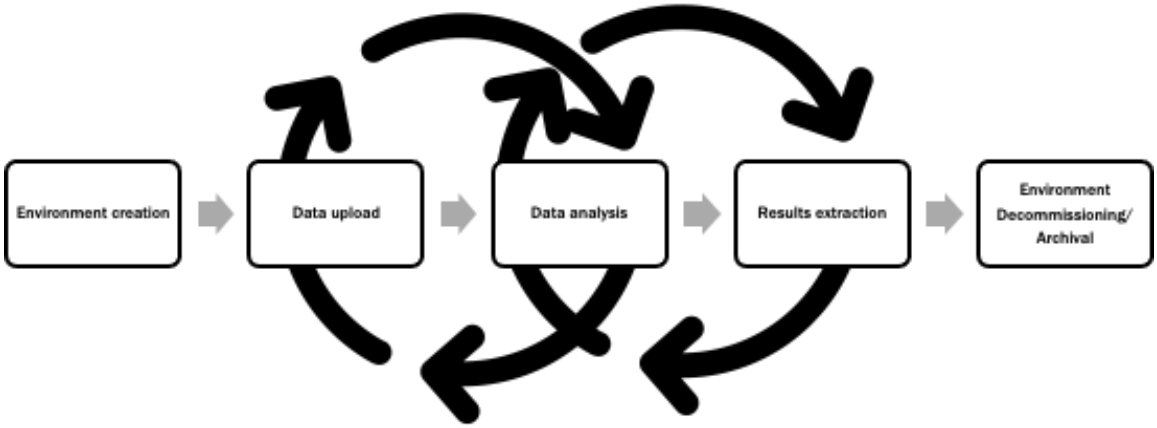


Figure 1: Secure Processing Environment lifecycle

Please note that the present SPE life cycle refers to a scenario where a single SPE is assigned to a data permit, and all the data sets allowed in that permit are then pooled together in such a SPE. There might be cases that for technical reasons, e.g., amount of data to be transferred, or political reasons, e.g., mandatory limitations to transfer data between specific jurisdictions or countries, a single data permit indicates that multiple SPEs are assigned and the specific mapping on which data sets should be transferred to which SPE. In this specific situation, the SPE life cycle is still applicable, but it will be replicated according to the number of SPEs involved.

Finally, it is also important to note that computing systems used for the processing of highly sensitive data exist, being the ones used in the military and the one used in banking one of the closest examples. In any case, for the sake of simplifying SPEs to be used in the EHDS context from the previous one, it is recommended that the regulation focuses on attaching general cybersecurity regulations (EU Cybersecurity Acts and

[2] SPE operator is the company or organisation in charge of providing the SPE services to the Health Data Access Bodies.

national Cybersecurity Acts), usually applied to the banking system, rather than the regulations for classified information used in the military or the intelligence context.

C.3 Legal and Regulatory Considerations

The EHDS Regulation provides the legal basis in accordance with Articles 9(2) (g), (h), (i) and (j) of Regulation (EU) 2016/679 for the secondary use of health data, establishing the safeguards for processing, in terms of lawful purposes, trusted governance for providing access to health data (through health data access bodies) and processing in a secure environment, as well as modalities for data processing, set out in the data permit. At the same time, the data applicant should demonstrate a legal basis pursuant to Article 6 of Regulation (EU) 2016/679, based on which they could request access to data pursuant to this Regulation and should fulfil the conditions set out in Chapter IV. More specifically: for processing of electronic health data held by the data holder pursuant to the EHDS, this regulation proposal creates the legal obligation in the sense of Article 6(1) point (c) of Regulation (EU) 2016/679 for disclosing the data by the data holder to health data access bodies, while the legal basis for the purpose of the initial processing (e.g. delivery of care) is unaffected. The EHDS also meets the conditions for such processing pursuant to Articles 9(2) (h),(i),(j) of the Regulation (EU) 2016/679. Further, the EHDS proposal assigns tasks in the public interest to the health data access bodies (running the secure processing environment, processing data before they are used, etc.) in the sense of Article 6(1)(e) of Regulation (EU) 2016/679 to the health data access bodies, and meets the requirements of Article 9(2)(h),(i),(j) of the Regulation (EU) 2016/679. Therefore, the EHDS provides the legal basis under Article 6 for the activities mentioned above and meets the requirements of Article 9 of GDPR on the conditions under which electronic health data can be processed.

C.3.1 Data protection - GDPR considerations

As the regulation considers the possibility of using both anonymised and pseudonymised data, it will require that specific GDPR role arrangement between data holders, HDABs and SPEs operators are in place to provide access to the data sets covered in a data permit. Article 51 of the legislative proposal establish the joint controllership of the health data processed, so in the case that the SPEs are provided by a HDAB, this will be understood also as a joint controller. In the case of external SPE providers, these are expected to be arranged as processor of the data on those projects that make use of them.

It would be very important that, to ensure a reliable quality of service to data users, the management of such agreements and any other obligations that might be derived from them is performed seamlessly, reducing as much as possible the data user interaction.

Complementary to the personal data from the datasets to be used in the SPEs, it will also be necessary to take into consideration the personal data of the data users authorised to access to such systems. In the current conception of the HealthData@EU architecture, there won't be necessary for SPE operators to store any data users' personal, only the credentials (non-personal public identification keys of a public key infrastructure service) required to access to the SPE, which will be provided by the

Authorisation and Authentication Infrastructure services. The management of data users' personal data will be the responsibility of the health data access bodies.

C.3.2 Regulatory intensity on service provision

For the sake of the harmonisation of the SPE provision, the implementing acts regulated by the Article 50(4) of the EHDS legislative proposal could possibly cover minimum functional capabilities of such environments, for example the set of basic analysis tools to be installed or the processes to control the data extraction (including in federated analysis scenarios).

In addition, it is expected that, as part of the EU core platform, described in the Article 52(10), there may be also an EU-level SPE, to facilitate the exploitation of the HealthData@EU infrastructure.

C.4 Organisational and Policy Considerations

C.4.1 Enablers for the implementation - Security requirements and certifications

Functional requirements of an SPE will mainly be determined by data user needs rather than compliance needs. Non-functional requirements such as security requirements are however to a higher degree determined by compliance needs. Currently the GDPR is the regulation that HDABs, data users and SPEs must comply with. The requirements defined in GDPR are very flexible as they rely mainly on a risk-based approach. This also means that the level of security may vary depending on what level of risk the data controllers are willing to accept. In the context of the cybersecurity, HDABs may be also obliged to comply with NIS2 Directive (2022/2555), as detailed in Art.(2)(f), and the national transpositions.

Although security and trust are important to data users, they will most likely be helped by clear directions on a minimum level of security that is accepted. Minimum security requirements should be harmonised on an EU-level, considering a framework that relies on existing security standards and specific developed regulations, if needed. It is important to also consider that the specific security requirements should never contradict the local cybersecurity regulations.

In terms of the standards, frameworks and schemes related to security that include requirement sets within security areas that are also relevant for SPEs. Existing SPE-like systems most already use ISO27001, and many are also certified. Therefore, it is relevant to investigate the possibilities to use and build on existing requirement sets when defining detailed security requirements for SPEs. Following table includes a listing of possible security standards to be adopted.

Standard/Guideline	Issuer
ISO/IEC 27001	International Standards Organisation (ISO)

Standard/Guideline	Issuer
Information security management systems[3]	
European Cybersecurity Certification Scheme for Cloud Services (EUCS)[4] - Draft	European Union Agency for Cybersecurity (ENISA)
Building Trusted Research Environments - Principles and Best Practices ("Five safes" report) [5]	Health Data for Research (HDR) UK
Data protection Code of Conduct[6]	Cloud Infrastructure Services Providers in Europe (CISPE)

One aspect to have in mind when choosing or developing requirement sets is that they should be able to verify through testing to make sure that compliance checking, or certification procedures can be done efficiently. This is especially relevant for implementation Art.50(3) where it is stated that the HDAB shall ensure regular audits of the SPEs. When looking into use of existing EU or international standards, frameworks, and schemes these are examples that have been identified that contain relatively specific requirements or controls.

To guarantee the appropriate enforcement of the security framework of selection, it will be necessary to require a certification procedure for all countries, controlled by an EU central body, designated by the EHDS board. Then, the verification of the compliance to such certification should be carried out by an external independent party, authorised by each country, upon request of the health data access body.

C.4.2 Enablers for the implementation - interaction with data holders

To guarantee the proper interaction between data holders and the secure processing environment for the data upload, a clear protocol should be defined to govern the data transfer between these two actors, in terms of who can initiate the communication. Basic uploading from data holders to secure processing environments is desired and covered by Art.50(2). Alternative interfaces where data is pulled from the secure processing environment might be decided at country level, for example to serve use cases where regular data retrieval is expected (e.g., continuous monitoring).

C.4.3 Education, training, and awareness

To ensure SPE effectiveness, education, training, and awareness are vital. SPE operators should provide regular training to their staff and data users to make them

[3] https://www.iso.org/standard/73906.html
[4] https://www.enisa.europa.eu/publications/eucs-cloud-service-scheme
[5] https://zenodo.org/record/5767586#.ZFkV3HZBwQ9
[6] https://www.codeofconduct.cloud/the-code/

aware of the importance of data security and privacy. Their teams should be equipped with the latest knowledge and skills to maintain such SPE as free of risks as possible. Practical scenarios should be included in the training to help staff understand the potential consequences of a data breach and the steps they can take to prevent it.

The SPE system administrators[7] should be trained to conduct regular audits and manual system inspections to identify any potential threats or data breaches. SPE operators should create a culture of continuous learning, training, and awareness to ensure staff remains vigilant and proactive in safeguarding health data. Awareness campaigns can also be directed towards data subjects, informing them about the role of the SPE in protecting their data. By fostering a culture of awareness, organisations can minimise the risk of data breaches and protect patient privacy and confidentiality in the SPE context.

C.5 Semantical Considerations

C.5.1 Data permit interoperability

One of the key elements that will drive the overall data access process within the HealhtData@EU infrastructure is the data permit. The data permit in this context refers to the digital object issued and stored in a health data access body that contains the information described in Art.46(6), as well as other necessary information to guarantee the SPE service provision according to SPE lifecycle. SPE-related information should cover the following elements:

- Identifier of the HDAB issuer.
- Related to data sets: persistent identifiers of the data sets granted, included the dataset locators, public certificates of the data holders where the data set reside. Optionally, persistent identifiers and locators of user provided data.
- Related to analysis tools: narrative description or persistent identifiers of the tools foreseen for the analysis. Optionally, tools locators, public certificates, or equivalent information to verify the tools' location.
- Related to computing capabilities: formal definition of the computing capabilities, ideally an Infrastructure as Code (IaC) format.
- Related to data users: public certificates of the data users authorised to access the data.
- Related to SPE operators: public certificates of the SPE operator in charge of instantiating the actual SPE.

It is worth noting that, in cases that for technical reasons, e.g., amount of data to be transferred, or political reasons, e.g., mandatory limitations to transfer data between specific jurisdictions or countries, a single data permit indicates that multiple SPEs are assigned and the specific mapping on which data sets should be transferred to which SPE.

[7] SPE system administrators are those individuals contracted by the SPE operator whose responsibility is the correct execution of the SPEs, according to the regulation and the contractual bindings with the data users. Their main duties will be administering the computing systems and communication networks to guarantee the availability of the SPE, with a primary focus on the security concerns.

C.5.2 Data upload interfaces

APIs for data transport between data holders and SPEs should be clearly specified. Depending on the directionality of such communications, i.e., data push (upload) from data holders to SPEs and data pull from data holders by SPEs, such API should include the verbs listed in the following tables.

A special case for data user provided datasets is also considered.

This general definition of the uploading/pulling interfaces is compatible with the scenario of having multiple SPEs.

Such interfaces may benefit from using secure transport solutions, such as the eDelivery.

SPE Interface

data_push (data_permit_id, data_holder_id, dataset_id, data):push_status
Description Starts a data upload from a data holder to a SPE. Its termination does not guarantee the full upload
Inputs: • data_permit_id Identifier of the data permit that allows the data upload • data_holder_id Identifier of the data holder that will transmit the dataset • dataset_id Identifier of the dataset to be transmitted • data The contents of the dataset
Outputs: • push_status OK if the data upload was started correctly, NOK if the data upload was not started correctly, WRONG_DATASET if data was not allowed in the permit

data_push_status(data_permit_id, data_holder_id, dataset_id):completed
Description Checks the completion of a data upload
Inputs: • data_permit_id Identifier of the data permit that allows the data upload • data_holder_id Identifier of the data holder that transmitted the dataset • dataset_id Identifier of the dataset to be transmitted
Outputs:

- `completed true` if the data upload correctly, `false` otherwise

`verify_data_set``(data_permit_id, dataset_id, CRC):correct`

Description

Checks the correction of a dataset uploaded or pulled in a SPE

Inputs:
- `data_permit_id` Identifier of the data permit associated to the dataset
- `dataset_id` Identifier of the dataset to be verified
- `CRC` Signature of the original data to verify the copy

Outputs:
- `correct true` if the verification was positive, false otherwise

`data_push` `(data_permit_id, data_user_id, dataset_id, data):push_status`

Description

Uploads a data provided dataset. Its termination implies the full upload

Inputs:
- `data_permit_id` Identifier of the data permit that allows the data upload
- `data_user_id` Identifier of the data user that will transmit the dataset
- `dataset_id` Identifier of the dataset to be transmitted
- `data` The contents of the dataset

Outputs:
- `push_status OK` if the data uploaded correctly, `NOK` if the data was not uploaded correctly, `WRONG_DATASET` if data was not allowed in the permit, `NON_AUTH` if data did not pass the auditing

Data holder interface

`data_pull``(spe_operator_id, data_permit_id, dataset_pid):pull_status`

Description

Pull the data from a data holder to a SPE. Its termination does not guarantee the pull
Inputs: • `spe_operator_id` Identifier of the SPE pulling the data • `data_permit_id` Identifier of the data permit that allows the data upload • `data_holder_id` Identifier of the data holder that transmitted the dataset • `dataset_id` Identifier of the dataset to be transmitted
Outputs: • `pull_status OK` if the data pull was started correctly, `NOK` if the data pull was not started correctly, `WRONG_DATASET` if data was not allowed in the permit

`data_pull_status``(spe_operator_id, data_permit_id, dataset_id):completed`
Description Checks the completion of a data pull
Inputs: • `spe_operator_id` Identifier of the SPE that is receiving the dataset • `data_permit_id` Identifier of the data permit that allows the data upload • `dataset_id` Identifier of the dataset to be transmitted
Outputs: • `completed true` if the data was pulled correctly, `false` otherwise

C.5.3 Data Analysis interfaces

It would be mandatory to offer interactive interfaces to data users to access SPEs, i.e., window-based interfaces remotely available or remote command-line interfaces. This interface will benefit from using interactive analysis tools as well as programmatic tools.

Optionally, it would be necessary that SPEs offer API based analysis interfaces. API interface will be obligatory for those SPEs that offer federated analysis capabilities, i.e., those SPEs that are able to coordinate with other SPEs to analyse data sets without the requirement of moving it out of the premises.

It is out of the scope of the present guideline to describe the analysis in detail, but its general conception should include:

- For non-federated analysis: the verbs to manipulate the data (filter, selection, aggregation, join, etc.) and the verbs to invoke different analysis models.
- For federated analysis: the same verbs to manipulate the data in the different SPEs and the verbs to coordinate the invocation of the analysis models in the

different SPEs involved. For this coordination, multiple approaches are possible (fork and join, map reduce, etc.).

C.6 Technical Considerations

C.6.1 Secure access

Secure access implies the use of reliable authentication and authorisation infrastructure (AAI) solutions, complemented with secure communication channels. Desirably, it should support federated AAI assisted by the EU core platform, to facilitate the reuse of existing AAI solutions of the different authorised participants. AAI solutions should be compatible with the secured API-based interfaces.

Multiple factor authentication should be mandatory to increase the security levels of the access.

Regular penetration tests should be included as part of the security protocols.

C.6.2 Secure computing

For economic reasons, it is considered a service provision based on the allocation of the SPEs in virtual machines. To guarantee the process and data isolation it would be desirable to use advanced processors supporting trusted execution environment (TEE)[8], to facilitate the isolation of the processes and data allocated to the SPEs' virtual machines.

C.6.3 Secure communications

Secure transportation refers to the impossibility of an attacker to observe the messages transmitted between external actors and a SPE, including the communications to upload the data from data holders, the communications with data users (both for interactive and API-based interfaces) and the communications within other SPEs in the context of the federated learning analyses.

Secure communications will rely on public key infrastructure.

C.6.4 Secure storage

The secure storage refers to the technical solutions to guarantee that the data, which is stored in the SPE, is only accessible to the data users authorised to.

Even it is recommended to count on encryption at rest for the datasets, it would be necessary to evaluate the operational cost of using such solutions per project basis. The associated cost to maintain such solutions and its scalability when dealing which large amounts of data may become a bottleneck for the analysis process. In any case, it should be minimised the exposure of the datasets only to authorised data users, even minimising the accessibility to SPE administrators, when possible.

The secure storage should be complemented with the protocols to separate the decryption endpoints, i.e., where the encrypted data from the data holders is decrypted (using the data holder public keys) for its further encrypted storage using a different

[8] M. Sabt, M. Achemlal and A. Bouabdallah, "Trusted Execution Environment: What It is, and What It is Not," 2015 IEEE Trustcom/BigDataSE/ISPA, Helsinki, Finland, 2015, pp. 57-64, doi: 10.1109/Trustcom.2015.357.

(using the data users public keys). These endpoints may be for at SPE level or a project level, but its further application should be agreed at EU level.

C.6.5 Secure analysis

The secure analysis refers to the use of such analysis tools that minimise the risk of exposing individual level data.

It is important that, even using high security measures in place for the SPEs operation, this security should be increased, wherever possible, by using privacy enhancing technologies (PETs) related to data analysis, for example homomorphic encryption, i.e., encryption mechanisms that permit analysing encrypted data as it was non-encrypted[9] or secure multiparty computation, i.e., cryptographic approaches of computing a model in a distributed manner, guaranteeing the privacy of the parties involved in such computation[10]. The applicability of such technologies should consider its maturity level and its use may be driven by per-project and per SPE basis.

It would also be relevant to evaluate the tools provided to data users. Well known statistical tools, programming languages and libraries should be safe to go, but beyond a basic toolbox, serving as an example the list of software available in the Findata's Kapseli SPE[11]), a methodology to determine which tools can be installed in the SPE provided for a given data access application is recommendable. The certification of such tools, or the container images created ad-hoc to be used in the EHDS context, could be a desirable requirement, but the generation of certification protocols and the standardisation of the certification processes would incur in an unaffordable cost, both for the tools' developers and the SPE operators, and limiting the tools availability for data users. Ideally, limiting the potential damage caused by an inappropriate tool to the extent of the project which is using it might be a minimum requirement.

Finally, a logging system of the analysis operations is recommended for auditing purposes and traceability, in parallel with the access logging system. This logging system should be defined to be as less-intrusive as possible, so as to avoid interactions with the analyses themselves, especially in terms of a performance degradation and a overuse of resources to store the log.

C.6.6 Secure exports

Secure export refers to the technical solutions to verify the extraction of data outside of the SPE. This extraction would be mainly the analysis results at a milestone or the end of the analysis process or the partial analyses results transmitted between SPEs when performing federated analysis.

At this current stage, the analysis results are usually manually checked. This suppose a bottleneck in the SPE operation, so heuristics to speed up this process are required.

[9] Acar, A., Aksu, H., Uluagac, A.S. and Conti, M., 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, 51(4), pp.1-35.

[10] Zhao C, Zhao S, Zhao M, Chen Z, Gao CZ, Li H, Tan YA. Secure multi-party computation: theory, practice and applications. *Information Sciences*. 2019 Feb 1;476:357-72.

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

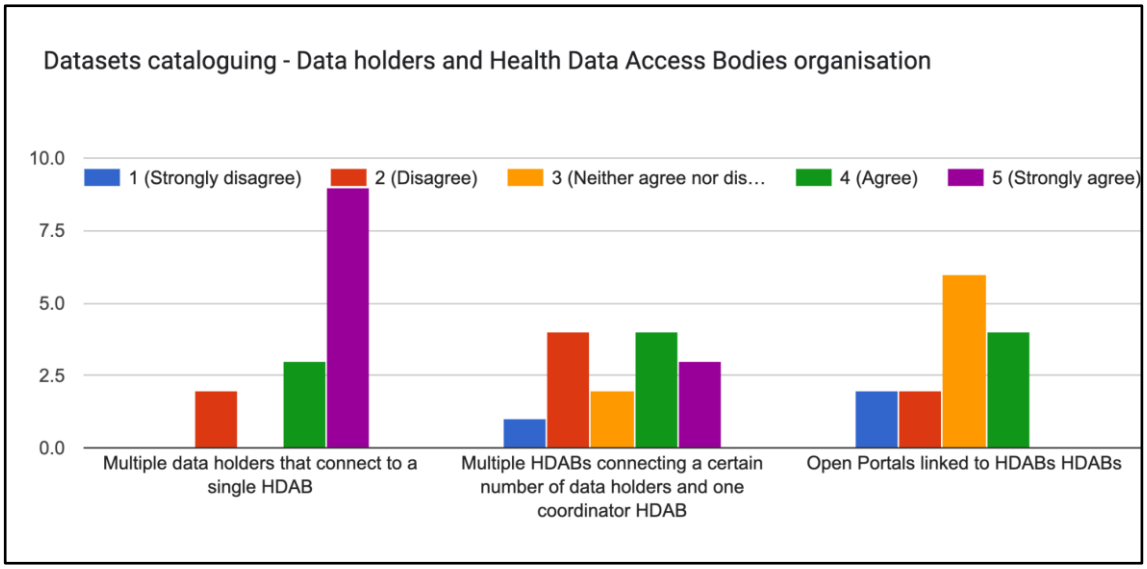
Some assistance is foreseen, whenever possible, using tools or models yet to be defined, e.g., deep learning models that evaluate the profile of the data to be exported. In the federated analysis scenario, it would be possible to relieve the data export burden, as the data transfer between SPEs should be considered secure.

Annex D Voting results

Results of the voting process. The general questions D.1 to D.8 and D.22 to D.27 correspond to the scenarios proposed in the main text of the deliverable, initially introduced in Milestone 7.6.

Questions related to SPEs, D.9 to D.21 were elaborated in the dedicated SPE writing group, to clarify those specific decisions that didn't reached a clear consensus, during the elaboration of the Guideline available in Annex C.

D.1 Datasets cataloguing - Data holders and Health Data Access Bodies organisation



	Multiple data holders that connect to a single HDAB	Multiple HDABs connecting a certain number of data holders and one coordinator HDAB	Open Portals linked to HDABs HDABs
Average	4.357	3.286	2.857
Stdev	1.082	1.326	1.027

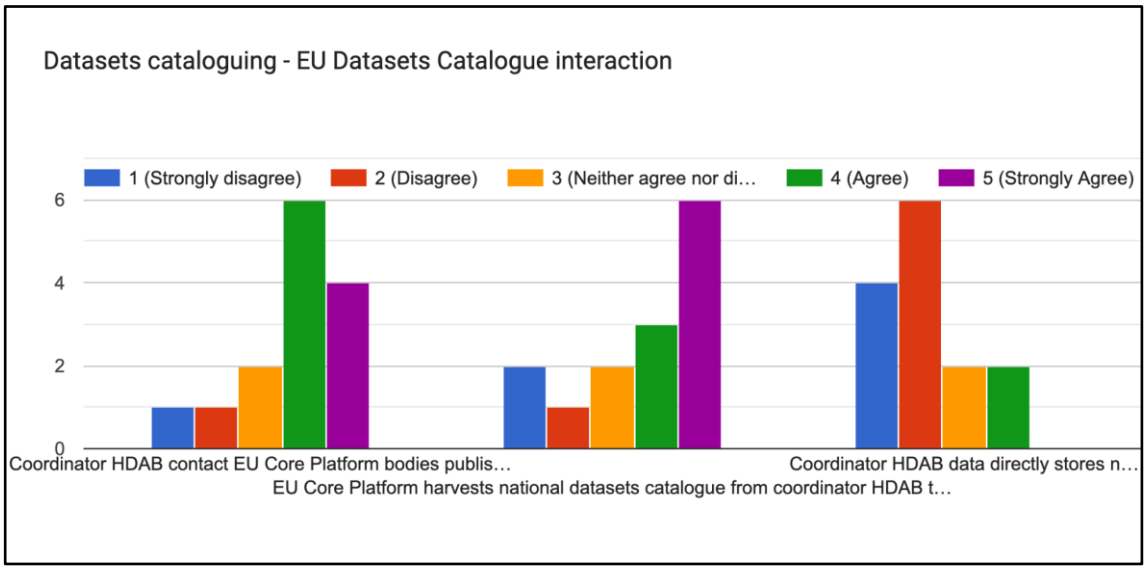
D.1.1 Comments

- Options 1 and 3 arise issues of organisation/performance and sustainability of the processes.
- Depending on national contexts, different MS may have different needs and priorities. E.g., countries that are very centralised versus countries that are very regionalized/decentralised.
- Data is located decentralised; management of norms and standards is managed centrally.
- Our answers are assuming that multiple HDABs here means one HDAB per MS, and one HDAB means one central EU level HDAB. We prefer the solution with

only one HDAB per MS thus, and one of these coordinating when multiple MS data holders are involved.

- We assume that if multiple HDABs are in place then the coordinating HDAB would make sure that all HDABs use the same standard (same training and technology) for the dataset catalogue.
- Open Data should not be part of datasets catalogues scenarios.

D.2 Datasets cataloguing - EU Datasets Catalogue interaction

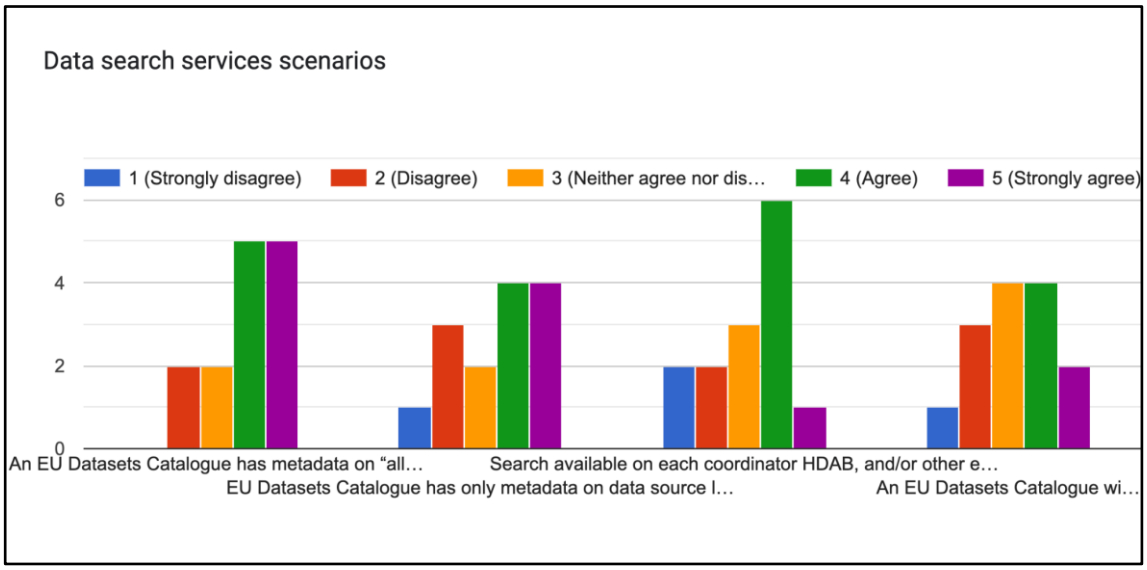


	Coordinator HDAB contact EU Core Platform bodies publish their metadata to the EU Dataset Catalogue	EU Core Platform harvests national datasets catalogue from coordinator HDAB to generate the EU Dataset Catalogue	Coordinator HDAB data directly stores national datasets catalogue in a dedicated space of the EU Core Platform
Average	3.786	3.714	2.143
Stdev	1.188	1.490	1.027

D.2.1 Comments

- The best governance is in the middle solution.
- Information is not collected centrally, it remains in the individual countries.
- Contact here should be an automatic process between computers ("M2M") - not a manual process.
- Going for a harvesting scenario may relieve the burden to HDABs
- Training is a very important aspect and we believe it should be added. For the second option the updating frequency can increase.

D.3 Data search services scenarios



	An EU Datasets Catalogue has metadata on "all levels"	EU Datasets Catalogue has only metadata on data source level and URL to more detailed metadata catalogues at national datasets catalogue	Search available on each coordinator HDAB, and/or other entry points, independently to the metadata capabilities of choice	An EU Datasets Catalogue with or without metadata on data source level, but with open data of different kinds
Average	3.929	3.500	3.143	3.214
Stdev	1.072	1.345	1.231	1.188

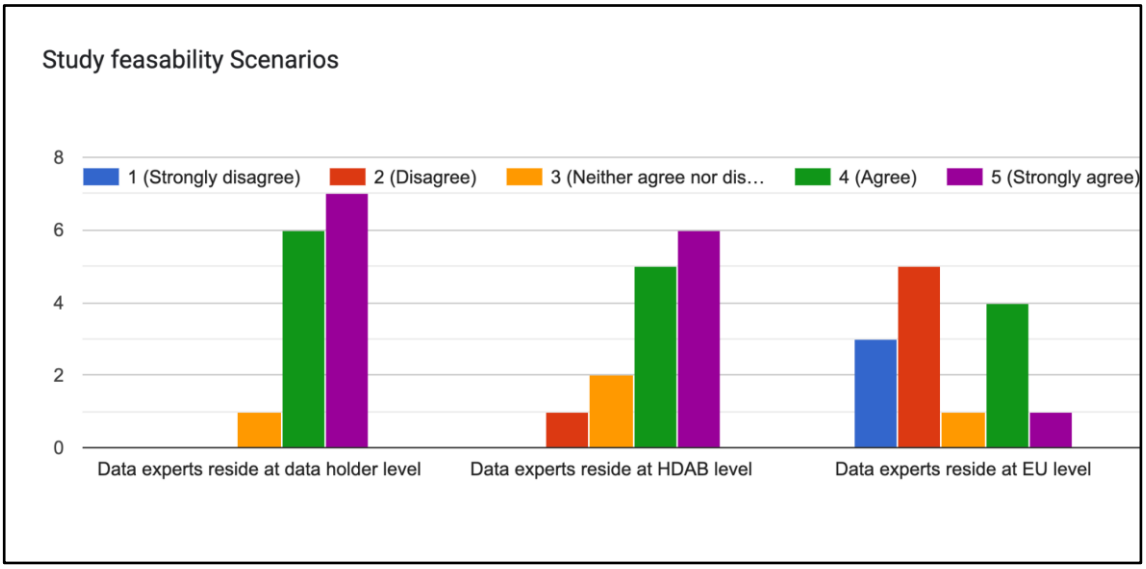
D.3.1 Comments

- The 2nd and 3rd options are a good compromise to allow flexibility while preserving. The 1st is not clear. The last is reasonably not sustainable.
- Searching in 27+ sources of catalogue data requires a sort of simplification and hierarchical structure may help faster converge to desided datasets. Detailed information about the data referenced in the catalogue may be easier to maintain on a national level. This opinion also slightly considers the currently discussed opt-out (of data subjects) and its impact on the datasets maintenance that will be made available by HDABs.
- For the convenience of the users, scenario 1 would be ideal - however, the feasibility of this solution may not be possible even within a longer timespan.
- Metadata must be standardised, in a transitional phase open standards may be necessary
- It would probably be sufficient to host high-level metadata for data discovery at EU platform level, with links to more detailed information at the MS level. The most frequent user is a national user of data from one MS. However, if feasible

(computationally), we have nothing against a solution with all levels of metadata at EU platform level, as long as it is a replication of the metadata available at each MS level (i.e. metadata not only available through the EU platform).

- Concerning open data: difficult to maintain and not in the scope of EHDS2.
- It will important to decide whether there will be a single entry point to the HealthData@EU (a EU portal) or every MS may have its own
- The first option would be feasible if there are data profiles available for all dataset catalogues. Option 4 is badly copy pasted from the box above!

D.4 Study feasibility Scenarios

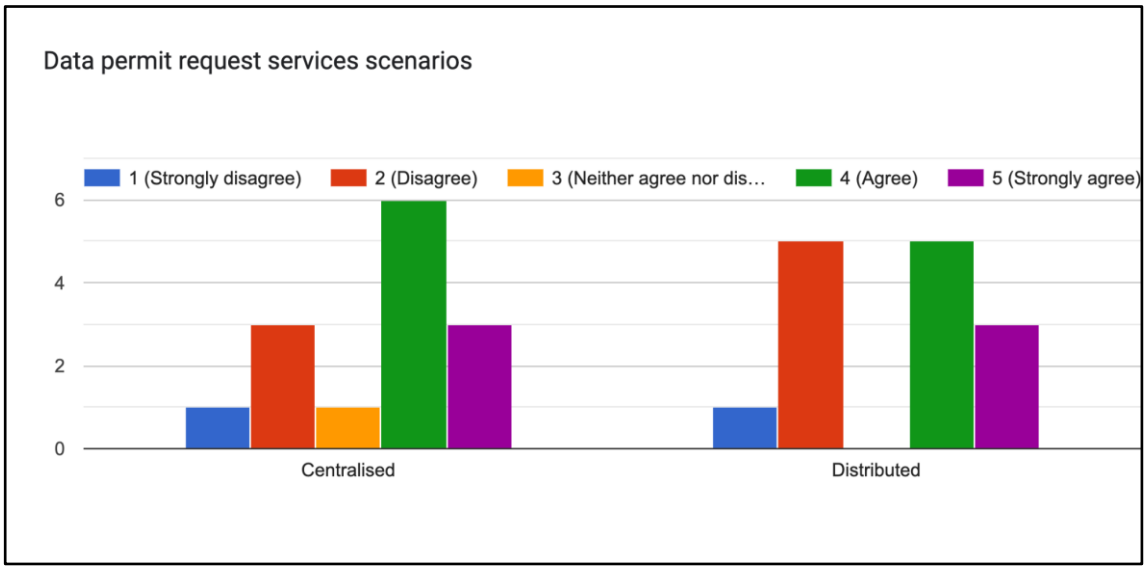


	Data experts reside at data holder level	Data experts reside at HDAB level	Data experts reside at EU level
Average	4.429	4.143	2.643
Stdev	0.646	0.949	1.336

D.4.1 Comments

- Data Experts should be available at all the levels with different efforts according to the roles, the amount of data/type of datasets and the relevance and responsibility of their tasks.
- In the case of data scientists at data holders, the issue is with a broad range of the data holders, e.g. large and also small hospitals. For smaller data holders it would be very difficult to keep data scientists. It was also proposed during negotiation of the Regul. on the EHDS that HDABs should provide expertise / support smaller data holders.
- Data experts must reside at HDAB level. However, they would ideally also be available at other points - the financial capability of institutions will determine viability of other scenarios.
- Data must be stored as close to the source as possible, thus also knowledge of data.
- Some data expertise must reside within the HDAB level such that they can perform their tasks, but it must also be present at data holder level of course, including subject expertise.

D.5 Data permit request services scenarios

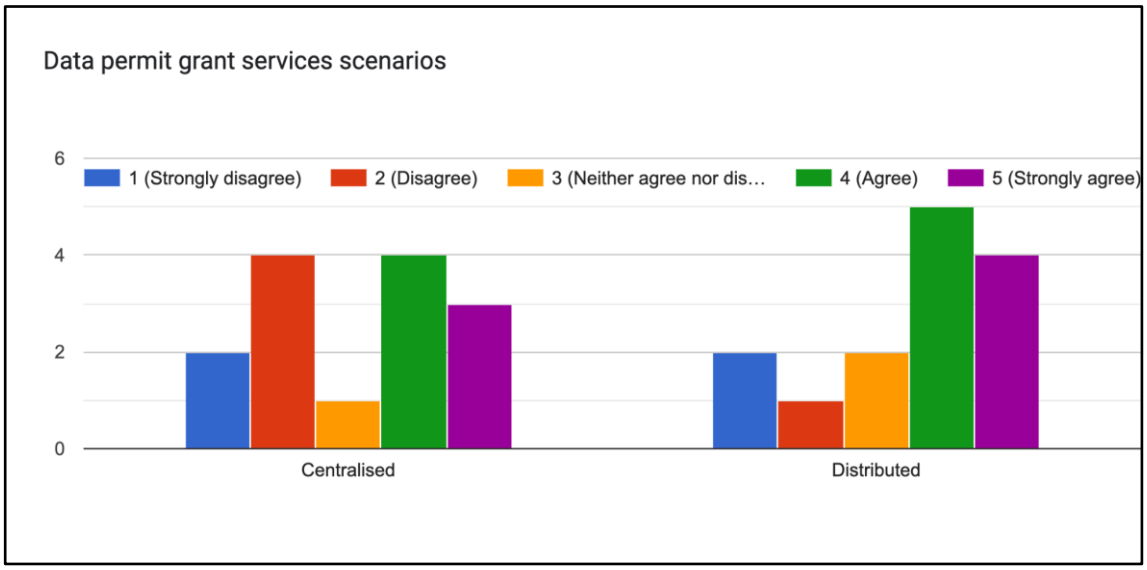


	Centralised	Distributed
Average	3.500	3.286
Stdev	1.286	1.383

D.5.1 Comments

- The answer to this question is consistent with the previous ones
- Either scenario is valid, given that the systems communicate appropriately and in a timely fashion. Interoperability should be insured and seamless.
- Rules are given centrally, handled decentralised.
- Still distributed (local) services may be needed for several years before centralised services are fully operational.
- both possible, centralised system preferable for data users

D.6 Data permit grant services scenarios

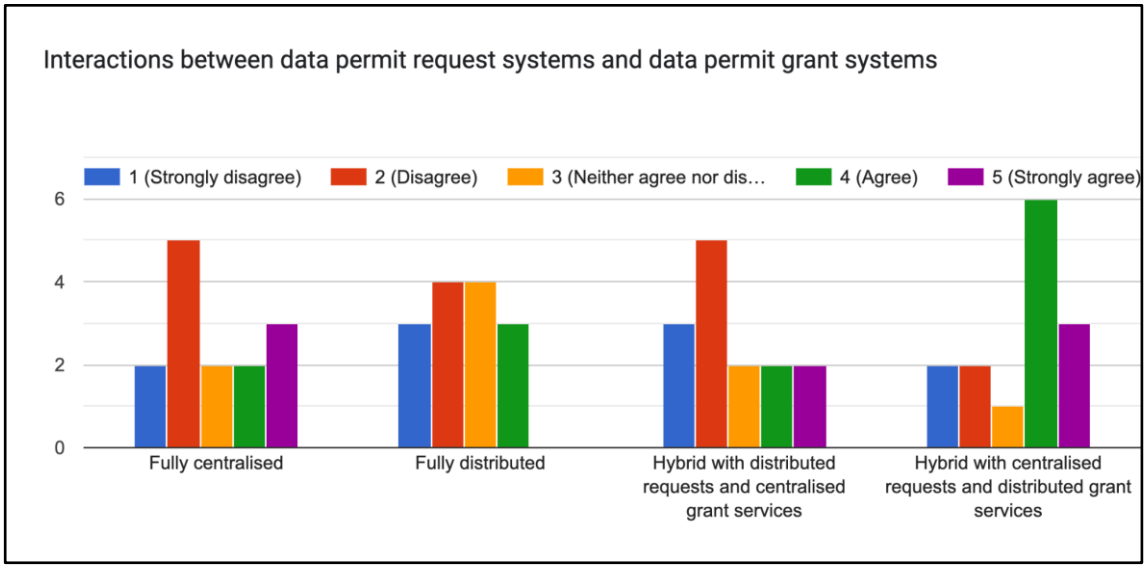


	Centralised	Distributed
Average	3.143	3.571
Stdev	1.460	1.399

D.6.1 Comments

- Same comment as above.
- Local deviations are for exception handling.
- Each MS must retain its right to grant permits for the use of its own data. But for most data sources, data permit granting should be centralised at the national HDAB level.
- Customisation of approval probably necessary

D.7 Interactions between data permit request systems and data permit grant systems

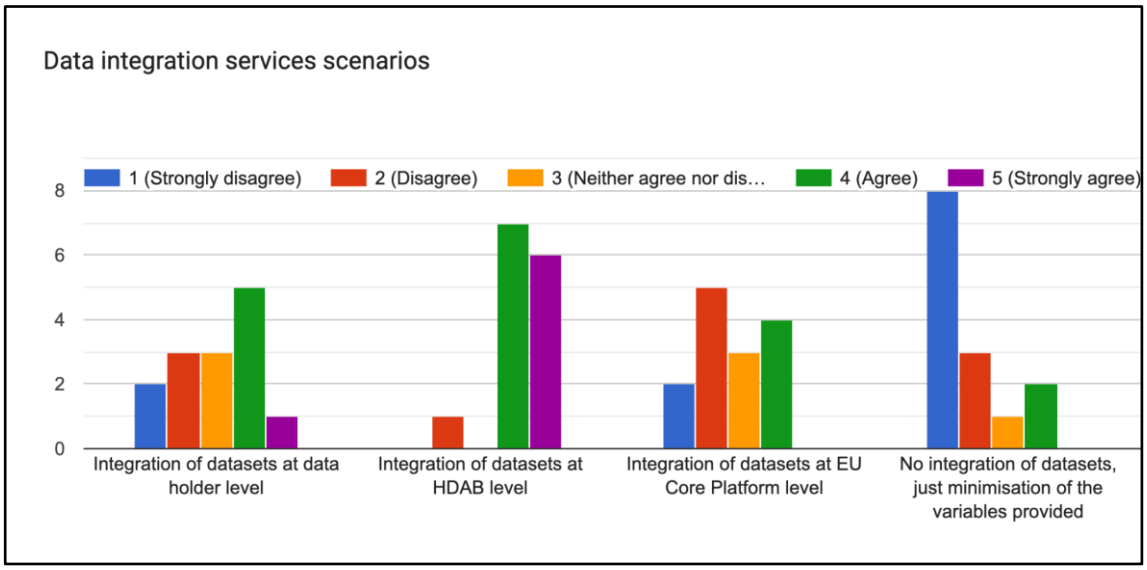


	Fully centralised	Fully distributed	Hybrid with distributed requests and centralised grant services	Hybrid with centralised requests and distributed grant services
Average	2.929	2.500	2.643	3.429
Stdev	1.439	1.092	1.393	1.399

D.7.1 Comments

- We follow the same rules and procedures for application, approval is handled locally
- Important that each MS remains in control of data permit grant functionality. However, any HDAB involved in the same project must be able to check/validate if an applicant has been granted or denied permit by the other HDABs in any MS.
- Fully distributed can be a long term target if we want to avoid any SPOF.

D.8 Data integration services scenarios

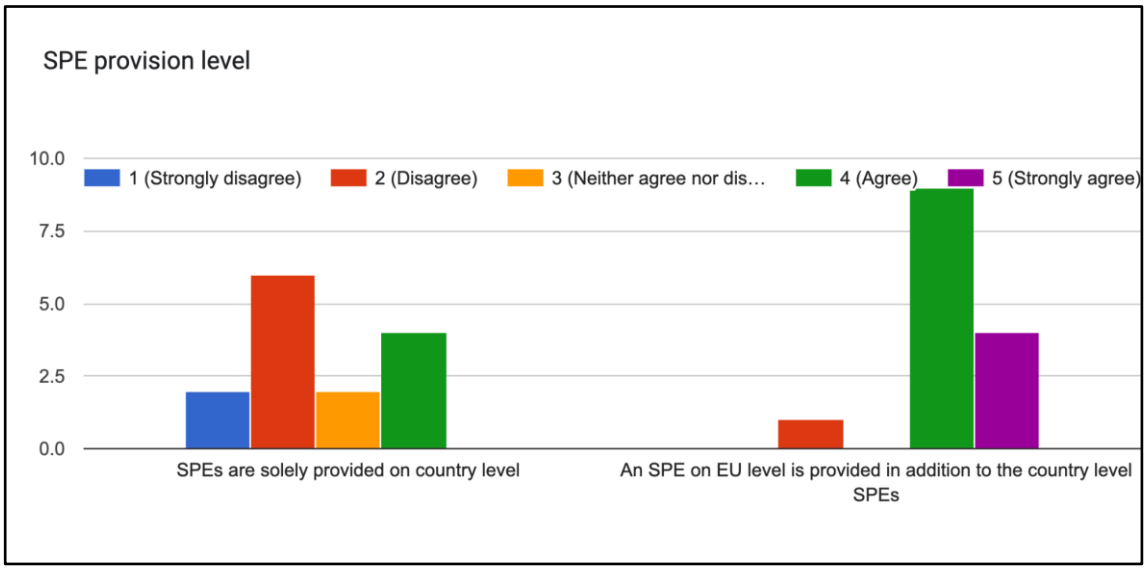


	Integration of datasets at data holder level	Integration of datasets at HDAB level	Integration of datasets at EU Core Platform level	No integration of datasets, just minimisation of the variables provided
Average	3.000	4.286	2.643	1.786
Stdev	1.240	0.825	1.082	1.122

D.8.1 Comments

- The integration is important at the different levels with a different effort
- Local processing may be required before data or results are delivered.
- We interpret "data integration" to mean joining/matching different data sources from different data holders within a project. We have not interpreted the alternatives as mutually exclusive, n.b. In some cases, there might be a need to have data integration at EU core platform level, but the first option should be federated analysis or local analysis at HDAB level.
- Need to clarify if the integration also includes the anonymisation/pseudonymisation
- HDAB or EU level depending on whether cross-border use is needed
- Will depend on competence and capacity needed for different projects.
- In France, as defined by regulation, some transformations must be performed by data holders (e.g.,pseudonymisation). To avoid any extra burden on DH, other transformations (e.g., standardisation) are done by the HDAB.

D.9 SPE provision level

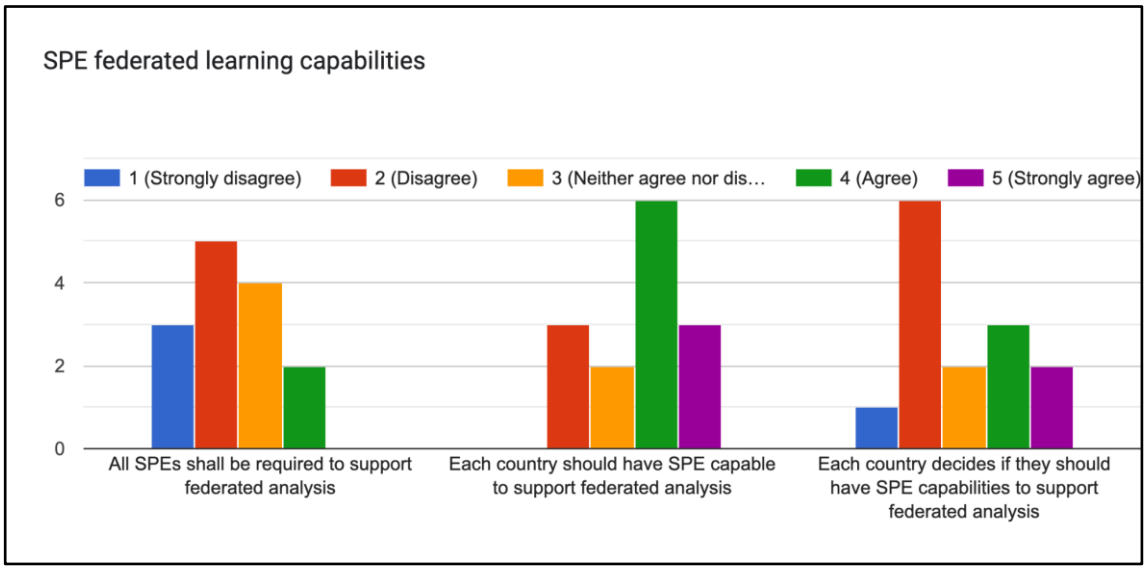


	SPEs are solely provided on country level	An SPE on EU level is provided in addition to the country level SPEs
Average	2.571	4.143
Stdev	1.089	0.770

D.9.1 Comments

- National SPEs are a necessity to ensure MS autonomy and adequacy to country-level idiossyncracies. However, an EU-level SPE may be provided for special purposes (such as for the use of transnational organisations - like OECD, WHO, ECDC, HERA, et cetera).
- Application/technique for SPE should be standardised as much as possible.
- SPEs with different services and software/tools mighht be needed to serve the wide range of users, since the types of data and associated tools vary a lot between various fields of work. Some SPE providers may specialize in e.g. tools for genetic analysis.
- Clarifications should be provided for third-country data users, data holders and SPEs
- Both levels needed
- In Norway the country level SPE's will be most important, but are ok with both options.
- An Open-Source-Solution for SPEs should be provided by the EU (as SPEs are also needed for other Data spaces), member states can modify it according to their national needs.

D.10 SPE federated learning capabilities

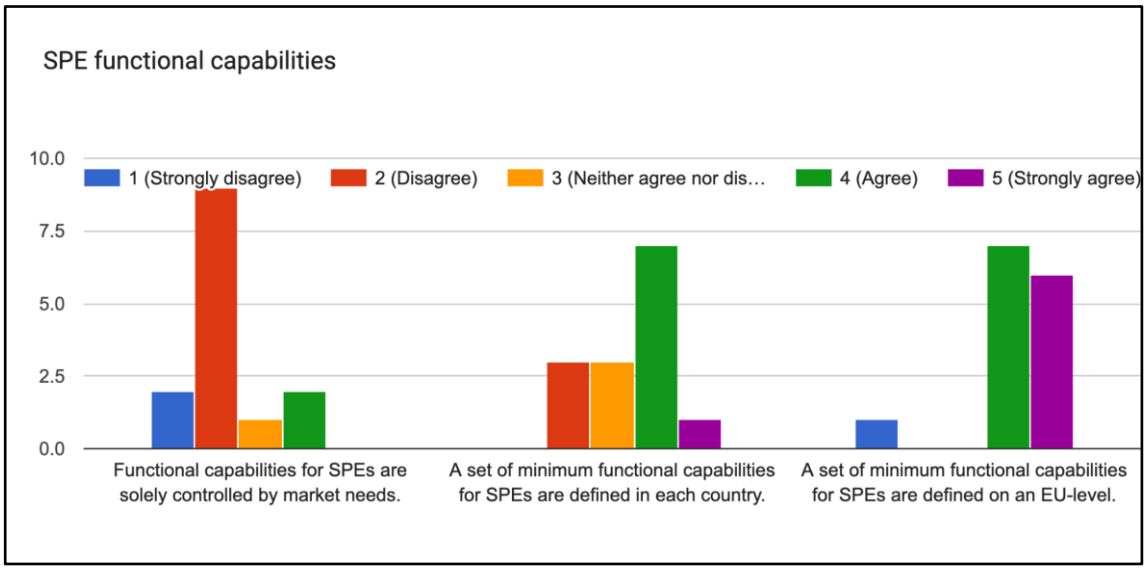


	All SPEs shall be required to support federated analysis	Each country should have SPE capable to support federated analysis	Each country decides if they should have SPE capabilities to support federated analysis
Average	2.357	3.643	2.929
Stdev	1.008	1.082	1.269

D.10.1 Comments

- Country-level autonomy on the option to support federated analysis would allow MS to make a cost/benefit decision - this would be a strong recommendation in the short/medium term, making sure this element will not add unecessary difficulty to initial national implementation. However, there should be a commitment to provide at least one SPE capable of federated analysis per country, in the long term (this would incentivize cohesion and homogeneity of services).
- We must promote cooperation; it must be possible with exceptions in a transitional phase.
- Initially, it may be that each MS shoud be merely encouraged to have at least one SPE supporting federated analysis but not forced. However, in the long run, the EHDS proposition should aim towards MS reciprocity here, with federated analysis offered in each MS.
- It is not realistic that all SPE's shall support federated learning, but it may be necessary to have requirements from EU on a country level to ensure further development of this capability.
- I guess you mean federated learning everywhere and not federated analysis, right? For federated learning the SPEs would need to be connected in real time. Having this capability is very resource dependent.

D.11 SPE functional capabilities

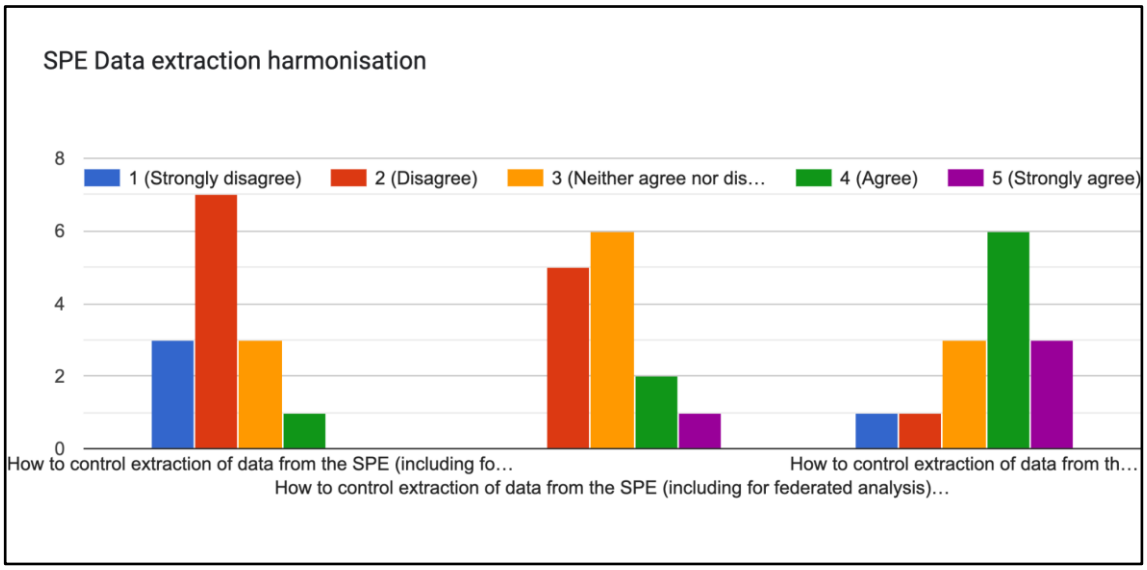


	Functional capabilities for SPEs are solely controlled by market needs.	A set of minimum functional capabilities for SPEs are defined in each country.	A set of minimum functional capabilities for SPEs are defined on an EU-level.
Average	2.214	3.429	4.214
Stdev	0.893	0.938	1.051

D.11.1 Comments

- It is critical that SPE functionalities are not determined by market needs, as the major focus should be on generating added value for the public good - however, there should be a keen attention on following market needs and trends to ensure up-to-date tools and capabilities for users. An EU-level consensus on minimum functional capabilities will help MS in navigating implementation and maintaining cohesion, by giving a clear outline of what is expected to be built and maintained. Minimum functional capabilities determined by each country gives them more autonomy in decision-making, but also creates a potential issue by enabling a patchwork of different specifications across Europe.
- Harmonisation is important.
- In order to serve the research community, a minimal set of requirements should be agreed at EU level. Ideally there would also be agreements at EU level on technical and semantic interoperability in the long run.
- Minimum functional capabilities should be defined on an EU-level, national requirements could be added
- It may be difficult to identify common traits for minimum requirements since SPE's can have different specialities. The answers above are dependent on that such minimum requirements are identified. It is also important that they are kept to a minimum to allow for flexibility in developing SPE services.

D.12 SPE Data extraction harmonisation

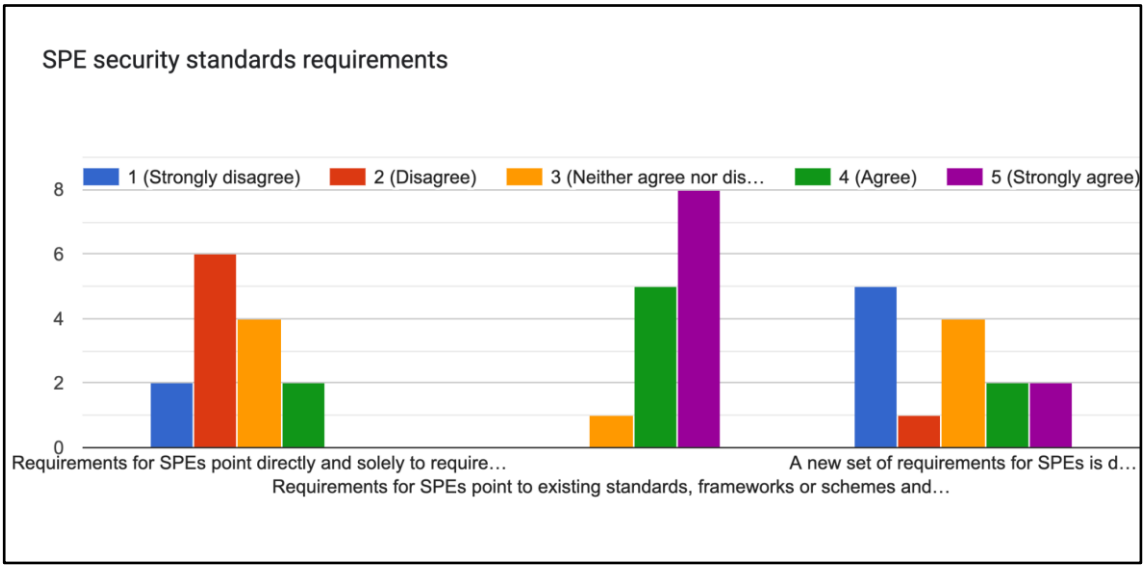


	How to control extraction of data from the SPE (including for federated analysis) should be determined by each SPE.	How to control extraction of data from the SPE (including for federated analysis) should be determined by each country.	How to control extraction of data from the SPE (including for federated analysis) should be harmonised within the EU.
Average	2.143	2.929	3.643
Stdev	0.864	0.917	1.151

D.12.1 Comments

- For reasons similar to the previous question (SPE Functional Specifications), EU-level agreement on data extraction allows for better homogeneity. Discussions between Member States on which approaches are best suited is a necessity not to be avoided.
- Harmonisation is important.
- It is important with common minimum requirements to build trust that is essential for the data holders to share their data. It is however important that the requirements are not too strict and kept at a minimum level. Optional requirements could be added locally.
- Harmonisation is necessary and important.

D.13 SPE security standards requirements

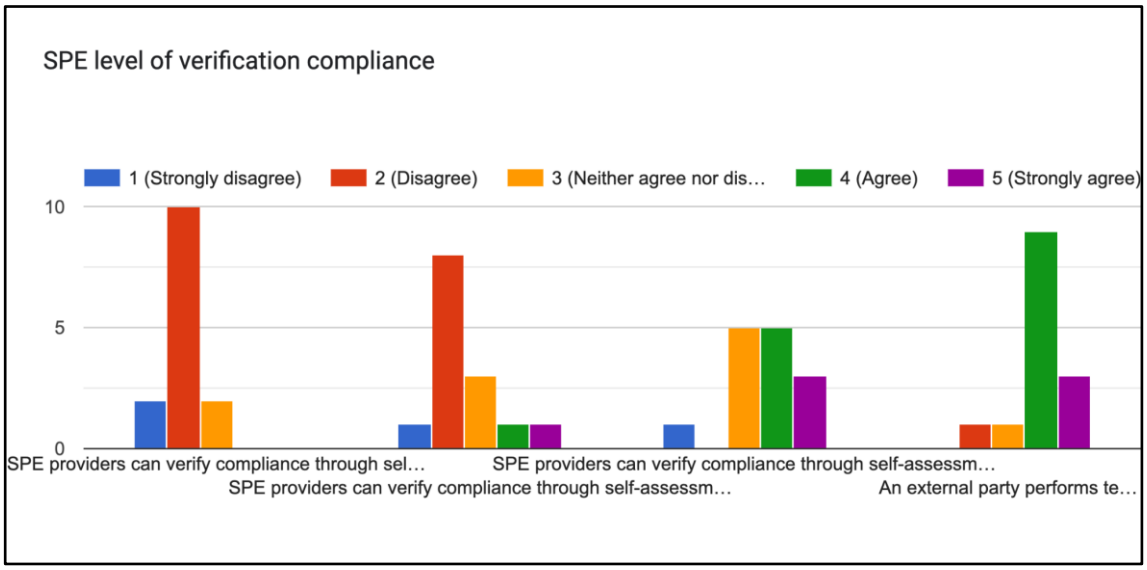


	Requirements for SPEs point directly and solely to requirement sets in existing standards, frameworks or schemes. E.g. one or several ISO-standards.	Requirements for SPEs point to existing standards, frameworks or schemes and are complemented with EHDS-specific requirements.	A new set of requirements for SPEs is developed based on existing standards.
Average	2.429	4.500	2.643
Stdev	0.938	0.650	1.499

D.13.1 Comments

- It's a good principle not to start from zero, and to use existing standards that are applicable. However, there should be an acknowledgement that EHDS has certain particularities that merit the creation of specific new requirements.
- Harmonisation is important.
- As far as possible, existing international standards should be employed. Where none exist, amendments with new standards that take into account EHDS-specific requirements should be made. We do not recommend inventing the wheel with brand new standards all through.
- SPE cybersecurity standards should be compatible with national cybersecurity standards

D.14 SPE level of verification compliance

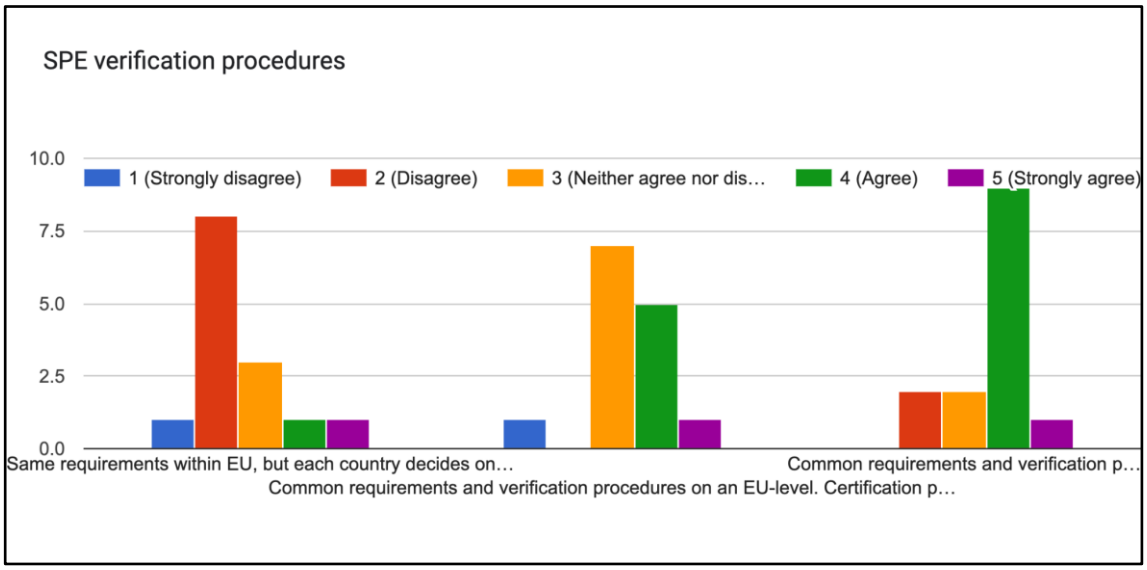


	SPE providers can verify compliance through self-assessment	SPE providers can verify compliance through self-assessment. Voluntary compliance testing is also available. (e.g., similar to CISPE Data Protection Code of Conduct)	SPE providers can verify compliance through self-assessment. External audits are performed (e.g., similar to the Data protection agencies audit procedures today).	An external party performs testing against requirements to verify compliance and provide certification. (e.g., similar to the certification process in Finland)
Average	2.000	2.500	3.643	4.000
Stdev	0.555	1.019	1.082	0.784

D.14.1 Comments

- To ensure maximum level of trust, there should always be external compliance audits. Having a system entirely based on external party testing may be costly and difficult for some MS.
- We take it to mean "verification of compliance" in this context. For enhanced trust among MS citizens, external audits are preferred. There may be a need for a progression towards this end, since it may take time to build such an external audit body/procedure.
- Audits can be in addition to voluntary testing. The level of verification is dependent on the previous question on requirements. If a high level of verification is required, the requirements to be verified should not be too «heavy» to comply with and verify.

D.15 SPE verification procedures

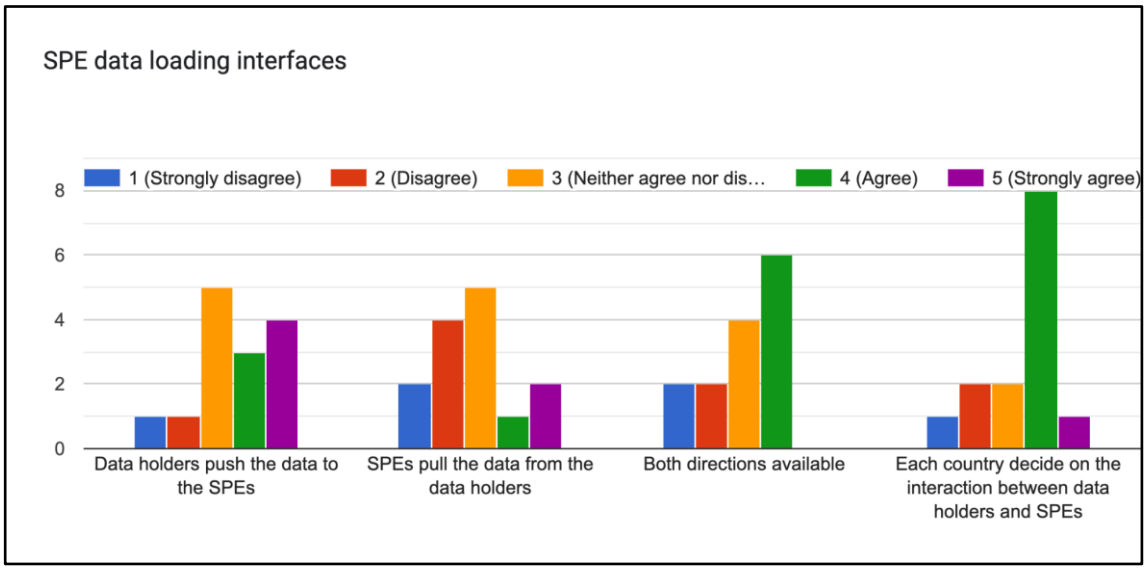


	Same requirements within EU, but each country decides on verification procedures	Common requirements and verification procedures on an EU-level. Certification processes controlled by local bodies in each country	Common requirements and verification procedures on an EU-level. Certification process controlled by EU body
Average	2.500	3.357	3.643
Stdev	1.019	0.929	0.842

D.15.1 Comments

- EU-level verification and certification improves homogeneity and trust.
- We don't really understand the difference between this and the previous question

D.16 SPE data loading interfaces

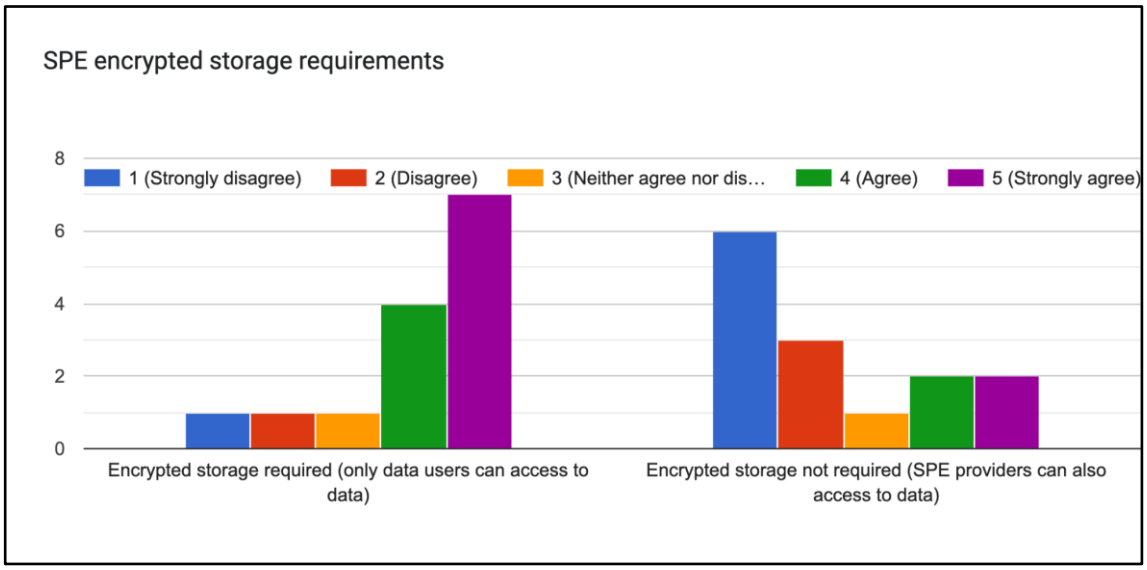


	Data holders push the data to the SPEs	SPEs pull the data from the data holders	Both directions available	Each country decide on the interaction between data holders and SPEs
Average	3.571	2.786	3.000	3.429
Stdev	1.222	1.251	1.109	1.089

D.16.1 Comments

- Flexibility in data loading is preferable, since 1) each MS can adapt the SPE to their existing operations, and 2) "less harmonisation" in this regard does not mean more heterogeneity in security and trust.
- This should be a choice made at MS level depending on existing data holders/sources and associated legislation.
- The push method is currently used and will be needed also in the future. However, in the case of well-defined data sets and interfaces (e.g. OMOP) it will be possible to also enable the pull method (in practice an API interface). This approach could be more easily automated, thereby reducing the data holder's workload.
- We expect that pull is strictly controlled.
- Recommendation should be to prioritise push from DH to HDAB to guarantee security by default. If security can be ensured by the data holders when exposing an endpoint to pull his data, then it could be allowed.

D.17 SPE encrypted storage requirements

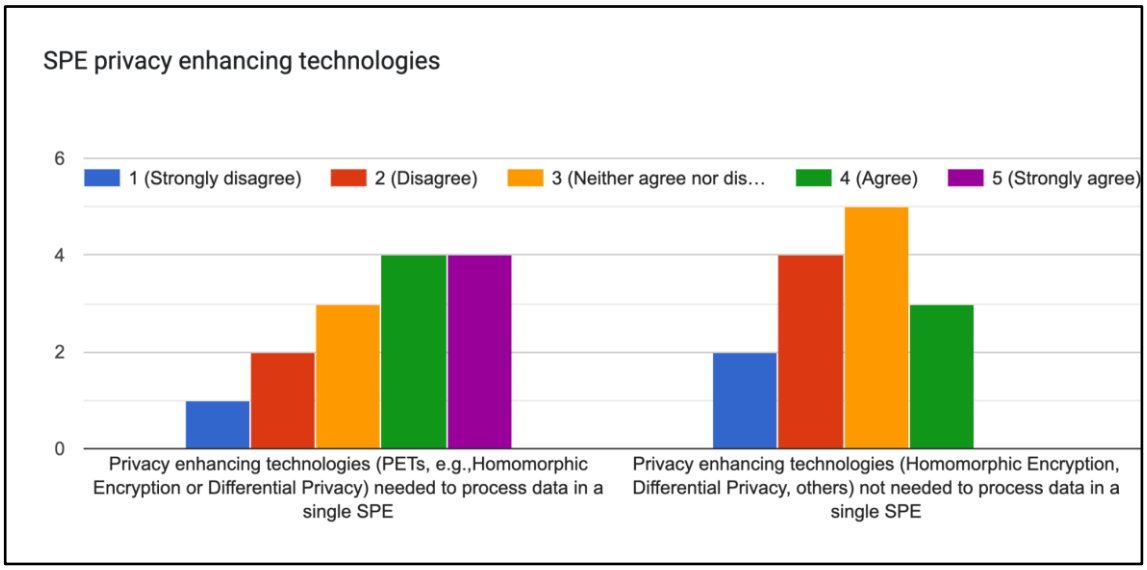


	Encrypted storage required (only data users can access to data)	Encrypted storage not required (SPE providers can also access to data)
Average	4.071	2.357
Stdev	1.269	1.550

D.17.1 Comments

- Encryption of stored data is the ideal solution to protect data subjects and promote trust. However, given that some MS may not have the in-house expertise and resources to implement and maintain such a measure, non-requirement would also be acceptable (especially in the short-term, to speed up implementation of EHDS2). Non-sensitive data may not require storage encryption.
- Encryption must be able to be used appropriately
- As long as compliance verification is not hindered.
- To the best of our knowledge: (i) access to SPEs by data holders is not foreseen by the EHDS Regulation proposal; (ii) access is needed not just for data users, but for the personnel of the HDAB in charge of the tasks foreseen in in article 31(1) of the EHDS proposal. Also, encryption in the SPEs is not necessarily linked to data access by a specific stakeholder, it rather depends on who holds the encryption key. Therefore, we don't understand the options proposed in this question
- Encryption at rest as standard, but there may be exceptions, e.g. large volumes of data. Other security measures would then be necessary.

D.18 SPE privacy enhancing technologies

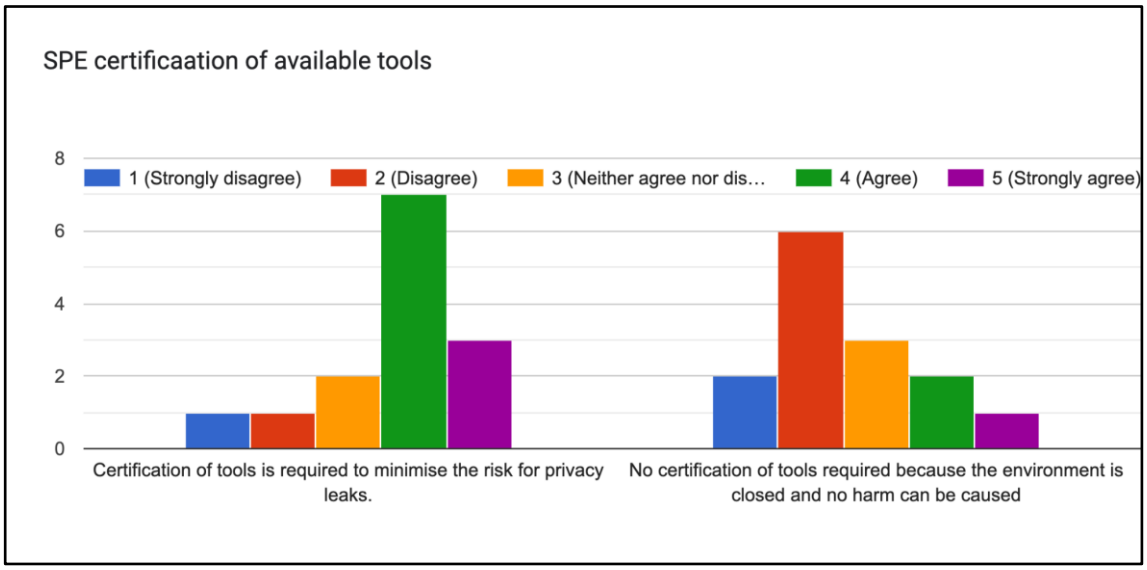


	Privacy enhancing technologies (PETs, e.g.,Homomorphic Encryption or Differential Privacy) needed to process data in a single SPE	Privacy enhancing technologies (Homomorphic Encryption, Differential Privacy, others) not needed to process data in a single SPE
Average	3.571	2.643
Stdev	1.284	1.008

D.18.1 Comments

- The use of privacy enhancement technologies should not be negotiable, even if it implies certain efficiency drawbacks.
- Privacy must be respected when required
- Privacy by design could be a potential role model
- According to EHDS regulation data has to be provided in an anonymised format, therefore privacy enhancing technologies will be needed. However, it must be thoroughly investigated which ones are suitable. Some techniques may not be feasible. In some cases data can be provided in a pseudonymised format (according to regulation proposal).

D.19 SPE certification of available tools

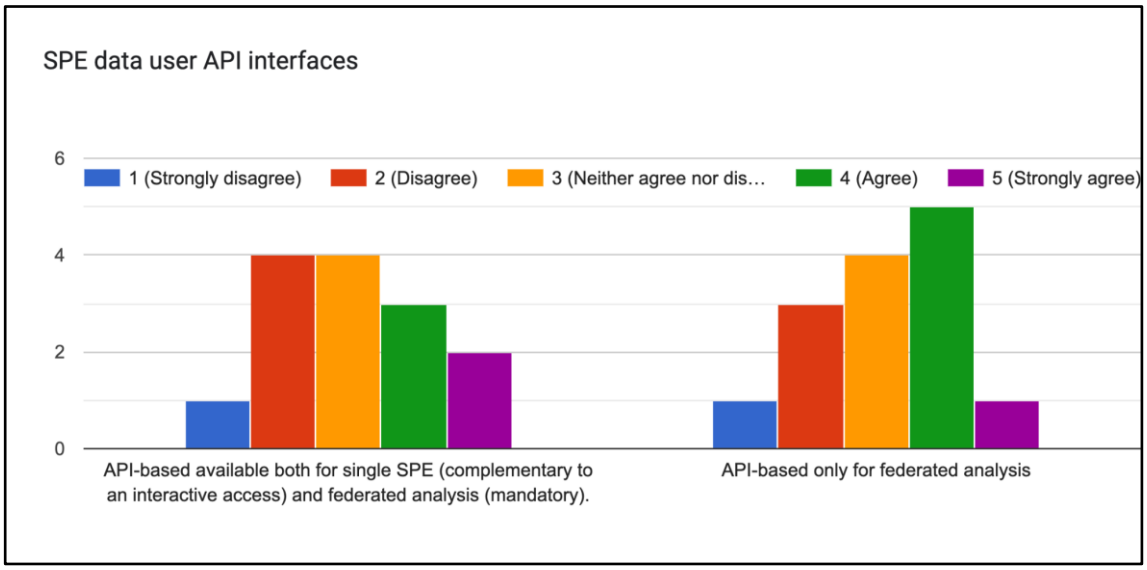


	Certification of tools is required to minimise the risk for privacy leaks.	No certification of tools required because the environment is closed, and no harm can be caused
Average	3.714	2.571
Stdev	1.139	1.158

D.19.1 Comments

- As mentioned in the response above, privacy and security measures should be prioritised and considered non-negotiable. A certification could be seen as a seal of trust in specific software.
- Important to consider the certification process. It should be as simplistic as possible so as to not cause unnecessary delays in access to analysis. Yet, there is a danger in believing the SPE is fool-proof and no prior vetting of tools to be needed. Perhaps a test environment for new tools could be an option.
- Certification of SW is not needed. However, harmonised quality and security principles for accepting pre-installed tools should be agreed.
- What do you mean to certify a tool?

D.20 SPE data user API interfaces

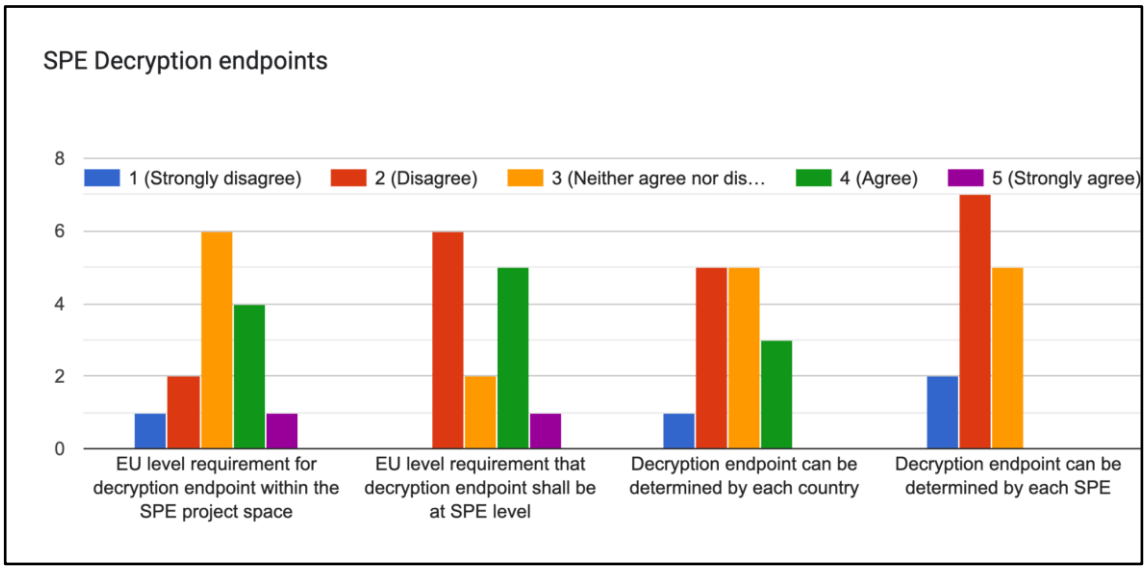


	API-based available both for single SPE (complementary to an interactive access) and federated analysis (mandatory).	API-based only for federated analysis
Average	3.071	3.143
Stdev	1.207	1.099

D.20.1 Comments

- We see a need for API interface harmonisation. Our answers are based on the listed pros and cons.
- Limiting only to federated analysis is an unnecessary limitation. Similar API functionality can support both cases.
- More time to think about question needed
- Federated learning not analysis

D.21 SPE Decryption endpoints

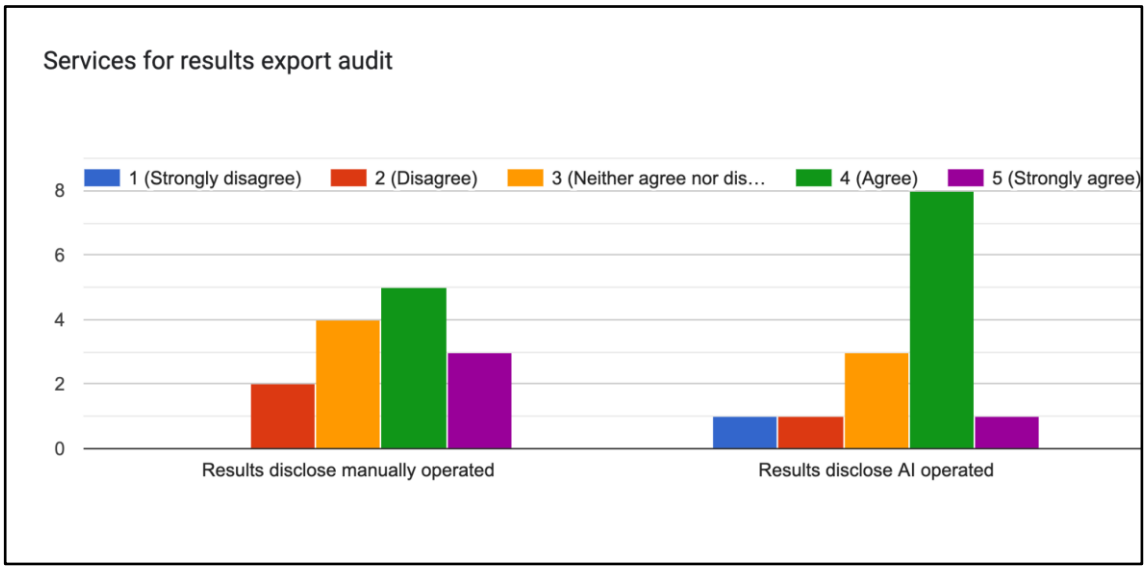


	EU level requirement for decryption endpoint within the SPE project space	EU level requirement that decryption endpoint shall be at SPE level	Decryption endpoint can be determined by each country	Decryption endpoint can be determined by each SPE
Average	3.143	3.071	2.714	2.214
Stdev	1.027	1.072	0.914	0.699

D.21.1 Comments

- EU level requirements for decryption endpoints promote homogeneity and trust. A centralised scenario for credentials management would be advantageous.
- We must harmonise, but it is also important to avoid being imposed on bad solutions.
- More time to think about question needed

D.22 Services for results export audit

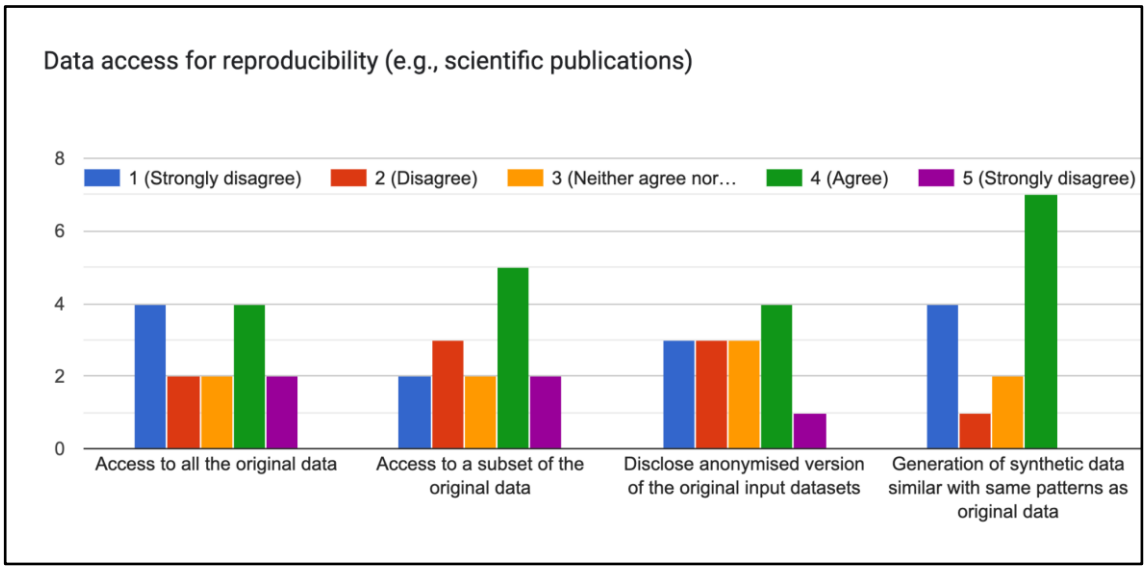


	Results disclose manually operated	Results disclose AI operated
Average	3.643	3.500
Stdev	1.008	1.019

D.22.1 Comments

- It depends on timing needed for AI tools development
- Until AI solutions are widely tested, and certified, manual operation should be prioritised. Although this may be less scalable, we should not recommend solutions that involve "unknowns". Nonetheless, a hybrid solution using both approaches will be the best possible scenario.
- At the moment, AI is premature for the task. However, manually operated procedures are resource consuming and the long goal should be to strive towards automation (AI or other algorithmic solution).
- Automatic disclosure should be the final target. Manual operation may be additionally needed for some time. May also depend on case-by-case, but harmonised criteria and methods are needed.
- AI-assisted disclose should be manually audited in some cases
- currently, manual audit seems to be the most suitable solutions, however AI may be an interesting approach for the future
- Scenario 2 may be a possible solution in the future but is not yet mature enough.

D.23 Data access for reproducibility (e.g., scientific publications)

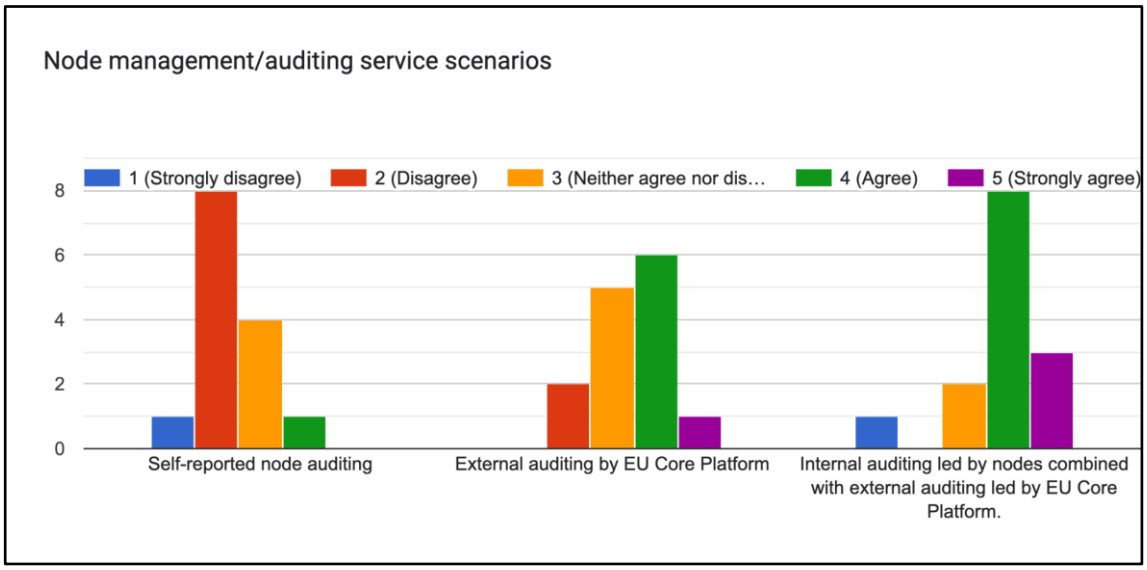


	Access to all the original data	Access to a subset of the original data	Disclose anonymised version of the original input datasets	Generation of synthetic data similar with same patterns as original data
Average	2.500	2.833	2.615	2.857
Stdev	1.314	1.193	1.193	1.351

D.23.1 Comments

- This question does not seem entirely clear. For scientific reproducibility, the only real solution is having access to the complete original data. Synthetic data does not allow for real reproducibility. Please add a reference to EHDS data access models in question.
- The individual research projects should be assessed based on relevance.
- (We suppose answer level 5 has a typo and should be "strongly agree"). A difficult question, considering legislative demands for verification of published scientific results years after. We do not believe anonymisation is ever possible, hence disagree. Synthetic data is also not enough to verify results from important publications.
- Note: (5 should "strongly agree"). In most cases it is not possible to anonymize without considerable modification of the data set. Synthesising in some cases is sufficient, but not generally, especially if the idea is to reproduce the original results.
- All possibilities may be sensible in certain cases (not only for reproducibility questions, but for data access and analyses in general)

D.24 Node management/auditing service scenarios

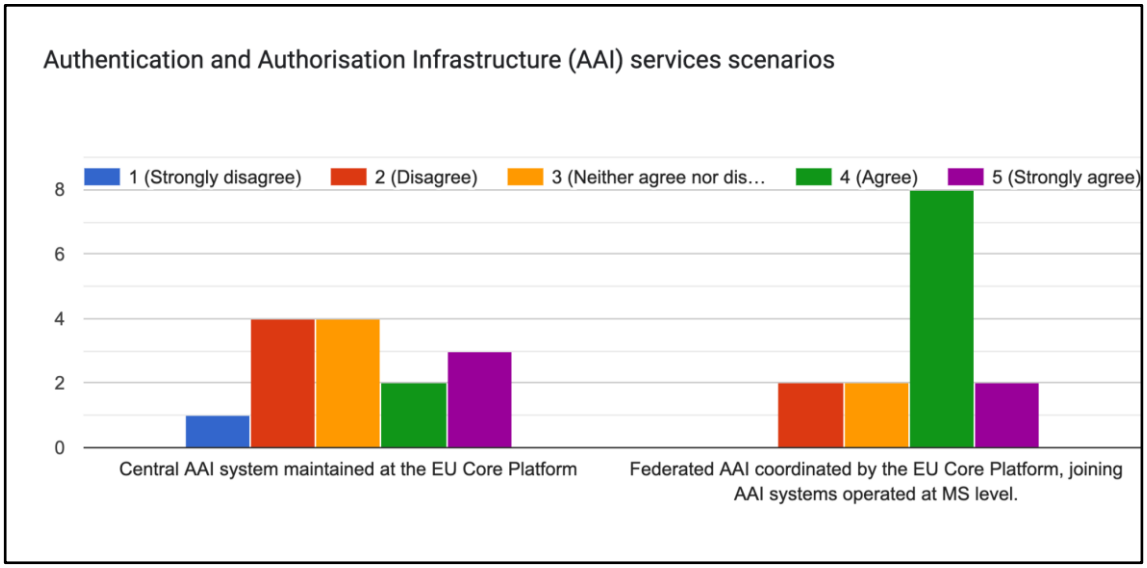


	Self-reported node auditing	External auditing by EU Core Platform	Internal auditing led by nodes combined with external auditing led by EU Core Platform.
Average	2.357	3.429	3.857
Stdev	0.745	0.852	1.027

D.24.1 Comments

- Sharing audit responsibilities between MS and Core Services ensures more transparency and equitability. Inter-institutional coordination will always be a necessity - use MyHealth@EU as a case study since regular external auditing is already a reality.
- Self-auditing is not sufficient.
- We need more information ideally to know what the audit concerns. In general, external audits are preferred for increased trust in the process.
- Is the node here an HDAB?

D.25 Authentication and Authorisation Infrastructure (AAI) services scenarios

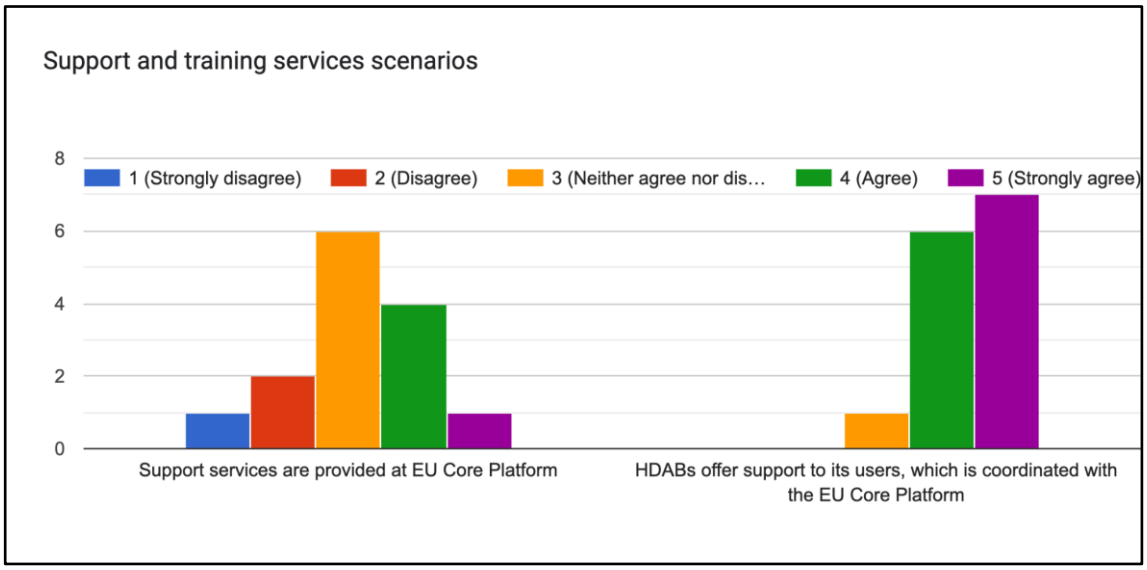


	Central AAI system maintained at the EU Core Platform	Federated AAI coordinated by the EU Core Platform, joining AAI systems operated at MS level.
Average	3.143	3.714
Stdev	1.292	0.914

D.25.1 Comments

- A federated solution would promote equitability and ensure eIDAS solutions in operation at a national level are leveraged to benefit the system.
- It must be a check of identity and approvals across borders
- National solutions that are already implemented could be useful

D.26 Support and training services scenarios

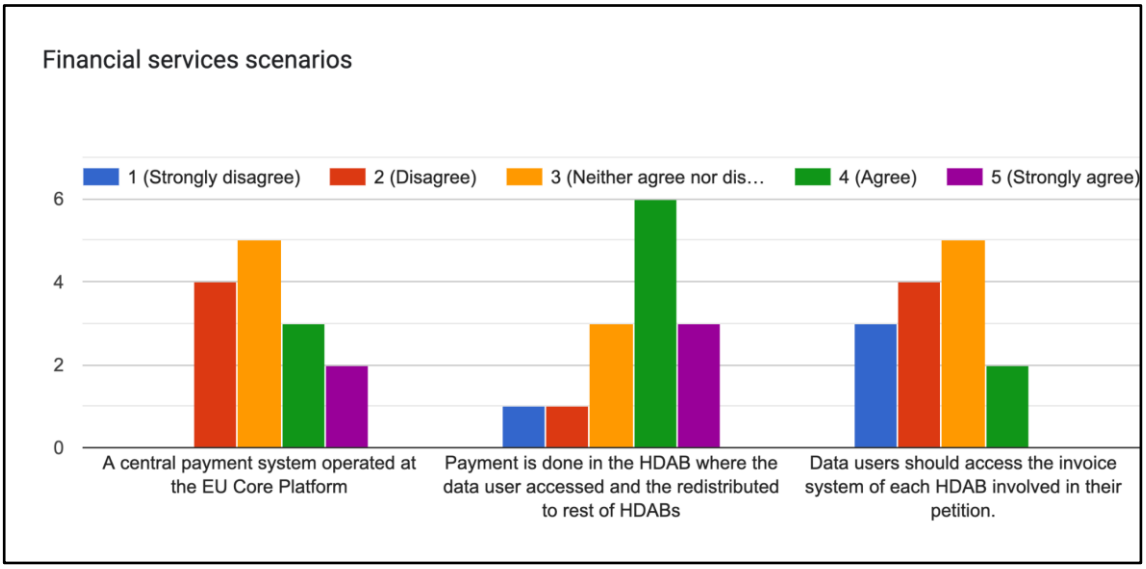


	Support services are provided at EU Core Platform	HDABs offer support to its users, which is coordinated with the EU Core Platform
Average	3.143	4.429
Stdev	1.027	0.646

D.26.1 Comments

- Support services by HDABs will be better suited to consider and address national-level idiosyncrasies - national-level HDABs must have their own "voice" when offering support. High coordination will be an unavoidable necessity.
- Training should be as close to the expertise as possible and tailored to the data user. Thus, possibly a need for training at both levels, but with different components. In our experience in SE it is crucial to identify and train super users that can build networks and spread their training in their own organisations.
- A combination could be sensible

D.27 Financial services scenarios



	A central payment system operated at the EU Core Platform	Payment is done in the HDAB where the data user accessed and the redistributed to rest of HDABs	Data users should access the invoice system of each HDAB involved in their petition.
Average	3.214	3.643	2.429
Stdev	1.051	1.151	1.016

D.27.1 Comments

- It would be counterintuitive to users if their national HDAB offers support, provides data access application management, and offers the use of an SPE - but does not take their payment. However, the matter of fees and payments is a complex issue which will likely merit a high-level negotiation in the near future; for that reason, it is difficult to strongly endorse a single solution.
- Payment of consumption to those who provide the service
- One-stop-shop is most user friendly, however, the one-stop-shop may need to be at MS level.
- I have no opinion on financial services