D6.3 Specifications for data access

Version 1

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M26, April 2023 (extended to M27, May 2023)
Dissemination Level	Public
Nature	Report
Author(s)	Eva Garcia (BBMRI), Petr Holub (BBMRI)
Contributor(s)	Celia Alvarez-Romero (SAS) Christian Ohmann (ECRIN) Danielle Welter (UNILU) Dario Livio Longo (EuroBioImaging/IBB-CNR) Davit Chokoshvili (UNILU) Dylan Spalding (CSC) Helena Lodenius (CSC) Jaakko Leinonen (CSC) Juan Gonzalez Garcia (IACS) Laura Portell Silva (BSC) Lidia López (BSC) Marco Roos (LUMC) Salvador Capella-Gutiérrez (BSC)

***Notice**: The HealthyCloud project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement Nº965345*

	Teresa D'Altri (CRG)
Reviewer(s)	Jerome de Barros (EC) Josep Ll. Gelpí (UB) Licinio Kustra Mano (EC) Sérgio Dinis (SPMS) Vanessa Lima (SPMS) Vanessa Mendes (SPMS)
Keywords	Data access, health data portal

Change log

Version	Author	Date	Description of Change
0.1	Eva Garcia (BBMRI) Petr Holub (BBMRI) Celia Alvarez-Romero (SAS) Lidia López (BSC) Laura Portell Silva (BSC) Teresa D'Altri (CRG) Juan Gonzalez Garcia (IACS) Marco Roos (LUMC) Christian Ohmann (ECRIN) Helena Lodenius (CSC)	28/02/2023	Table of contents
0.2	Eva Garcia (BBMRI) Petr Holub (BBMRI) Celia Alvarez-Romero (SAS) Lidia López (BSC) Laura Portell Silva (BSC) Teresa D'Altri (CRG) Juan Gonzalez Garcia (IACS) Dylan Spalding (CSC) Dario Livio Longo (IBB-CNR) Davit Chokoshvili (UNILU) Jaakko Leinonen (CSC) Marco Roos (LUMC) Christian Ohmann (ECRIN) Helena Lodenius (CSC) Danielle Welter (UNILU)	28/04/2023	First draft
1	Eva Garcia (BBMRI) Petr Holub (BBMRI)	18/05/2023	Final version after revision

Table of contents

Executive summary

The present deliverable takes as input the previous work done within work package (WP) 6 to identify specifications for data access through the FAIR health data portal. In order to do so, a landscape analysis was performed based on the work from other HealthyCloud WPs together with reports from related projects.

Interests of user personas, existing data hubs situation and six hypothetical scenarios were used as the basis for proposing specifications easily adaptable to different circumstances of a FAIR health data portal. Depending on such circumstances, the portal should apply the different specifications on data access application, negotiation, conditions and implementation described here.

Introduction

HealthyCloud overview

The creation of a European Health Data Space (EHDS) is a critical element of the six strategic priorities for 2019-2024 of the European Commission[1]. The European Health Research and Innovation Cloud (HRIC) will be one of the future cornerstone pieces for this area. HealthyCloud will deliver a Strategic Agenda including a Ready-to-implement Roadmap for the HRIC ecosystem.

The Strategic Agenda will incorporate the consolidated feedback of a broad range of stakeholders: the European Commission, the Member States and regional, national, European and international relevant initiatives. These agents will be invited to be part of the HealthyCloud's Stakeholders' Forum, designed to facilitate the dialogue among them and the Consortium, and to act as an umbrella to bring together similar efforts in specific domains. The draft Strategic Agenda is already available in Zenodo[2] and it includes a set of 10 services for the future HRIC to cover. The service related to the FAIR health data portal is service 7: An "EOSC Health" catalogue service. This service includes identifying and helping to recruit and coordinate services to EOSC, in the particular domains of health-related research.

The ultimate goal of HealthyCloud is to propose an ecosystem that builds and reinforces the trust of patients and citizens in the use of their health data for research.

[1] Priorities 2019-2024 (europa.eu)

[2] HealthyCloud Strategic Agenda for the Health Research Innovation Cloud (HRIC) - First Draft

Background

The following projects, related to data access, were kept in mind when shaping this deliverable.

TEHDAS

The goal of TEHDAS, as it is explained in the project's website[3], 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.

TEHDAS, the joint action Towards the European Health Data Space, helps EU member states and the European Commission to develop and promote concepts for the secondary use of health data to benefit public health and health research and innovation in Europe.

The results of the TEHDAS project will provide elements to the European Commission's legislative proposal on the European Health Data Space as well as support the pan-European dialogue that will follow the proposal.

HealthData@EU Pilot (EHDS2 Pilot)

The HealthData@EU Pilot project[4] will build a pilot version of the European Health Data Space (EHDS) infrastructure for the secondary use of health data "HealthData@EU" which will serve research, innovation, policy making and regulatory purposes. The project will connect data platforms in a network infrastructure and develop services supporting the user journey for research projects using health data from various EU Member States. It will also provide guidelines for data standards, data quality, data security and data transfer to support this cross-border infrastructure. Priority services include a metadata discovery service and a common health data access request. The consortium will collaborate closely with the European Commission and their team working on developing the central services for secondary use of health data.

FAIR principles

Even though the FAIR principles were the focus of the previous deliverable of this WP (D6.2 "Specifications for the FAIR data portal"), all of them are interlinked. Indeed, Findability is

[3] TEHDAS project website
[4] HealthData@EU Pilot project

closely related to Accessibility, which is the main topic of this deliverable. Actually, accessibility has to be considered before findability in cases where the risk of identifying individuals by a minimum amount of information is high, such as is the case for rare diseases. In those cases, a stepwise process of gaining access to increasingly detailed data by increasingly elaborate assessment of data use conditions and user profiles is appropriate. Capturing data use conditions 'for machines' to enable responsible automation is therefore an active research topic for implementing FAIR principles for health data.

Regardless of the domain of study, different granularity levels of discoverability allow to maximise the number of users that can search for data, because they may have different technical skills (as indicated in D6.2) but also because they may have different rights to access to detailed metadata. For instance, a low entry-barrier based on general statistical descriptors can be usually open to all users, in contrast to federated queries on structured data, that require more restrictive measures. The latter also requires the standardisation of the data at source, applying different widely used standards appropriate to each domain. Standardisation of data use conditions is also needed for ensuring machine-actionable search and access to data, using Open Digital Rights Language[5] (W3C standard) or DURI[6] (Data Use & Researcher Identities) and DUO[7] (Data Use Ontology) from GA4GH (Global Alliance for Genomics and Health)[8], among others, as explained in D6.2.

Objectives and linkage with other work packages

This deliverable is focused on describing specifications of how the process of data access should look like and follows the lines of the previous ones from this WP, building on the specifications that a FAIR health data portal should have (for definition of data portal please see D6.2). As this definition is broad, different accumulative scenarios in terms of data access are possible (Figure 1), from very basic functionality to most advanced interaction (full provision):

- Scenario 1 (S1):
    - The portal acts as an aggregator, gathering together all the different data hubs and providing links to them, being a catalogue of data hubs. Of note, in this project and according to HealthyCloud's glossary[9], data hubs are those infrastructures that meet the following minimal criteria:

[5] Open Digital Rights Language (ODRL) Version 1.1
[6] DURI (Data Use & Researcher Identities)
[7] DUO (Data Use Ontology)
[8] GA4GH (Global Alliance for Genomics and Health)
[9] Glossary of commonly used terms in the field of health data research - developed by the EU project HealthyCloud

- A digital technical infrastructure with the core mission of enabling health data sharing.
- It provides health data from different sources.
- It allows discovery of health datasets.
- It has a metadata discovery service.
- It has a data accessibility mechanism in accordance with existing regulation.
- It has an authorization functionality, provided by the same Data Hub or by an external institution.

A parent definition for "data hub" is "infrastructure provider", which is the responsible organisation to support the physical management of health-related data following existing regulations. Infrastructure provider is not only the parent definition for data hub, but also for data collection and secure processing environment (SPE).

- Scenario 2 (S2):
  - In addition to the functionality described above, the portal will provide a description and comparison of data access conditions of each of the data hubs.
- Scenario 3 (S3):
  - This scenario adds a new feature, making possible the search of datasets' descriptions through their metadata.
- Scenario 4 (S4):
  - The portal will provide a single access form that the user fills in and is sent directly to the hubs, which continue with the rest of the data access process.
- Scenario 5 (S5):
  - In this case the portal, besides providing the form, acts as facilitator during the whole process, being an intermediary between the users and the data hubs.
- Scenario 6 (S6):
  - The portal facilitates actual authorization and access to the data, which could only incorporate data hubs with at least joint controllership over their data.
    - 6a: the portal manages authorizations and access, facilitating/supporting the data access when it is granted.
    - 6b: the portal is somehow involved in the access process (e.g., supports expedite access).

  These two possible scenarios are being developed by the European Commission, in a joint effort of the participants in the HealthData@EU Pilot project[10] and the Central Services. This endeavour aims also to ensure that the access forms include information needed for data access in the countries where the application will be sent to, including the features from less complex scenarios (e.g., S2).

[10] HealthData@EU Pilot project

Of note, it is out of the scope of this deliverable to discuss which one is the best scenario or even the number of possible scenarios or their sustainability, they are just presented here to facilitate the discussion on the different recommendations that could be needed in the different approaches. Similarly, responsible research[11], ethical aspects and global data policy about legal aspects for data accessibility are not the focus of this deliverable, being addressed by WP2. Finally, this deliverable does not intend to give guidelines or information about quality, provenance, or interoperability of data.

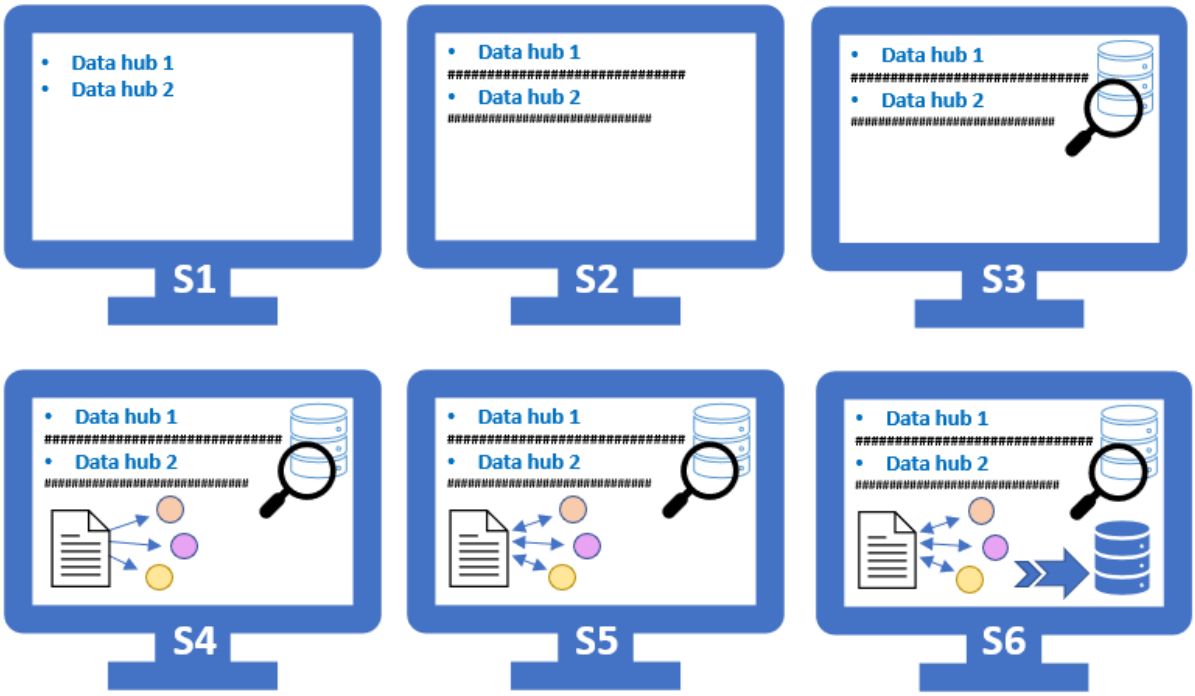

Figure 1. Schematic representation of the six hypothetical scenarios.

As mentioned above, access recommendations in this deliverable are based on the different levels of complexity that can be envisioned for the portal. In addition, we have selected the goals, challenges, needs and expectations of each user persona from D6.1[12] concerning data access to act as the basis of this work (Table 1). New personas that were deemed to be relevant for the present report (namely donor, patient, patient representative and research participant) were added to the table together with their interests and are being taken into consideration to be included in an update of D6.1, together with minor modifications in some of the interests (all changes are denoted in the table with asterisks). These definitions of personas taken from 6.1 are helpful to address the access specifications that a FAIR data portal should have. However, it is important to take into account that the role of a concrete

[11] Responsible Research & Innovation (RRI) – what exactly is it?

[12] FAIR Health Data Portal expected users' interactions

user can include more than one persona definition in terms of access and, as a result, that user will have the interests of several personas at once.

Finally, data hubs' responses on accessibility aspects to the survey made by WP3 and WP4, together with the extracted conclusions collected in deliverables D3.2[13] and D4.2[14] were consulted as part of the landscape analysis, to have the perspective of some existing data hubs also included in this deliverable. Concretely, deliverable D3.2 provided an analysis to assess the FAIRness maturity levels. Regarding accessibility, the strong recommendation is related to data access procedures or protocols defined and publicly accessible. That is, establishing formal procedures to data access and transfer through a secure processing environment. In this sense, deliverable D4.2 included the analysis of the results of the survey conducted between January 2022 and July 2022. In total, 42 health data hubs across Europe answered the survey and a set of best practices were gathered from them, concluding that data must be accessible; the access conditions must be published; the users must register previously to access; the sensitive data must be encrypted; and approvals must be managed.

[13] HealthyCloud D3.2 Guidelines to standardise metadata templates and assessment of FAIRness maturity levels

[14] HealthyCloud D4.2 Report on current discoverability solutions and FAIR adoption level

Table 1. Personas' interests (from D6.1[15]) relevant for data access.

Personas	Type of interest	Description
Citizen Donor Research participant Patient and patient representative Brief description from D6.1: *Individual who wants either to get information from one or more scientific biomedical disciplines or to contribute to a citizen science initiative.*	Goals	[No access-related items identified from D6.1]
	Challenges	[No access-related items identified from D6.1]
	Needs	Know where their data is being used and how.
		Understand how secure the ecosystem is (data and communication).
		Know who is accessing the data (researchers or companies).
		Decide whether they want to be notified about the use of their data or not, since it could be overwhelming.**
	Expectations	Want to be informed and potentially involved in data release, modulate use of data. Also want to know what data is being used.*
		Depending on the informed consent, they may want to be informed about the use of data and the results of a research/study for which the consent was given (according to GDPR[16]).*
Researcher *Individual that will interact with the future FAIR health data portal to obtain, process, produce, analyse, deposit or share research*	Goals	Want to do analysis with the data
	Challenges	Do not know how to ask permission to use the data found through the FAIR health data portal.
		Waiting too long to get access to the data

[15] FAIR Health Data Portal expected users' interactions
[16] GDPR (General Data Protection Regulation)

data and its potentially associated outcomes.		Programmatic (computational) access to health-related data may be too cumbersome
		Complex user-interfaces might be overwhelming
	Needs	[No access-related items identified from D6.1]
	Expectations	Straightforward user interfaces for achieving their goals [...] better understanding how to share/access to data [...]
Technical-oriented researcher *Researcher with higher technical expertise, which include for instance software engineers and data scientists.*	Goals	Want to perform complex data analyses using their own algorithms.
	Challenges	Do not know how to ask permission to use health-related data found through the FAIR health data portal.
		Have regularly updated documentation of the existing APIs for discovering and accessing health-related data across the different providers.*
		Waiting too long to get access to the data.
	Needs	A reference portal where to find information about, and direct links to, the infrastructure

		providers.*
		Effective programmatic means to discover/access/process relevant metadata and data for their research.
	Expectations	Machine actionable FAIR data
		Documentation on protocols for requesting access to sensitive data.*
Policy and decision maker *Individuals that gather information through consultation and research.*	Goals	[No access-related items identified from D6.1]
	Challenges	Access to heterogeneous data sources, which might be geographically distributed and may fall under different legal frameworks.
	Needs	A reference place to gain access to heterogeneous health-related data sources, including aggregated information about specific healthcare aspects or data usage patterns.
	Expectations	[No access-related items identified from D6.1]
Healthcare professional *Person that works in the healthcare sector and has an active role in providing health-related data, which can eventually have a second use for research purposes.*	Goals	[No access-related items identified from D6.1]
	Challenges	How to handle incidental findings (procedures must be included in the DTA)**
	Needs	[No access-related items identified from D6.1]
	Expectations	[No access-related items identified from D6.1]

Data curator *Individual responsible for the quality and FAIRness of health-related data, and to make sure data is discoverable and accessible.*	Goals	[No access-related items identified from D6.1]
	Challenges	[No access-related items identified from D6.1]
	Needs	[No access-related items identified from D6.1]
	Expectations	Easy-to-contact with the primary data providers for better understanding how data were collected and generated.
Data manager *Person that ensures a correct flow of the data, which implies a holistic approach to how data is collected, used, re-used and potentially shared.*	Goals	Provide guidance to researchers for proper management of health-related data, including the implication of data access and sharing
	Challenges	[No access-related items identified from D6.1]
	Needs	[No access-related items identified from D6.1]
	Expectations	[No access-related items identified from D6.1]
Infrastructure provider *Individual working in the responsible organization to support the physical and digital*	Goals	Want to facilitate data access to those who have the rights for it.
		Want to enable access control to data providers to manage access to available data.
	Challenges	Limited awareness of the existing mechanisms for trustworthy and secure data access and sharing by data providers.

management of health-related data following existing regulations.	Needs	The communication between evaluators of the application from different infrastructure providers to be facilitated.**
	Expectations	Single sign-on mechanisms available through the portal for facilitating users authentication and authorization on the connected data providers.*

*Slightly modified from D6.1.

**New, will potentially be added to an update of D6.1.

Methods

For a landscape analysis, a search was performed through materials provided by other projects that also address data access and usage. Two projects were found especially useful for this goal, EGI-Engage[17] and CORBEL[18]. As task 6.3 is closely related with the Joint Action TEHDAS, the work that is being done within its framework was also taken into account. In addition, two research infrastructures, namely ECRIN and BBMRI-ERIC, provided further information on this topic.

Synergies can also be found within HealthyCloud. The surveys done for milestones M4.2 "Study: patterns of governance of selected data hubs" and M4.3 "Study: data hubs usage current metrics" and deliverables D3.2[19] and D4.2[20], provide useful links to information about access policies of data collections and data hubs. Thus, we read through them, making a summary of the most common procedures. Notably, when a data collection or hub is used as an example, it does not mean that it is the only platform applying the described methodology.

The aforementioned steps led to the achievement of the milestone M6.3 "Study: existing mechanisms for usage of and access to already structured and organised datasets". After it, several discussions took place during the WP6 regular meetings and *ad hoc* meetings in order to shape the content of this deliverable identifying:

- The different scenarios for the portal in terms of data access.
- The interests from user personas that are relevant for accessibility purposes.
- Further connections with other WPs.
- The key points of accessibility that are related with the above points.
- The content of each section of the deliverable.

[17] EGI-Engage
[18] CORBEL
[19] HealthyCloud D3.2 Guidelines to standardise metadata templates and assessment of FAIRness maturity levels
[20] HealthyCloud D4.2 Report on current discoverability solutions and FAIR adoption level

Data accessibility

Landscape

The access mechanisms highly depend on several features of the data and the sources.

Data characteristics

Publicly available data and controlled access data

Depending on data types, the openness of the data changes, including levels of granularity[21]. Different types of secondary use are described in a report from EOSC-Life WP14[22] on policies for secondary use of data from the COVID-19 Repository:

- Publicly available data, which is out of the scope of this deliverable.
- Publicly available data after user identification.
- Data under controlled access, which can be managed directly by the data provider or by the repository following the indications of the data provider.
- Data that can not be downloaded and can only be accessed through the repository.

Data size

It is important to take into account this property of the data in order to determine the best way to implement the access to them. For instance, if it is a small dataset it could be easily downloaded, so this mechanism could be considered, even though it is a practice to be extinguished, especially with personal data (independently of being anonymous or not). Indeed, the EHDS legislative proposal[23] foresees access to health data to be provided exclusively through a secure processing environment (see below), regardless of the size of the dataset. In addition to legal considerations, for larger amounts of data downloading may not be feasible and other computational solutions should be chosen, including data visualisation or bringing algorithms to the data without actually accessing them. Of note, in this reasoning we are just considering the size, but other factors should be examined when implementing the access to the data, namely the access conditions that infrastructure providers and data hubs apply to them, and data protection or issues regarding sensitive (health) data.

[21] Open Data Platform: Requirements and Implementation Plans
[22] EOSC-Life WP14: COVID-19 Repository Data Sharing Policy
[23] European Health Data Space (EHDS) legislative proposal

Organization of the data sources

According to a published TEHDAS report[24] that gives an overview of data access of five health data platforms based on different strategies, not only the difference in the data types must be considered, but also the differences among the institutions and their organisation (centralised systems vs distributed systems).

Traditional vs machine-driven data access

Characteristics of data sources are also key when talking about the transition to digitalised access of data. We use this term to refer to the change from traditional (human mediated) to computational (machine actionable) ways of accessing data. Here, we would like to reflect on the main pros and cons of both approaches (summarised in Table 1):

Machine actionable data access mechanisms are those that, based on pre-established conditions and rules, manage and handle applications with minimal human intervention. This would fulfil some needs and expectations of technical-oriented researchers, namely effective programmatic means to discover/access/process relevant (meta)data for their research and machine actionable FAIR data (Table 1), as they present the following advantages:

- Transparency of access.
  The portal must ensure transparency of access with regards to the protocols (technical approach to request and access data) and conditions (the requirements that a user must meet and the rules he/she must follow in order to access the data) that applies to the data access.

  The extent to which the portal can improve transparency of access depends on the different scenarios. In scenario 1, this can be approached in a way that all data hubs included in the portal must have available information about the protocols and conditions for data access. In the next scenario, this information should be made available through the portal. For the remaining scenarios in which the portal is somehow involved in the data access procedure, apart from the information coming from the infrastructure providers, the portal should make its own documentation about both aspects, ideally in a human and machine-readable way. Additionally, in these scenarios the portal should keep track of all access applications, granted and rejected ones, and the actual access to data. This information must be kept up-to-date and be made available upon request. On top of that, the portal could extract and make public some metrics of data access, without individualising per infrastructure provider, since this is something more sensitive and will be briefly discussed later in this

[24] TEHDAS scrutinises data access processes in four countries

document (see Applications' metrics). Notably, all these activities related with logs and tracking are simplified if the majority of the work is done in a standardised way by machines with non or minimal human intervention.

- Speed.
  As the process can be automated, the access application is usually faster than the traditional ways, especially after putting it in place for the first time. To speed up the complete process to the provision of the data, this could be complemented with solutions as those proposed in the EHDS, mandating specific timeframes for the handling of an application and for making the data available once access has been authorised.

- Reproducibility and low variability.
  Machine controlled data access can apply DUO (Data Use Ontology)[25] to the data application information and check if it is approved or not. Thus, it is algorithm deterministic, reducing variability. However, tagging all the original datasets is an entry barrier of this approach.

On the contrary, the main drawbacks of this mechanism are:

- Barrier on the adoption (users and hubs).
  Here the data hubs organisation is again a key aspect, as the implementation of machine actionable data access heavily depends on their structure. Nonetheless, the user's perspective is also important for this point, as programmatic access to health-related data may be too cumbersome as reflected in the challenges envisioned for researchers (Table 2).

- Install and deploy technology in the hubs.
  Even for those data hubs where the implementation of machine actionable data access is feasible, it requires an important effort at the deployment and implementation phases.

Broadly, traditional ways are the ones that are almost entirely managed by humans, in the sense that they do all the steps needed for the data application, from the reception of the data to granting access. As can be told by their name, these are already in use in the majority of the data hubs and are widely adopted by the community. Indeed, this is the first of their advantages:

- Commonly used approach.

[25] The Data Use Ontology to streamline responsible access to human biomedical datasets

- Individualised interpretation of each application.
  It may be the case that access applications are not straight-forward and a dedicated specific analysis of them is needed. In these cases, traditional ways allow *ad hoc* interpretation of each application.

Despite their widespread adoption, these traditional methods have the following disadvantages:

- Time consuming.
  Conversely to machine actionable methods, traditional ways can not be automated and, as a consequence, the amount of time spent on each application can not be reduced.

- Lower degree of reproducibility.
  The individualised interpretation of each access application means that the results are not as reproducible as they are using computational approaches. Nonetheless, there are efforts trying to reduce this variability, such the policy from GA4GH on DAC (Data Access Committee) procedures[26].

In any case, access conditions must be clearly established, defined and explained. Otherwise the likelihood of not gathering the right information from applications and, as a consequence, the possibility of not getting an access approval increase. Ideally, a data access model should be in place, providing guidelines and details of how it works (e.g., EGA Data Access Model[27]).

Table 2. Advantages and disadvantages of traditional and machine-driven approaches.

	Machine-driven approach	**Traditional approach**
Pros	Transparency of access	Commonly used
	Speed	Individualised interpretation of each application
	Reproducibility and low variability	
Cons	Barrier on the adoption	Time consuming
	Install and deploy technology	Lower degree of reproducibility

[26] Global Alliance for Genomics and Health: Data Access Committee Guiding Principles and Procedural Standards Policy
[27] The European Genome-phenome Archive in 2021

Data access application

Controlled access for sensitive data is the main focus of this deliverable. This kind of data needs to be requested before being accessed. Importantly, when applying for access to these data it is necessary to perform specific queries to get access to those data that are needed for a concrete purpose, meaning that these applications are not supposed to grant access to all data available. Data minimisation requirements are defined by the infrastructure providers and dictated by the nature of data discovery and analysis services enabled by them (assuming they retain controllership for access) and not by the FAIR Health data portal.

The platforms included in the aforementioned TEHDAS report[28], have an electronic request system, where candidates should fill documentation including the purpose of the requested data. This finding is in concordance with what can be extracted from the surveys in WP3 and WP4, since most infrastructure providers and data hubs have access forms as the preferred mechanism for making these applications[29]. In general, at least the following information has to be provided in the electronic application:

- Purpose for which the data is requested (the research plan/project)[30,31,32]. Similarly to users, this can be limited, for instance, to statistical or scientific research, even though this limitation can be flexible[33]. The purpose and the users that can have access to a dataset could also depend on the dataset itself[34,35].
- Description of the requested data.
- Ethical approval if needed due to the research purpose and the type of data requested.
- Further information can be requested such as funding information, research team, publication or data management plan[36,37].

However, application forms differ among institutions. As explained in the above scenarios, the portal could act in different ways with regards to this requesting process:

- It could bring together different data hubs (S1) or even give guidelines to the users about how the access procedure works and how to apply for access, including for instance the variables they should fill in order to apply for data access (S2 and S3). In such a case, the portal should retrieve this information from the infrastructure

[28] TEHDAS scrutinises data access processes in four countries
[29] HealthyCloud D4.2 Report on current discoverability solutions and FAIR adoption level
[30] Statbel
[31] Finnish Social Science Data Archive
[32] THL Biobank
[33] Statbel
[34] Finnish Social Science Data Archive
[35] European Genome-phenome Archive (EGA)
[36] THL Biobank
[37] National FinHealth Study

providers. This can be done in several ways. It could lie on the portal side, for instance, running periodical surveys, or on the infrastructure provider side that should update the portal information every time it changes. Actually, according to D4.2[38], 40/42 of the data hubs that participated in the survey publish the data access conditions with tangible information and the 32/42 describe in their website their data access protocols.
Another way to keep this information up-to-date is establishing communication channels with the infrastructure providers (see below). Ideally, this information should be automatically updated with minimal human intervention, but this implies that the infrastructure providers have this information in a machine-readable way. For this option, it is important to consider incentives for infrastructure providers to actively contribute and regularly update this information. The EHDS proposes two set of incentives: mandating participation by law (i.e. for data providers to make datasets known, and for Health Data Access Bodies (HDBAs) to maintain an up-to-date datasets catalogue) or by proposing a fee structure for making data available (to be paid by data users to data providers and intermediaries - HDABs).

This way the portal would take care of the needs, expectations and challenges (Table 1) from health care professionals and researchers, especially technical oriented ones, having:

- A reference portal where to find information about, and direct links to, the infrastructure providers.
- Information about how to ask permission to use health-related data found through the FAIR health data portal.
- Documentation on protocols for requesting access to sensitive data.
- Updated documentation of the existing APIs for discovery and accessing health-related data across the different providers.

In addition it would help citizens, donors, research participants and patients to understand how secure the ecosystem is (data and communication). The goal of data managers can also be achieved this way, as their guidance to researchers for proper management of health-related data, including the implication of data access and sharing could be included here. Finally, this would mitigate the challenge of infrastructure providers on limited awareness of the existing mechanisms for trustworthy and secure data access and sharing by data providers (Table 1).

- The portal could also act as a broker, in terms that it has a single form that users fill out and reaches the different infrastructure providers in the portal (S4-6). Actually, this would fulfil the needs and challenges of policy and decision marker personas, who

[38] HealthyCloud D4.2 Report on current discoverability solutions and FAIR adoption level

need a reference place to gain access to heterogeneous health-related data sources, including aggregated information about specific healthcare aspects or data usage patterns and see as a challenge to access to heterogeneous data sources, which might be geographically distributed and may fall under different legal frameworks (Table 1). For having such a platform, the information of what should be collected in the forms is still needed. In addition, it has to fulfil the requirements from the different providers, which could be significantly different. This can be done by two different approaches:

- Finding a common ground that covers the minimum information needed from each infrastructure provider. This aligns with one of the main propositions of the EHDS, harmonising the legal frameworks in the countries of data providers.
- Having dynamic forms that adapt depending on the target infrastructure provider.

In both cases, a machine-readable, standardised form with established vocabulary and semantics would need to be developed first to make sure that the information provided is understood in the same way by all actors and, once the application is filled, the portal should programmatically send it to the different providers. It is likely that both approaches, especially the first one, would require providers to request additional information to complete the access application and this can be done in the data access negotiation step.

Data access applications' status

To support scenarios S4-6 it is important that the portal has in place a method to resume and check access applications.

Save ongoing data access applications

This refers to ongoing non-submitted applications (Figure 2a). Basically, this feature would allow a user to leave the application form at a specific moment and continue later on, without losing the work done. To make this possible, the portal should save this kind of requests during a short period of time, without sending them to the infrastructure providers.

See status of the data access application

The portal should inform the user about the status of the application, before and once it is submitted to the infrastructure providers (Figure 2a,c). A dashboard that shows the status of the application, especially if it involves more than one provider, is a user-friendly approach to give a quick overview to the different actors (the portal, infrastructure providers and users). The information provided in this aspect comes from the infrastructure providers and an alignment with them about what and how the information should be shared is needed.

Applications' metrics

An aspect that is not specifically about access but is very tight with the above topic and scenarios S4-6 is how to track applications in terms of number of requests per infrastructure provider, time to process each application, etc. Here it should be defined who provides and has access to this information as well as if it can be made publicly available or not (D4.3[39], D4.1[40]). Generally speaking, even though the portal might have its own metrics, the information on the measures from the infrastructure provider should be provided by each of them to the portal so it is always consistent. On the other hand, the portal must make sure that the infrastructure providers receive information and then take into account the applications received and/or managed from the portal so they do not see the figures drop due to the applications made through it.

These metrics are also important when looking at one of the needs of the first persona definition in Table 1, as they need to know who is accessing the data (researchers or companies) and where their data is being used and how. Finally, this feature could be used to raise awareness of project results deriving from data access facilitated through the portal, increasing its value for individuals/general public and ensuring transparency.

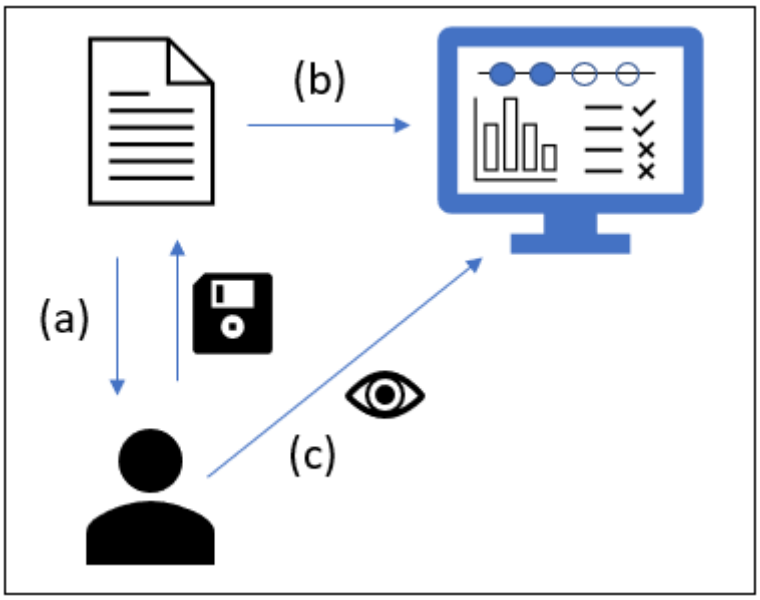

Figure 2. Data access application steps.

[39] Reference to be added when published.

[40] HealthyCloud D4.1 Recommendations for integration in HealthyCloud, including an analysis of data hub patterns of governance

Data access negotiation

Communication channels and project follow-up mechanisms are necessary for those scenarios in which the portal acts as a facilitator of the whole data access application process (S5 and 6).

How to handle communication

For the access negotiation step, one of the most important points to be taken into account is the interaction between the different actors. Actually, this recommendation aligns with the following data curator's expectation: easy-to-contact with the primary data providers for better understanding how data has been collected and generated. It also reflects the recommendation in D3.2[41] on including communications protocols in data access procedures.

Data applicant - Infrastructure providers

In those cases that communications between data users and infrastructure providers are made through the portal (Figure 3a), it should cover the following needs:

1. Refinement of the queries.

   As the search that can be made through the portal is on aggregated data, it is recommended to have a mechanism that allows the refinement of the queries: this implies bidirectional communication, since it could happen that the user needs to redefine the request (e.g., he/she needs more data or has more specifications of the data than in the initial step). Similarly, this refinement can also start from the infrastructure provider side, when more information is needed to fully understand the access application.

2. Information on availability.

   Once the application is clarified as much as possible, the infrastructure providers must show the availability of the data requested on their side. The timing of response of each infrastructure provider is a key factor here. If the infrastructure providers shown in the portal – or more precisely, the ones that are the target of the request – share similar reply times, then the situation where the portal communicates all decisions at once to the user is naturally the best option. Conversely, if the time of response from the infrastructure providers differs, the portal should update this information periodically, so the requester can have at least access to batches of data.

3. Selection of the available data by the user (by default all – but possibly a subset)

   The availability/unavailability of data must be communicated to the requester through the portal and, based on their description and access conditions, she/he must be able to choose which data are suitable for her/his project.

4. Communication of access decisions.

[41] HealthyCloud D3.2 Guidelines to standardise metadata templates and assessment of FAIRness maturity levels

After that, infrastructure providers must consider the application and proceed with their own access committees to decide if they grant access to the data and in which conditions (see section below).

5. Data Access Agreements preparation.

Usually the topics handled in this phase are quite sensitive and, because of that, in most cases they are handled directly between the requester and the infrastructure providers or the data hubs. However, the portal could facilitate this process providing a communication interface where both actors can discuss and share documents in a bilateral confidential way.

This channel of communication could be used as well when the requester profile matches the interests of the citizen, research participant, donor or patient personas, since the majority of their interests are focused on information and decisions on their data (Table 1). This way the portal could act as an intermediary between them and the infrastructure providers, while it is not involved in the agreements between these both actors. This could also apply for communications regarding incidental findings, a challenge for healthcare professionals. Nonetheless, how to proceed in such situation must be stated in the DTA (Data Transfer Agreement), including decisions on communication aspects.

Data access committee between requester and infrastructure providers/data hubs: on the data hub side.

Importantly, as each data hub works internally with their access committee/s, likely the portal does not intervene/facilitate anything here in any sense.

Communication among different infrastructure providers

Another specification that could facilitate the process in such an ecosystem is the communication between different infrastructure providers (Figure 3b). This could help them with the decisions, for instance if they share the same access committee, so they do not have to ask twice for the same access application. Indeed, this aligns with the need of facilitating the communication between evaluators of the application from different infrastructure providers (Table 1).

Communication portal - infrastructure providers

Finally, as the portal will be working closely with the infrastructure providers, a suitable communication channel must be set up (Figure 3c). This should be easily distinguished from the communication with the users and should also keep track of the discussions about the features of the portal, such as the application metrics discussed above.

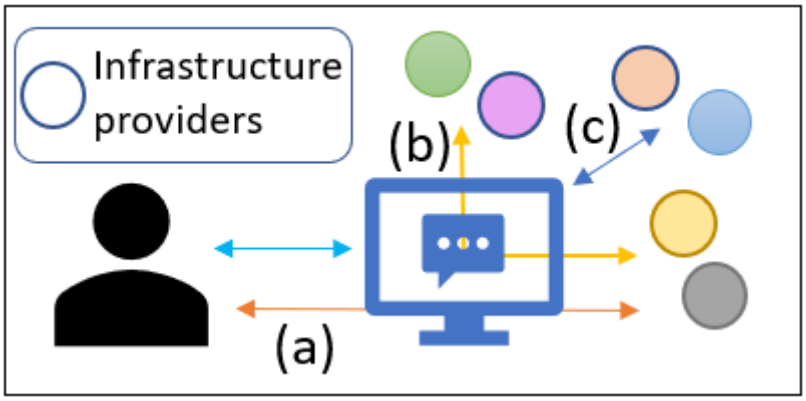

Figure 3. Schematic representation of the communications channels during data access negotiation.

Application (project) progression

Applications are based on projects/activities that users want to carry on using the data they are requesting access to. As such, the applications made through the portal must accept amendments and follow-ups. This procedure could speed up the process, since neither the users nor the infrastructure providers need to start from scratch. This should be done at both ends of the access application:

- User:
  - In order to be able to follow up with an application, they must be saved in a "history of applications". For providing this feature, the portal must have a mechanism to register users. If the idea is to register users only to provide this kind of tracking service, a low barrier registration system should be sufficient.
- Infrastructure providers:
  - The portal should maintain the history of applications and the derived applications that could arise from an initial one.

Data access conditions

Broadly, access conditions are the requirements that a user must meet and the rules he/she must follow in order to access the data. The users that can access health data are not always the same, i.e., some institutions only allow access to other institutions and not individual users[42]. According to the EOSC-Life WP14 report[43], data object providers have the control over the data objects as well as the decision of how they should be shared, complying with the corresponding regulations. This can be handled by the data repository, but it can never make a decision (but proceed according to the agreement met with the data provider).

[42] TEHDAS scrutinises data access processes in four countries
[43] EOSC-Life WP14: COVID-19 Repository Data Sharing Policy

Nonetheless, the portal can give support and features when it comes to applying the data access conditions after formal agreements with data providers[44].

Controlled access

Different levels of granularity

When a portal acts as a display that shows metadata and/or data that are present at the infrastructure providers (from S3), it has to deal with the sensitivity levels of such (meta)data. For instance, publicly available data can be made directly reachable or can be visualised without any restriction. However, this does not apply to restricted access data and the portal should provide the infrastructure providers with different levels of granularity, showing in each of them a description of the data at different levels. Indeed, integrated search of public and restricted data in the same portal is a good practice, so users can easily access both, depending on their authorization[45].

Even though this is on the edge between findability and accessibility, it is true that for some types of data (e.g., in the field of rare diseases), data searches are only possible after complying with some access conditions, so both are naturally interconnecting. When a user accesses the metadata of either a data hub or one of its datasets, he/she must be registered and agree with the terms and conditions of it as a first approach. As said, these must be set by each of the data hubs and, based on the results from the WP3 and 4 surveys, they reflect the responsibilities for those accessing the data[46], which normally include, among others, the prohibition of sharing the data without further approval and the restricted usage of them to the purpose described in the request step[47,48]. A practical consideration to this point, that will be further discussed in D2.4 “Guideline on ELSI compliant governance models”, is that data hubs participating in a HRIC can be allowed to impose contractual conditions, such as various responsibilities when accessing and using their metadata. However, these data hub-imposed conditions are not strictly GDPR-specific, because the metadata the data hub provides for resource-level discoverability purposes in a HRIC are not personal data and should be treated accordingly.

Usually, when accepting the Terms and conditions, the requester also agrees on being compliant with national laws and acknowledging the use of the data in the results of the project. In that document, the users must be informed if the data provided by them in the

[44] BBMRI-ERIC Colorectal Cancer Cohort (CRC-Cohort): Data Protection Policy
[45] Open Data Platform: Requirements and Implementation Plans
[46] Research Services at Statistics Finland
[47] THL Biobank
[48] EUROCAT Central Registry

request process is stored and retained by the platform[49]. The criteria of request approval can also be reflected[50].

How to handle restricted access

The surveys done by WP3 and WP4, show that for accessing restricted data (e.g., sensitive data such as individual data[51]), users must be registered. According to the results reported by TEHDAS, the users that have rights to access the data can be restricted, for instance, to those belonging to an institution[52,53,54].

Hence, for cases that require higher data protection, users must be authenticated and authorised to access the (meta)data they are aiming for. Here the portal can also be of help for the data infrastructure providers, using an AAI (i.e., Authentication and authorization infrastructure). Interestingly, as part of the EOSC-Life project, the Life Science Login is already in production[55,56,57] and provides a common AAI for different research infrastructures. This would allow infrastructure providers to facilitate data access to those who have the rights for it and to have a single sign-on mechanism available through the portal for facilitating users' recognition and authorization on the connected data providers, addressing some of their goals and expectations (Table 1).

Access control organisation

Data controllers that are part of the portal always keep their sovereignty on the data and are the ultimate responsible for data controllership, being the portal just a platform to facilitate and/or accelerate access, depending on its architecture. Actually, the role of the portal is different depending on the access control organisation. Further alignment with the HealthData@EU infrastructure access control processes is not possible at the current moment due to the status of the legislative process as the technical specifications of the data access will be developed as part of the implementing acts listed in Article 45(6) and Article 52(13) of the legislative proposal, among others.

[49] Finnish Social Science Data Archive
[50] THL Biobank
[51] GDPR (General Data Protection Regulation)
[52] Finnish Social Science Data Archive
[53] THL Biobank
[54] BIFAP (Pharmacoepidemiologic Research in Public Health Systems)
[55] Life Science Login
[56] EOSC-Life Access and User Management System for Life Science – the implementation and usage report
[57] EOSC-Life Access and User Management System for Life Science – the blueprint update

Committee-controlled access in federated settings where there are many controllers

In this situation the portal is not involved at all in the access control of the data itself (all proposed scenarios but S6). It can facilitate communications as shown above. In addition, in case that the application is filed through the portal, it could apply a high-level review of it, just checking that it is not obvious spam. This check could be programmatic (e.g., using some key words) or by hand (usually slower but more precise).

Support for expedite access modes for datasets

However, the data portal could be involved in the process of granting access, facilitating the actual access to the data (S6). This is especially useful for concrete datasets, collections, cohorts or *ad hoc* applications that have data from different data sources. In terms of the governance set up, the portal would act as a processor with respect to data access decisions, acting on the instructions of the data hub. The operator/s of the portal screens and assesses the eligibility of access requests based on data use conditions & restrictions defined by the data hubs, and then generates an eligibility assessment report for the data hubs' review and approval. In this situation, data hubs are the sole controller for making access decisions. It may or may not agree with the assessment of the portal. Indeed, data hubs' non-response within a defined time-frame means the access request has been denied.

This would decrease the waiting time to get access to the data, addressing one of the challenges that researchers and technical-oriented researchers face (Table 1). However, the conditions must be established with the data hubs, ensuring that they do not lose visibility due to this approach.

Implementation of data access

Once the access to the data is granted, a Data Access Agreement (DAA) or a Data Transfer Agreement (DTA) is signed[58], containing all the details of data access and usage. Afterwards, the actual access to the data has to be implemented (S6).

Technical implementation

Technically speaking, data can be accessed in two different ways:

Secure download

In this case the data is downloaded by the user through a temporary link, the Collaborative Spanish Variant Server (CSVS)[59] can be taken as an example. Using this method the data is

[58] THL Biobank

[59] Collaborative Spanish Variant Server (CSVS)

completely transferred to the data requester, who is bound by the clauses of the DAA/DTA in terms of data use.

Processing in place

The above solution could work for some data hubs, such as the 20/42 that are part of D4.2 and responded "Yes" when asked about data leaving the infrastructure or the 14/42 that indicated "Yes" providing specific conditions (e.g., DTA, only for aggregated data, etc). Conversely, 8/42 responded a hard no for data leaving the Infrastructure[60]. Therefore, a different solution is needed for such cases, where the data is processed in place and here two different situations can still be distinguished:

1. Secure processing environment (EHDS[61], Article 50).

The first one implies that the requester accesses the actual data, without downloading it. This must be done through a Secure Processing Environment (SPE, a.k.a. TRE - Trust Research Environments). The specifications of these environments are defined based on the requirements of the data and several efforts at European level are planned to further define them. Apart from the specific features that SPEs could need depending on the data, it is also important to consider who is going to provide such services:

i. Data hub provided: According to a recommendation from the EGI-Engage project[62], a method for accessing the data directly in the source is convenient for sensitive data as well as for large datasets with large files.
ii. Research provided.
iii. Third party provided.

2. Bring algorithms to the data without actually accessing them.

Here the users do not have access to the data, but send their algorithms to where data are available and they are run there. Even if requesters can not access the data, it is important to check the scripts that are sent. This approach is not always feasible from the users' perspective, since they need access to the data for intensive development of algorithms or in situations where data exploration is needed before even designing an algorithm.

These solutions are enabled by 'FAIR at source' (meaning that FAIR principles are followed from the generation of the data) and their consequences are still underexplored. Depending on the circumstances, one of them would help researchers and technical-oriented ones achieve their goal of performing analyses with the data, including complex data analyses using their own algorithms (Table 1). Nonetheless, the portal as it is described within the framework

[60] HealthyCloud D4.2 Report on current discoverability solutions and FAIR adoption level
[61] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space (EHDS)
[62] Open Data Platform: Requirements and Implementation Plans

of this deliverable and WP6, is not foreseen to host or manage the access to the secure environments in any case. However, it could act as a way to keep track/publish the availability or usage of these facilities if requested by the infrastructure providers, as 15/42 data hubs provide a safe space for users to analyse data without downloading[63].

Software deployment central/federated/hybrid

Recently, the minimum set of services in HealthData@EU have been established within the framework of TEHDAS[64]. Regarding data access applications, the software that helps in the process of requesting access can be deployed whether in the central node or distributed in the different nodes. A similar situation is envisioned for the software needed for granting access, which can be deployed by the HealthData@EU nodes in a central, distributed or hybrid mode where both parts are involved in this deployment. However, it is likely that existing data infrastructures would have their own software already in place and in this case the key point would be to have common, well-described/specified and stable interfaces rather than software implementations.

In any of the cases the user interface should be user-friendly and oriented to a broad audience with different expertise since, according to Table 1, complex user-interfaces might be overwhelming for researchers and they expect straightforward interfaces for achieving their goals (Table 1).

Project progression

Data access applications based on previous requests

This topic was already discussed above (see Application progression). Still, it is highlighted again since this might happen not also at the time that the requesters are preparing the access application, but at any time during the project development in scenarios S5 and S6. This is important to consider because of the time that the history of applications must be saved for each user.

Any clarifications on the data

It might happen that while working with the data, some questions will arise and the knowledge at the source is the best way to answer them. The history of request is also key for providing support to such situations and a communication channel through the Portal can be used for this purpose.

[63] HealthyCloud D4.2 Report on current discoverability solutions and FAIR adoption level

[64] TEHDAS suggests minimum technical services for the European health data space

Offer the results back to the data sources

It is frequent that a project results in the creation of new data that can enrich the initial dataset. A good practice in this aspect is offering these results to the data sources, so they can add them to their collections, facilitating data reuse. A way of offering and negotiating the conditions to provide these data back could be through the portal, via the bilateral communication channels between the data user and the infrastructure providers or the data hubs. This feature of the portal would also allow the infrastructure providers and the data hubs to keep track of the results of projects that have been performed with the data coming from them.

Conclusions

One of the main important takeaways from this deliverable is that, as the features of a FAIR health data portal are not completely defined yet, the specifications need to be flexible. Hence, they are based on six hypothetical scenarios that allow us to provide specifications that can be easily adapted to several situations. In addition, not only the scenarios but also the organisation of the infrastructure providers and the data hubs and the interests of the user personas are key points to consider when identifying these specifications.

Briefly, it is important that as many infrastructure providers as possible are brought together in the portal. Then, depending on the level of complexity of the portal, some requirements must be fulfilled. The most relevant ones have to do with the information about the access procedures of the infrastructure providers and data hubs, the communication channels between different actors of the data access applications and the facilitation of the actual access to the data.

Next steps

This is the last of the three deliverables of WP6 about the FAIR Health Data Portal. During the work performed to achieve these specifications, new user profiles deemed relevant in terms of data access and will be reviewed as potential additions to an updated D6.1. In addition, this deliverable aims to gather useful information regarding data access for the future versions of the HRIC Strategic Agenda.