HEALTHYCLOUD
Health Research & Innovation Cloud

D6.2 Specifications for the FAIR data portal
Version 1.0

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M21, November 2022
Dissemination Level	PU
Nature	R
Author(s)	Danielle Welter (UNILU) Wei Gu (UNILU) Venkata Satagopam (UNILU)
Contributor(s)	Laura Portell Silva (BSC) Lidia López Cuesta (BSC) Salvador Capella-Gutiérrez (BSC) Eva García (BBMRI-ERIC) Marco Roos (LUMC) Celia Alvarez-Romero (SAS) Irène Kesisoglou (SCIENSANO) Luiz Bonino da Silva Santos (LUMC) Juan González Garcia (IACS) Petr Holub (BBMRI-ERIC) WP6 meetings participants WP6 workshops participants

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

Reviewer(s)	Pascal Derycke (SCIENSANO) Jordi Rambla (CRG)
Keywords	FAIR data portal, meta-catalogue, interoperability

Change Log

Version	Author	Date	Description of Change
v0.1	Danielle Welter Laura Portell Silva Marco Roos Lidia Lopez Eva Garcia-Alvarez Petr Holub Wei Gu	2022/09/29	Table of contents
v0.2	Danielle Welter Laura Portell Silva Lidia Lopez Eva Garcia-Alvarez Petr Holub Luiz Bonino da Silva Santos Juan González Garcia Celia Alvarez Romero Irini Kessissoglou Wei Gu Venkata Satagopam	2022/11/04	First draft
v0.3	Danielle Welter Laura Portell Silva Lidia Lopez Salvador Capella-Gutiérrez Venkata Satagopam Celia Alvarez Romero	2022/11/28	Addressing reviewer comments
v1.0	Danielle Welter Wei Gu Venkata Satagopam	2022/11/30	Final version submitted to coordinators
			(Final Change Log entries reserved for releases to the EC)

Table of contents

Executive Summary

The objective of HealthyCloud deliverable D6.2 is to lay out the specifications for a FAIR health data portal, taking into account findings from previous deliverables such as D6.1[1], D3.1[2] and D4.1[3]. The key focus of these specifications is the compliance of data and metadata with FAIR principles in all aspects of the portal, from how the portal acquires metadata to how it models and presents it, to how the portal presents itself, both to humans consuming information directly and to machines tasked with finding and aggregating information on humans' behalf.

The European health data space is undergoing rapid expansion and ensuring the long-term interoperability of health data while respecting the specific access and privacy needs of this context is a key challenge. A FAIR-compliant data portal that is able to expose harmonised metadata from a range of data sources is therefore an essential component for tackling this complex and multi-faceted challenge.

In this report, we discuss the key requirements of the FAIR health data portal, in particular with respect to the features identified as essential in the deliverable D6.1 about user profiles. We explore the importance of a semantically interoperable metadata model and processes involved in bringing metadata from diverse sources into the metadata catalogue that lies at the core of the FAIR health data portal, with a view to building a central community resource that connects together and leads the health data space not just in terms of its data but also in terms of knowledge, guidance and research best practice.

1. Introduction

This report presents a detailed set of specifications for a future FAIR health data portal. It builds on previous work and deliverables from WP3, WP4 and WP6, including D6.1 regarding user profiles, D3.1 regarding existing data collections and D4.1 regarding data hubs. The recommendations made in this report will be mostly implementation agnostic, i.e. no technical specifications such as specific programming frameworks or hosting platforms will be included or any specific tools mentioned to stay solution-neutral. The major exception to this are FAIR metadata models and strategies, where some domain-appropriate solutions will be discussed.

[1] https://healthycloud.eu/wp-content/uploads/2022/11/D6.1_Updated.pdf
[2] https://healthycloud.eu/wp-content/uploads/2022/11/D3.1.pdf
[3] https://healthycloud.eu/wp-content/uploads/2022/11/D4.1.pdf

While the FAIR health data portal is of course based in the wider context of health data work in Europe, the unique focus of the portal centres around its FAIR features.

2. Summary of previous findings

2.1. User profiles

HealthyCloud WP6 is focused on defining the reference architecture for a FAIR health data portal. This portal is conceived as an access gateway for existing resources and a place for providing references to different users. The first step to reach this goal was to define the different user profiles that interact with the portal. Indeed, different users such as citizens, researchers or infrastructure providers have different expectations from the portal and what it can offer. In deliverable D6.1, the needs and objectives of the user profiles were defined in order to later on detect the functionalities that the portal should have based on their needs.

Eight different profiles and their corresponding sub-profiles (Figure 1) were considered to define the user interactions with the FAIR health data portal, which were grouped into the six orthogonal categories – 1) data generation and usage; 2) legal roles health-related infrastructures; 3) career development of a given professional; 4) intended use of health-related data; 5) professional sector of the main activities of a given professional; 6) temporal scale of the activities considered. These orthogonal concepts were designed to capture relevant aspects for the users' profiles when interacting with the FAIR health data portal.

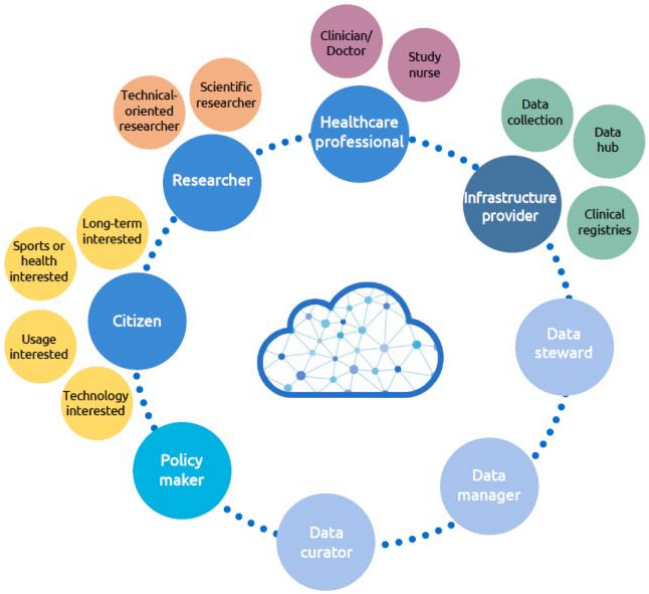

Figure 1: FAIR health data portal user profiles and sub-profiles.

In addition to the categorization of the user profiles, user personae were also defined as archetypical users whose goals and challenges represent the needs for a larger group of users. The definitions of the user personae include their skills, goals, challenges, needs and expectations from the FAIR health data portal. The definition of these user personae helped to better understand future users of the FAIR health data portal and support the designing of the reference architecture for the FAIR health data portal. Based on the user personae definition, it was possible to map the user needs and expectations to the exemplary user interactions with the FAIR health data portal, as shown in Table 1.

The user expectations extracted from the FAIR health data portal user personae definition can all fall into five different big categories:

- Share data in a secure environment.
- Find individual and aggregated data from different domain and sources.
- Find infrastructure providers for data management and analysis.
- Find guidance and best practices in different aspects.
- Access data quality validation mechanisms.

In the different sections of this deliverable the five categories are considered in order for the FAIR health data portal specifications to cover all the expectations of the user personae extracted previously.

It should be noted that some aspects of the user profiles described in deliverable 6.1, such as some of the reference requirements in Table 1, do not relate directly to the FAIR-focused scope of these specifications. These aspects will not be explicitly discussed in this work.

Table 1. Expected interactions of the different user profiles with the future FAIR health data portal.

	Citizen	Researcher	Policy and decision maker	Healthcare professional	Data curator	Data steward	Data manager	Infrastructure provider
A place to share their data in a secure environment, including easy-to-follow instruction on how to do it	x	x		x				x
Find research outcomes about a specific topic	x	x		x				
Do analysis with the data		x						
Reference place for identifying existing cohorts and creating new (virtual) ones, including documentation		x		x				
Effective programmatic means to discover/access/process data		x						
Access summarized information about healthcare trends in the general and/or disease-specific population	x	x	x					
Easy-to-combine aggregated information from different sources and/or domains			x					
Aggregated information for data usage patterns from different sources and/or domains			x		x	x	x	x
Reference place for identifying geographically distributed infrastructure providers for data management and analysis, including documentation		x						x

	Citizen	Researcher	Policy and decision maker	Healthcare professional	Data curator	Data steward	Data manager	Infrastructure provider
Reference place for best practices, guidelines and tools for working with sensitive data, including data quality		x			x		x	
Reference place for best practices and guidelines for developing and validating data-driven protocols for preventing, diagnosing and treating individual diseases				x				
Reference place for best practices, guidelines and tools to evaluate the FAIRness of datasets and contribute towards their FAIRification						x		
Access to description on data quality validation mechanisms, ideally driven by community standards					x		x	x
Reference place for best practices, guidelines and tools to work with domain-specific controlled vocabularies and ontologies					x	x		x
Reference place for best practices, guidelines and community-driven common data models							x	x
Reference place for best practices, guidelines and tools for creating and maintaining up-to-date Data Management plans (DMPs)							x	x

2.2. Data infrastructures in Europe

Deliverables D3.1 and D4.1 provided an overview of existing data collections and data hubs in Europe. Briefly, European data infrastructures constitute a complex ecosystem comprising various components with various types of interactions. The most prominent components are:

- Upcoming European Health Data Spaces (EHDS)[4], which has a component for reuse of health data for research and policy making purposes (called EHDS2). The EHDS2 consists of health data hubs that act as nodes in the network and are interfacing to all other nodes to fulfil access requests of the EHDS2 users.
- European Research Infrastructures (RIs), such as BBMRI-ERIC or ELIXIR, which are typically federated systems consisting of data and service providers and the RIs provide fabrics for data discovery and accessibility. In specific cases, some of the RIs (e.g., BBMRI-ERIC) can act also as data hubs with the responsibilities of data controllers and hence release the data directly.
- Other upcoming European Data Spaces beyond EHDS, with the most prominent examples being European Genomic Data Infrastructure (GDI)[5] and Federated European Cancer Imaging Infrastructure (EUCAIM)[6].
- There are also domain-specific health data infrastructures, such as the rare diseases, where European Joint Programming for Rare Diseases (EJP-RD) is providing a Virtual Platform[7] which provides generic services including data discovery and accessibility mechanisms and which is used by the disease-specific European Reference Networks (ERNs).

First of all, a survey was designed and developed in WP3 and WP4, joining efforts and sharing outcomes. The main objectives of the survey were:

1. To perform a landscape analysis of the different governance models in those data infrastructures; and
2. To evaluate the feasibility of linking individual-level data between data collections.

Deliverable D3.1 focused on analysing the FAIRness levels of a selection of European health-related data collections containing datasets essential to answer the research questions of the cancer and atrial fibrillation use cases. Using a catalogue matrix/survey, information about the format of the data and data quality aspects

[4] https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en
[5] https://gdi.onemilliongenomes.eu/
[6] https://eucanimage.eu/
[7] https://www.ejprarediseases.org/what-is-the-virtual-platform/

along with their compliance with the FAIR principles of each examined data collection were presented and analysed. Then a FAIRness evaluation tool was adapted and published in an open-access format on Zenodo[8] in order to assess the FAIRness level of each data collection according to the information collected during the survey. This HealthyCloud FAIRness self-assessment tool is a 2-in-1 tool allowing the publication of the HealthyCloud FAIRness evaluation survey and the production of a report including pie charts demonstrating the percentage scores for each FAIR principle as well as an overall score.

Deliverable D4.1 described the analysis of the governance patterns of the interviewed data hubs, aiming to provide specific profiles of these dedicated data infrastructures. In total, 42 out of the 99 contacted data hubs answered the survey. The work done in D4.1 is being completed with deliverable D4.2 “Report on current discoverability solutions and FAIR adoption level”, to be submitted in December 2022, and which describes recommendations on how newly created data hubs can enable the exploitation of the FAIR data benefits by design and how those hubs can integrate with the HealthyCloud ecosystem, including a set of best practices for enabling the exploitation of data collections at different FAIRness levels and with pre-established maturity levels to increase their FAIRness level.

3. FAIR data portal considerations

Before we can elaborate the specifications of a FAIR health data portal, we first need to establish a common understanding of what we mean by "FAIR data portal". This includes defining what we mean by "data portal" and how a "FAIR data portal" differs from a normal data portal.

3.1. What is a data portal?

For the purposes of this report, we define the following concepts in line with and as an extension to the HealthyCloud Glossary[9]:

- **Data portal**: in the present context, a data portal is a single point of access to data from different sources. Data are usually organised into subsets or categories based on defined characteristics to make them easier for users to find. The portal could in theory store data but in the present specification, we will not consider the storage of any data itself, only representative metadata about the data, including access links and parameters. Where possible, the metadata is automatically aggregated from the sources with little or no manual interventions, at least after some initial setup.
- **Data repository**: a data repository is an infrastructure to collect, manage and store data for the purpose of analysis, reporting and sharing. Data

[8] https://zenodo.org/record/7038397#.Y2UIXOzMKAo
[9] https://doi.org/10.5281/zenodo.6787119

repositories store both the data files themselves as well as associated metadata describing the parameters and context of the data.

- **Data hub**: the HealthyCloud glossary defines the concept of a "Health data hub" as a technical infrastructure which provides data from different sources. It is effectively synonymous with the data portal concept defined here.
- **Data collection:** a data collection is a compilation of datasets with no associated governance or technical infrastructure.
- **Data registry**: a data registry is an interactive system that collects, organises and displays information about data. Like a data portal, it only stores metadata but not the data itself. Unlike a data portal, a registry involves direct submission of metadata to the registry. Metadata submissions may be edited by the submitter over time and initial submissions may not be linked to any actual data yet. Examples of registries include clinical trials registries, tissue banks or patient registries.

3.2. How can a data portal support FAIR?

The FAIR health data portal should support the data FAIRness both up- and downstream. The FAIRer the initial data sources are, the easier it will be for the data portal to pull in compliant metadata. Equally, the inclusion or the potential for inclusion of a data source in a FAIR health data portal may serve as an incentive to the owners of these data sources to improve the FAIRness of their own data and metadata, clarify licensing and data use conditions or adopt community standards.
From the end user perspective, the offerings of a fully FAIR-enabled data portal will facilitate the interoperability and reuse of existing data. It can also serve directly or indirectly to improve downstream practices through the provision of explicit training materials and guides and the implicit "leading by example".
Through its central role in a community, a data portal is in a unique position to drive cultural change in terms of data management practices and data representation, through engagement with data providers and end users.
In order to support the FAIRness of its content, i.e. the metadata pulled from a variety of data sources, a data portal should cover the following aspects:

- **Identifier strategy**: every entry into the portal should be assigned a globally unique and persistent identifier. Although the source data will likely already have their own identifiers, identifier strategies may differ substantially between data sources, in terms of structure, policy and granularity. It is therefore necessary for the portal to assign its own identifiers following its own set of criteria, and mapping or cross-linking to the source identifiers
- **Metadata**: the primary utility of a data portal lies in the presentation of data from potentially very diverse sources under a common representation or metadata schema. In order to maximise findability, the metadata needs to cover a good range of relevant attributes, which need to be underpinned by standard vocabularies or ontologies wherever possible. If an appropriate

domain standard is available, it should be used directly for metadata representation or the metadata should at least be directly compatible with this standard. The metadata should be represented in or exportable to a formal and broadly applicable knowledge representation such as RDF that is interoperable with a wide range of other sources.

- **Indexing by search engines**: in order to maximise findability via common search engines, the data portal should expose relevant metadata using common markup strategies such as DCAT[10], SDO[11] or BioSchemas[12], for example via the implementation of a FAIR Data Point (FDP)[13,14].
- **Automated metadata retrieval**: In addition to being findable and in a standard structured format such as JSON-LD or RDF/XML, FAIR metadata should be machine retrievable, likely through an API. This also automatically fulfils the accessibility aspect of standard communication protocols as a web-based access portal will use HTTP or HTTPs communication protocols.
- **Authentication & authorisation**: it is very important to remember that FAIR is not synonymous with "free & open". Importantly, FAIR aims to provide data as open as possible, and as closed as neccesary. This is especially relevant for health-related data, which tends to be of sensitive nature. If required, a data portal needs to offer appropriate authentication and authorisation procedures to meet data access restrictions, either as a built-in feature or through the use of external identity providers and authorisation management services. In this present case, the primary responsibility for authentication and authorisation should lie with the data sources, with the portal exposing only public metadata, including the conditions under which data access is possible, so authentication is not a core requirement for the portal.
- **Formal obsolescence policies**: one often-ignored aspect of FAIR is the persistence of metadata beyond the lifetime of the data. This requires the establishment of formal obsolescence policies that define what level of metadata remains exposed following the removal of the data and how it is represented, including reasons for obsolescence, dates and, if applicable, links to replacement data records.
- **Ontologies and controlled terminologies**: a core pillar across all areas of FAIR lies in the annotation with terminologies that are themselves FAIR-compliant, e.g. use appropriate identifier, versioning and obsolescence policies, and are machine readable and actionable. Therefore, this aspect of

[10] https://www.w3.org/TR/vocab-dcat-2/
[11] https://schema.org/
[12] https://bioschemas.org/
[13] https://www.fairdatapoint.org/

annotations is essential for a FAIR health data portal because of the importance to achieve semantic interoperability. Both the portal's metadata model and the metadata represented through the model should be standardised against and annotated with community-adopted vocabularies and ontologies that meet the above criteria wherever possible. Where no existing appropriate vocabularies are available, the portal should ideally work with its community to develop, maintain and disseminate these.

- **Qualified cross-linking with data sources & other resources**: cross-linking with data sources has already been discussed but in addition, the data portal should make use of all appropriate public community resources for unambiguous identification and interlinking of concepts. Examples of this include ORCiDs[15] for authors/data owners and accessions from external databases. Links should be semantically qualified where possible to facilitate machine-actionability of metadata records.
- **Machine-actionable licensing & date use conditions**: licensing and data use conditions should be captured in machine-actionable formats within metadata records, e.g. using the Creative Commons Rights Expression Language (CC REL)[16], Data Use Ontology (DUO)[17] or the Open Digital Rights Language (ODRL)[18].

While a lot of the characteristics listed here can be found in many existing data portals, it is rare that they are all implemented to their full extent. In particular, machine-actionable licensing and data use conditions and formal obsolescence policies are commonly neglected. Truly interoperable semantically enabled metadata is another area where existing resources often fall short, with many resources implementing their own standards rather than reusing or extending existing ones.

3.3. What makes a data portal FAIR?

As well as supporting the FAIRness of its content, a data portal should conform to all requirements of FAIR itself. In other words, the data portal should be FAIR (to a certain degree) itself as well as provide FAIR-supporting features to its content. This means that on top of all the aspects mentioned previously, the portal should also provide human- and machine-actionable metadata of itself, such as how to find and access the portal, its metadata model, its licensing and versioning. There exist a range of metadata solutions to support this, including the aforementioned DCAT, SDO or BioSchemas. In addition, the FAIR health data portal shall endeavour to assess its own level for FAIRness using community standard approach and

[15] https://orcid.org/
[16] https://wiki.creativecommons.org/wiki/CC_REL
[17] https://doi.org/10.1016/j.xgen.2021.100028
[18] https://www.w3.org/TR/odrl-vocab/

automated tooling, such as the "FAIRsFAIR Research Data Object Assessment Service" (F-UJI)[19].

4. FAIR health data portal specification

This section lays out the actual specifications for the FAIR health data portal, taking into account the considerations detailed in the previous section, as well as any previous and concurrent deliverables.

4.1. Metadata catalogue requirements

The main component of the FAIR health data portal is the metadata catalogue of existing health data hubs and data collections. To build this, the minimal information needed to interconnect the existing data sources following the FAIR principles under the same ecosystem has to be defined. Hence, this implies a strong focus on existing data models and use of ontologies and controlled vocabularies as interoperability mechanisms.

This metadata catalogue should include relevant publicly funded health research data hubs, registries and infrastructures, which have been already listed by WP3 and WP4 in their respective deliverables (see table 2 in D3.1 and MS4.1). The meta catalogue specification will cover a range of requirements including:

1. An interoperable metadata model, including a description of available data type-specific metadata templates that the data sources might use.
2. Definition of data access policies and potential standardisation on data usage conditions and data access mechanisms.
3. An engine with ontology-based searching to facilitate easy findability of relevant databases, datasets and registries for the users.
4. A functionality that allows metadata recombination from different sources.
5. Machine actionability.

Together, all of these aspects result in a meta catalogue for health data that will allow users to find the data they need for their projects through a single gateway.

Metadata interoperability

A key challenge of the FAIR health data portal lies in the diverse nature of the data sources indexed in the metadata catalogue. The portal's metadata model needs to both provide a core set of metadata elements to allow harmonisation and integration of the metadata from different sources but also be flexible enough to accommodate new data types without the need for major manual intervention. A

[19] https://doi.org/10.5281/zenodo.4063720

stricter metadata model provides greater interoperability potential but also requires greater effort to align source and target models and put in place "Extract, Transform, Load" (ETL) processes. A more flexible model on the other hand reduces the burden of curation but at the cost of interoperability as there may be less alignment between metadata properties and greater reliance on generic properties without adequate semantic typing.

The FAIR health data portal metadata model needs to balance these two conflicting requirements. HealthyCloud D3.2 will provide guidance on standardising descriptive metadata templates, including guidelines to assess the FAIRness maturity levels. These descriptive metadata catalogue templates will be based on an extension of the DCAT-AP[20] standard for health-related data collections, called Health DCAT-AP extension. This is to also align with the work being done in the EHDS2 pilot project and according to the EU regulation on the EHDS for secondary use.
Ultimately, the FAIR health data portal will likely have to implement a multi-layered metadata approach, with a stricter model at the top level, to maximise the alignment of the greatest number of possible datasets on a limited set of properties, with more flexible sub-models nested underneath where appropriate to capture a higher degree of granularity in some areas. These sub-models will be particularly reliant on high-quality semantic typing of both variables and values to support interoperability.

FAIR data access policies

In order to be truly FAIR, the meta-catalogue needs to explicitly encode licensing and data access conditions in both human- and machine-readable formats. Specifically, the format of these metadata elements should be machine actionable, i.e. machines should not only be able to extract the data elements but also correctly interpret and act on them, e.g. by selecting or deselecting target datasets based on disease restrictions or geographical restrictions. A summary of a dataset's data access modalities should be available without the need for users to review free-text documentation such as data access request forms.
In addition to the actual data use conditions, the portal's FAIR data access policies should also cover the capture and return of decisions made by data access committees, again in both human- and machine-readable formats. This information should be easily reviewable in the portal at users' convenience. This and other data access considerations will be covered in detail in deliverable D6.3, a summary of which is provided below.

[20] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/211

Use case-driven searching

An engine with ontology-based searching is required to facilitate easy findability of relevant data hubs, data collections and registries for the users. In order to increase the usability and decrease the learning curve, the search language should be as expressive as possible.

The portal should support a wide range of users with different skills and technical knowledge levels, from citizens to experienced researchers. For this reason, the portal should provide search at different levels adapted to the different user profiles:

- **Generic**: The goal of this kind of search is providing general vision of the available data in a concrete domain. The user should be able to use generic fields (e.g. data types, diseases) combined with some generic demographics (e.g. sex, age) and geographic (e.g. country) variables. The results of this search should focus on the metadata associated with the data collections instead of the concrete set of data collections. The aim of this kind of search is not looking for data collections.
- **Basic data search**: The goal of this kind of search is exposing what data is available in a concrete domain and understanding its codification. Researchers need to know where to find the data they need and how this data is codified in order to be able to set up the concrete query they need to use. The user should be able to search data that contains a concrete set of variables or categories, i.e., a group of related variables (e.g. intermediate nodes in an ontology). The results of this search need to include the information about where the data is (data collection and data hub) complemented with the metadata related to the fields to understand how the data is codified (e.g. variable name, value type).
- **Advanced data search**: The goal of this kind of search is to identify the data collection that contains individuals with concrete characteristics, i.e., individuals with concrete values for concrete variables. This search can be used to create cohorts (related to the next section 4.1.4). Like in the basic data search, the results of this search need to include the information about where the data is (data collection and data hub) complemented with the metadata related to the fields to understand how the data is codified (e.g., variable name, value type). This kind of search needs some degree of harmonization in order to search in different data collections.

Ideally, the resulting information will be complemented with some demographics and geographic information for each data collection (e.g. number of data collection entries, totals by sex and/or sex, data hub country). This complementary information will support the user to understand the amount of available data and if it can be used because of geographic restrictions.

Metadata-based data recombination

A key functionality of the FAIR health data portal identified in D6.1 is the ability to utilise the portal metadata to create tailored recombination of datasets or elements of datasets based on the information available from the metadata alone, which again highlights the importance of a semantically interoperable metadata model. A representative example of such a recombination would be a virtual or "synthetic" cohort builder. This type of search portal allows the user to identify all data available through the portal from patients that conform to specific criteria of age, sex, disease- or medication status and for whom data types of interest such as sequencing data, vital signs or medication histories are available. Once such a synthetic cohort has been put together, the portal should then mediate data access to the component datasets and sub-datasets in a centralised and unified fashion.
While a synthetic cohort builder represents the most obvious example of data recombination, the overall functionality is generalisable to any other use case of this type, such as the combination of datasets based on experimental or clinical methodologies to study outcome differences for different patient profiles, or to study healthcare trends through metadata aggregation.

Machine actionability

Machine actionability is a core tenet of FAIR but it is often poorly understood in practice or confused with machine readability. The latter is obviously a pre-requisite of machine actionability but while there are many data formats that are machine readable, the data they contain may not be machine actionable. In order be machine actionable, data need to be supplied with metadata that is presented in a way that computers can understand without human input. In the case of data usage conditions for example, this might mean encoding the statement "This dataset may be reused only for non-commercial research on inflammatory bowel disease within the EU" in a way that allow a computer to identify the component conditions of "non-commercial research", "inflammatory bowel disease" and "EU only". Existing solutions such as DUO or ODRL address this particular scenario in a FAIR-compliant way and could be absorbed wholesale into the FAIR health data portal.

4.2. Metadata contribution

The primary source of metadata for the meta catalogue will be the data registries and repositories. Most of this metadata will be added to the FAIR health data portal automatically or semi-automatically, i.e. with minimal human involvement, but the portal also needs to have the capacity for one-off submissions of metadata from trusted sources such as project repositories wishing to share their metadata in a single batch at the end of the project. As the FAIR health data portal is not intended as a primary source of metadata or data but rather as a gateway to external repositories, direct submission facilities for these one-off submissions can be very lightweight. The portal should however be able to act as a gateway for direct

contributions of data or metadata to the appropriate repositories, for example by signposting potential submitters to the best place for their data.

Automatic metadata acquisition

The primary usage scenario for the operation of the FAIR health data portal is that the sources of the resources will make the metadata of these resources available in such a way that the portal can automatically retrieve and index them. In this use case, the metadata is controlled and updated outside the platform by whoever is responsible for the resource and the FAIR health data portal only retrieves and indexes the metadata. It is also expected that the portal keeps the indexed metadata synchronized with the source to guarantee that the users of the portal will have the most updated information.

In order to enable this level of automation, the following agreements have to be made:

- **Metadata access mechanism**: the sources should make their metadata available in a way that the portal is able to access them. The metadata could for example be provided as a web resource, accessible using the HTTP protocol that resolves to a document containing the metadata record. This is the most desirable option although others could be envisaged.
- **Metadata harmonisation**: as discussed in section 4.1.1., the portal needs to feature a metadata model that is sufficiently flexible to accommodate a wide range of data sources as well as being FAIR compliant. Before metadata from a source repository can be integrated into the data portal, the source metadata needs to be harmonised against the portal model. Assuming that both the source and target models are relatively stable, this is a one-off process, albeit a potentially labour-intensive one. If the source model is semantically enabled, its semantics can be leveraged to automate mapping to the portal model. In the absence of a semantic model, the mapping will be the responsibility of human curators with expert knowledge and a good understanding of both models. If a resource already implements DCAT or an extension of DCAT, this will greatly facilitate the metadata harmonisation process, as the catalogue metadata will likely implement the Health DCAT-AP extension. Additionally, the portal may maintain a set of minimal metadata schemas for a number of different types of resources that should be used by the source to improve interoperability.
- **Metadata conversion**: Once the metadata models have been aligned, ETL procedures can be set up to generate portal-compliant metadata. Depending on the quality of the mappings and the complexity of the source metadata, additional manual curation steps or spot-checks may be required on top of any automatic ETL processes. The responsibility for these processes should fall primarily on the source repositories in order to reduce the scalability load on the portal. Not only are the source repositories the

experts about their data and metadata, they also only need to handle what they provide while the FAIR health data portal having to do this for every resource it interacts with would make this work difficult to sustain.

- **Update responsibility**: While the responsibility for updates could lie with the portal, which would require a polling service that monitors all source repositories, this scenario would be difficult to scale up in the case of a large number of source repositories and would place an undue burden on the portal. The preferable scenario would be that the portal is notified by the source of updates in its metadata that the portal needs to synchronize. In this case, the portal will define a specific API endpoint for this notification. To avoid an overload on the portal and the source in the case of constant updates, the portal will schedule a batch synchronization of the source when it receives a large number of notifications from the same source in a short period.

Additional considerations that need to be taken into account in this context include:

- **Versioning strategies:** One important consideration when presenting metadata or data from different sources is the versioning of metadata. Although not explicitly mentioned in the FAIR principles, versioning is an integral part of provenance metadata, highlighting that changes may have occurred in the information that is presented and signposting when changes occurred and what they entail. As updates are generally expected to be batched, these batches can be versioned as a whole, for example per resource, data type or dataset.
- **(Meta)data quality assurance:** The quality of the metadata presented in the portal is intrinsically linked to the quality of the metadata provided by the original sources, mitigated by the harmonisation of the metadata models through carefully designed ETL processes. If the incoming metadata is of poor quality, for example, due to its sparseness, even the most well-designed portal model cannot really overcome these shortcomings. Well-designed ETL process may however be able to improve metadata by converting it to a semantically interoperable metadata model or by including automatic ontology annotations. At a minimum, the portal needs to have in place, as part of its metadata import processes, a solid validation process that flags and, ideally, rejects, metadata that is not compliant with the portal's model. In addition, the onboarding of new source repositories should include an evaluation of the source's data and metadata quality in terms of FAIR compliance, using any of the numerous FAIR assessment methodologies available, such as the previously mentioned one developed by HealthyCloud. The results of these assessments should be encoded in the resource's provenance metadata in order to enable users to trace and judge quality for themselves in an understandable and transparent manner. While the portal ultimately has no control over the data and metadata in the

source repositories, it can provide recommendations on how to improve these data, going as far as to refuse inclusion if minimum standards are not met by the sources.

Data contribution gateway

Although not the primary focus of these FAIR health data portal specifications, the user stories described in D6.1 highlighted the need for the portal to deal with some degree of direct submissions.

The portal's remit very clearly covers only metadata, not the storage of data as a primary resource. Instead, the portal should act as a gateway to repositories for potential data submitters. This should include signposting to help submitters identify the best repository for their data. As with other aspects of the portal, the signposting should be both human-readable in the shape of guidelines or search results, and machine-readable through metadata about the repositories and hubs whose metadata is indexed in the portal. In addition to signposting, the portal could also leverage the repositories' own authentication procedures or a community standard authentication process such as Life Science Login (LS-Login)[21] to provide transitive authentication, allowing users to reuse existing authentication processes in their data submission to repositories via the portal.

Conversely, trusted users of the portal should be able to perform one-off metadata submission directly to the portal, for example in the case of the aforementioned project repository wishing to expose their metadata at the end of the project. In this scenario, automated metadata import processes would represent an unnecessary effort. These types of submissions should however be limited to trusted users only in order to ensure that the source of the metadata is known and traceable, and to avoid the inclusion of substandard metadata from users unfamiliar with the models and standards used in the portal.

4.3. Data access

While facilitating and brokering access to the data for which it presents metadata is a core functionality of the FAIR health data portal, requirements related to data access are actually the remit of the separate deliverable D6.3 "Specifications for data access". This section summarises the work of the closely related milestone MS6.3 "Study: existing mechanisms for usage and access of already structured and organized datasets", as well as the preliminary findings for D6.3.

Data access modalities and requirements vary widely based on a number of factors including but not limited to the organisational structure of the data hosting

[21] https://lifescience-ri.eu/ls-login/

institutions, types of users, metadata and data types. For this reason, the recommendations and procedures for accessing the data have to be flexible and able to adapt to the different situations as much as possible, although findability and discoverability must be possible regardless of whether systems operate in a centralised or federated model. Data ownership will generally remain with the source repository, with only metadata to aid discovery and traceability exposed at the level of metadata catalogues.

Given the sensitive nature of most health data, access to the data - whether by humans directly or via machines acting on behalf of human users - needs to be subject to strictly controlled authentication and authorisation procedures. While aggregated metadata at the level of the metadata catalogue is likely public and openly available, access to the more fine-grained data in the source repositories not only requires users to authenticate themselves but also receive authorisation to access data. This process can be managed through authentication and authorization infrastructures (AAI), such as the one that is being developed in the framework of the EOSC-Life project, the Life Science Login[22,23], that will provide a common AAI for different research infrastructures.

On the side of the metadata catalogue, the primary requirements relating to data access are for human- and machine-readable encodings of any access and reuse conditions for each dataset. The health data portal's systems should enable the user to instigate a data access request from the portal to the source repositories and keep track of all their access authorisations gained via the portal in a central dashboard within the portal.

4.4. Infrastructure providers for computational resources

While the portal itself is not intended to provide computational resources for data-centric health research and analysis, it can serve as a discovery and entry point to existing infrastructures, an overview of which can be found in HealthyCloud deliverables D5.1 and D5.2. Providing this information will again require the collection of relevant metadata from target resources. Unlike metadata about data sources and the datasets they contain however, metadata about computational resources does not necessitate the set-up of automated indexing and update pipelines. Rather, the resources themselves should be able to submit the metadata to the portal via a web form or in structured format via a simple API endpoint.

[22] https://zenodo.org/record/4559400#.Youf-ahByzV
[23] https://zenodo.org/record/4633191#.Youf_6hByzV

4.5. Guidance & knowledge hub

Although not strictly related to the technical remit of FAIRness and FAIR data, the data portal should also act as knowledge hub providing guidance on best practice in data management for health data research. This can include documentation, guidance on tooling, FAIR-related guidance, information on metadata standards and how to apply them, and pointers to training materials. All guidance should be underpinned by FAIR-compliant metadata to ensure that the information is fully searchable in the same way as data-related metadata. As with computational resources, this metadata should be easy to submit by contributors so that the data portal can connect to as many external sources as possible.

5. Conclusion

This report presents some of the key features that need to be taken into consideration in the implementation of a FAIR-compliant health data portal. In particular, the requirements presented here address the key categories identified by the preceding deliverable D6.1 on user personae, namely

- Share data in a secure environment (sections 4.1, 4.2 & 4.3).
- Find individual and aggregated data from different domain and sources (section 4.1).
- Find infrastructure providers for data management and analysis (section 4.4).
- Find guidance and best practices in different aspects (section 4.5).
- Access data quality validation mechanisms (section 4.1 & 4.5).

FAIR-compliance relies primarily on high-quality metadata that, in a standard interoperable format such as DCAT, presents not only the portal content but also relevant information about the portal itself and its access and navigation modalities, to both human users and machines. We discuss how this metadata is used in the metadata catalogue, how it is acquired and how data access can be brokered by the portal. While this specification is not exhaustive in terms of the full features for a data portal, it highlights the specific requirements for increased FAIRness in the health data space.

In line with the HRIC strategic agenda, the FAIR health data portal described here will be a central community resource that connects together and leads the health data space not just in terms of its data but also in terms of knowledge, guidance and research best practice.

Acronyms and Abbreviations

- AAI - Authentication and Authorization Infrastructure
- API - Application Programming Interface
- D - Deliverable
- DCAT - Data CATalog vocabulary
- DCAT-AP - DCAT Application Profile
- DTA - Data Transfer Agreement
- DUO - Data Use Ontology
- EHDS - European Health Data Spaces
- EJP RD - European Joint Programming for Rare Diseases
- ERN - European Reference Network
- ETL - Extract, Transform, Load
- EUCAIM - Federated Cancer Imaging Infrastructure
- FDP - FAIR Data Point
- GDI - Genomics Data Infrastructure
- HRIC – Health Research & Innovation Cloud
- MS - Milestones
- ODRL - Open Digital Rights Language
- RI- Research Infrastructure
- SDO - Schema Dot Org
- WP – Work Package