HEALTHYCLOUD
Health Research & Innovation Cloud

D3.3 – Landscape analysis using a health-related data catalogue matrix

Document Information

Contract Number	965345
Project Website	http://www.healthycloud.eu/
Contractual Deadline	M27, May 2023
Dissemination Level	PU - Public
Nature	R - Report
Author(s)	Shona Cosgrove (Sciensano), Irene Kesisoglou (Sciensano), Pascal Derycke (Sciensano)
Contributor(s)	Celia Alvarez Romero (SAS), Lorenz Lorenz Dolanski-Aghamanoukjan (GöG), Mari Mäkinen (THL)
Reviewer(s)	Maria Panagiotopoulou (ECRIN), Eva García Alvarez (BBMRI)
Keywords	FAIR principles, data infrastructures, health research, metadata records, metadata catalogue, landscape analysis, European Health Information Portal

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

Change Log

Version	Author	Date	Description of Change
V0.1	Shona Cosgrove Irene Kesisoglou Pascal Derycke	01/04/2023	Initial Draft
V0.1	Shona Cosgrove Irene Kesisoglou Pascal Derycke	03/05/2023	Draft version sent to WP3 participants
V0.2	Shona Cosgrove Irene Kesisoglou Celia Alvarez Romero	10/05/2023	Draft updated based on WP3 participants' input
V1.0	Shona Cosgrove Irene Kesisoglou	30/05/2023	Draft updated based on reviewers' comments (Maria Panagiotopoulou)
V1.2	Irene Kesisoglou	07/06/2023	Final version submitted to EC

Table of contents

Contenido

1. Executive summary

This document presents the findings of the extended landscape analysis performed in HealthyCloud WP3. The aim of HealthyCloud WP3 is to carry out a landscape analysis of available health-related data infrastructures, in order to capture the European health data collections available for research purposes, evaluate their FAIRness levels and determine the feasibility to perform individual level data linkages.

To perform this landscape analysis, Task 3.1 focused on collecting information on available health-related data infrastructures, including their governance, health-related domains covered, structure of the data stored, quality assurance of the datasets and the adherence to the FAIR principles (Findability, Accessibility, Interoperability and Re-usability), among others.

Initially, to collect this information, a survey was designed in collaboration with the leaders of WP4, in the form of a catalogue matrix. The previous deliverable D3.1 'Landscape analysis of FAIRness levels of health-related data using a catalogue matrix' presented the initial landscape analysis performed with the catalogue matrix, which focused specifically on the scope of the HealthyCloud use cases on atrial fibrillation and cancer.

The deliverable D3.3 'Landscape analysis using a health related-data catalogue matrix' presents the subsequent extension of that landscape analysis. The extension was achieved through a collaboration between the WP3 team and the Population Health Information Research Infrastructure (PHIRI) project, which has developed the European Health Information Portal (HIP), a one stop shop for services for researchers, including a metadata catalogue of health data collections.

In terms of methodology, the HIP metadata template of data sources was compared to the catalogue matrix developed by HealthyCloud, allowing the identification of properties common to both as well as essential properties missing from the HIP metadata template, which were proposed to be added. Subsequently, the HIP metadata template was sent to HealthyCloud partners so that they could add a record of their data collection. The landscape analysis was performed by analysing all new and existing data source records in the HIP (over 330). The key properties to be analysed were identified to be in line with the previous methodology in D3.1. Finally, a FAIRness evaluation was carried out of the new records made by HealthyCloud partners, using the FAIRness evaluation tool developed in D3.1.

The results section of this deliverable D3.3 shows the analysis of the following properties: type of information, geographical coverage, target population, access information, updating periodicity, personal identifier, level of aggregation, linkage possibility, permanent identifier of the data source. Over 330 data collections were analysed. In addition, six additional properties were analysed for the data collections added by HealthyCloud partners: type of data, anonymisation, community standards used to structure data, data format for exchange, metadata record, unique identifier

for metadata. These properties were added as they are important properties to determine the FAIRness and linkability of data. The results of the FAIRness evaluation of the new records using the HealthyCloud FAIRness assessment tool are also presented.

Overall, the landscape analysis demonstrates that health-related data collections in Europe are highly heterogeneous. Most of the data collections currently included in the HIP are national or regional, but the collaboration with HealthyCloud allowed the addition of European level data infrastructures. This collaboration benefited both projects, allowing the completion of the HealthyCloud landscape analysis covering over 330 data collections, but also allowing the enrichment and future improvement of the HIP's European health-related metadata catalogue.

2. Introduction

The recent COVID-19 pandemic, amongst others, brought to light two key aspects about health data in Europe; the complex and very heterogeneous landscape of health-related data collections and the lack of adherence of these collections to the FAIR principles, compromising their re-use by researchers and policy makers. Therefore, the HealthyCloud project's Work Package 3 was tasked to provide an overview of already available health-related data collections and infrastructures in Europe, and collect more information on their structure, the data management and re-use of these data collections.

Within this work, Task 3.1 focused on performing a landscape analysis of available health-related data infrastructures, collecting information about the data governance, health-related domains covered, structure of the data stored, quality assurance of the datasets and the adherence to the FAIR principles (Findability, Accessibility, Interoperability and Re-usability) [1]. To collect this information, a survey was designed in an electronic tool (using typeform.com) and was conducted in collaboration with the leaders of WP4 [2].

As an initial step, the study focused on the health data collections that would be useful to answer the research questions of the two HealthyCloud use cases - one on cancer and one on atrial fibrillation. The aim of this initial focus was to analyse the feasibility of linking individual level data from these different data collections in order to perform the two research questions of the use cases. The results of this initial study were presented in Deliverable 3.1 'Landscape analysis of FAIRness levels of health-related data using catalogue matrix' [3].

Thereafter, the HealthyCloud landscape analysis of health-related data infrastructures has been expanded through a collaboration with the Population Health Information Research Infrastructure (PHIRI) project. PHIRI has developed the European Health Information Portal (hereafter referred to as HIP), an online one stop shop for services, such as a metadata catalogue of more than 330 available health-related data collections across Europe. This collaboration between HealthyCloud and PHIRI benefited both initiatives. It allowed, on the one hand, enrichment of the HIP through the addition of data sources from the HealthyCloud consortium partners and, on the other hand, the HealthyCloud landscape analysis was expanded through the analysis of the 330 already existing records on the HIP.

This document, D3.3, presents the final landscape analysis of health-related data infrastructures performed based on this collaboration. It builds on previous work presented in D3.1 'Landscape analysis of FAIRness levels using a health related-data catalogue matrix', and expands the analysis and FAIRness evaluation to more data collections.

3. Methods

3.1. Survey development and dissemination (D3.1)

For the initial landscape analysis, presented in D3.1, a survey was developed through a collaboration between WP3 and WP4, aiming to combine efforts and avoid sending multiple similar surveys to the same health data infrastructures.

To develop the survey, the following aspects were considered:

- The organisation and governance of the data infrastructures;
- The nature of the data;
- The type of data sources and level of detail;
- The data storage process;
- The findability, accessibility, interoperability and re-usability of the data and metadata. The compliance with the FAIR principles, as defined by the Research Data Alliance (RDA).

The survey (also referred to as catalogue matrix) included over 50 indicators (questions) under the following ten areas:

1. Administrative;
2. Data;
3. Completeness of the data collection;
4. Quality aspects of the data collection;
5. Metadata;
6. Findability;
7. Accessibility;
8. Interoperability;
9. Re-usability;
10. Data governance.

The survey was made available in an online tool (see here) for ease of completion by respondents [2]. The full survey was included in the Annex of D3.1 [3].

As mentioned above, the scope of the first landscape analysis presented in D3.1 was set around the two use cases of HealthyCloud, on cancer and on atrial fibrillation. In collaboration with the task leaders of WP7 responsible for these use cases, the relevant data collections containing the various data essential to conduct the studies of the use cases and answer the research questions were identified. The survey was sent to more than 28 data collections in the scope of D3.1.

The data collections covered in the initial landscape analysis are shown in Table 1.

Table 1: Data collections to which the survey was sent for D3.1.

Cancer use case	
Belgium	Belgian Cancer Registry
Belgium	Belgian Registry on Genomic Data
Belgium	Health Interview Survey and Health Examination Survey
Belgium	Statbel
Finland	Avohilmo, Register of Primary Care Visits
Finland	Care Register for Social Welfare
Finland	Findata - Social and Health Data Permit Authority
Finland	FinHealth 2017 Survey
Finland	Finnish Cancer Registry
Finland	Finnish Social Science Data Archive
Finland	National FinSote Survey
Finland	Research Services at Statistics Finland
Finland	THL Biobank
Germany	Survey was sent to German contacts in Charite and TMF for dissemination to relevant data infrastructures
Spain	Cancer Registry of Granada
Spain	Genomics registry SAS
Spain	Red Española de Registros de Cáncer (REDECAN)
Spain	Registro de Cáncer Poblacional de Castilla y León (RECA)
Atrial fibrillation use case	
European	BigData@Heart

Finnish	Biobank of Eastern Finland
French	Atrial Fibrillation registry
French	MICCAI 2012 Right Ventricle Segmentation Challenge
French	MICCAI 2017 ACDC
Germany	Study of Health in Pomerania
Spain	FANTASIIA Registry
Spain	FAPRES Registry
Spain	REVERSE Registry
European Research Infrastructures relevant to both use cases	
European	BBMRI-ERIC
European	EuroBioImaging

3.2. Expansion of the landscape analysis and collaboration with PHIRI 'European Health Information Portal'

The aim of D3.3 was to expand the landscape analysis of available health-related data infrastructures. The expansion of the landscape analysis included contacting the HealthyCloud consortium partners in order to include data collections from the consortium. HealthyCloud has the advantage of being a project with consortium partners from a wide spectrum of health-related domains, from public health to molecular biology, genomic data and clinical trials. In collaboration with the PHIRI project the same methodology as the one used in D3.1 was used to analyse over 330 available data collections that had already a metadata record in the metadata catalogue of the HIP.

3.2.1. Adaptation of metadata template

The European HIP has a metadata catalogue based on a dedicated metadata record template that follows the DDI metadata standard [4], which makes it interoperable with other metadata catalogues. Information is collected using the DDI metadata record template, interoperability with other metadata catalogues is achieved by using the DCAT-AP 2 standard (maintaining a FAIR Data Point instance). A FAIR Data Point instance is a metadata service that provides access to metadata following the FAIR principles. It uses a REST API for creating, storing and serving FAIR metadata records.

Interoperability with search engines is achieved using schema.org and hence all metadata records included in the HIP metadata catalogue are easily findable through the web with search engines.

In order to collaborate with the PHIRI team, the HealthyCloud WP3 leads had a joint meeting with the PHIRI coordination, in which the HIP metadata template was compared to the survey (catalogue matrix) developed by HealthyCloud WP3 for the initial landscape analysis. Through this comparison, questions/properties that were common to both were identified and essential questions/properties missing from the HIP metadata record template were proposed to be added.

The table below (Table 2) presents the six properties included in the HealthyCloud catalogue matrix that were not in the HIP metadata template and were proposed to be added.

Table 2: Properties proposed to be added to the HIP metadata template

Indicator	**Description of the indicator**	**Format of the input**
Type of data	Specify the type of data collected	Multiple choices possible: / Images / Text / Numbers / Files / Tissue samples / Sounds / Other (please specify)
	If other, please specify	Free text
Anonymisation	Is the data stored within the data infrastructure anonymised or identifiable?	Select a single response: / Anonymised / Identifiable / I don't know / This question doesn't apply to my data infrastructure

Unique identifier for metadata	Do you have a unique identifier for your metadata (e.g., uuid)?	Select a single response? / Yes / No / I don't know / This question doesn't apply to my data infrastructure
	If yes, what type of unique identifier (e.g., uuid)?	Free text
Standards used for data	Which community-recognised vocabularies, standards or methodologies are used for data to facilitate interoperability?	Multiple choices possible: / SNOMED CT / LOINC / ICD-10 / ICD-11 / Other / I don't know / This question doesn't apply to my data infrastructure
	If other, please specify	Free text

Data format for exchange	What is the format(s) for distributing data?	Multiple choices possible: / csv / xml / json / ld-json / pdf / R / SAS / Other / I don't know / This doesn't apply to my data infrastructure
	If other, please specify	Free text
Metadata record	Do you have a metadata record API endpoint (m2m) in place?	Select a single response: / Yes / No

These six additional questions were sent to the additional data collections in an Excel form in order for the responses to be analysed for this deliverable. In the future these six additional properties may be added to the metadata record template of the HIP and will hence support its enrichment.

3.2.2. Dissemination of metadata form

The HIP metadata record template form was then disseminated to HealthyCloud partners. From the total of 21 HealthyCloud partners, a selection was made for who to send the form to. The exclusion criteria were:

- No data (e.g., partners relevant to the computational work of HealthyCloud who are not data holders of health-related data): 5 partners excluded: BSC, de.NBI, CSC, EGI, TMF.
- Already included in the HIP (e.g., partners involved in PHIRI): 2 partners excluded: THL, Sciensano.

The metadata template was then sent to 13 partners, including instructions on how to complete the form, as well as the six additional fields in an excel form. Reminder follow up emails were sent over a period of 6 weeks. Excluding those from whom no

response was received, those whose organisation is not a data holder and those who were already included in the HIP, finally eight new records were added to the HIP.

The new records created were:

- BBMRI-ERIC Directory
- BIGAN
- CRC-COHORT from BBMRI
- Electronic Health Record of the Andalusian Public Healthcare System
- Erasmus Glioma Database
- European Genome-phenome Archive
- Study of Health in Pomerania
- UK Biobank

3.2.3. Analysis of new and already existing records

At the time of the analysis for this deliverable, the HIP had 332 records (including both previously existing and new records added through HealthyCloud).

In order to ensure the extended landscape analysis presented in this deliverable is consistent with the initial landscape analysis in D3.1, the properties selected for analysis were selected to be in line with those previously analysed in D3.1.

The properties selected for analysis were the following:

1. Type of information
2. Geographical coverage
3. Target population
4. Access information
5. Updating periodicity
6. Personal identifier
7. Level of aggregation
8. Linkage possibility
9. Permanent identifier of the data source

The six additional fields (not yet part of the HIP metadata form) were analysed separately only for the eight new records as this information was not available for the already existing metadata records of the HIP.

3.3 FAIRness evaluation of new entries in the HIP (using HealthyCloud FAIRness evaluation tool)

In addition, the new records made by HealthyCloud partners underwent a FAIRness evaluation using the FAIRness evaluation tool developed in HealthyCloud WP3. The methodology for the development of the FAIRness assessment tool has been previously described in D3.1 and we briefly describe it below.

After in-depth examination of available web-based tools endorsing the FAIR Data Maturity Model, the ARDC FAIR Data self-assessment tool published by the Australian Research Data Commons (ARDC) [5] was selected as a basis.

The ARDC FAIR Data self-assessment tool consists of a HTML Web page with functionalities coded in Javascript. The HealthyCloud WP3 team customised and integrated the existing tool in an Rmarkdown notebook and extended its functionalities. The developed **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 compliance with each FAIR principle as well as an overall score.

The HealthyCloud FAIRness self-assessment tool has been made freely accessible on a public BinderHub portal hosted by the community at mybinder.org allowing any user to produce the FAIRness evaluation and the general analysis of their data collections [6]. The FAIRness evaluation reports produced by the tool can be updated at any time as a csv file, which can be downloaded and will serve to produce a new updated report from the tool.

The HealthyCloud FAIRness self-assessment tool has been published on ZENODO [7]. The tool was developed in this user-friendly format as it offers the means to expand the landscape analysis to more data collections, and facilitates analysis by making it more user friendly. The HealthyCloud FAIRness self-assessment tool includes quick user instructions on how to proceed with the tool. A Readme file is also accessible on GitHub [8].

The FAIRness assessment tool questions and corresponding point allocations are included in Annex 2.

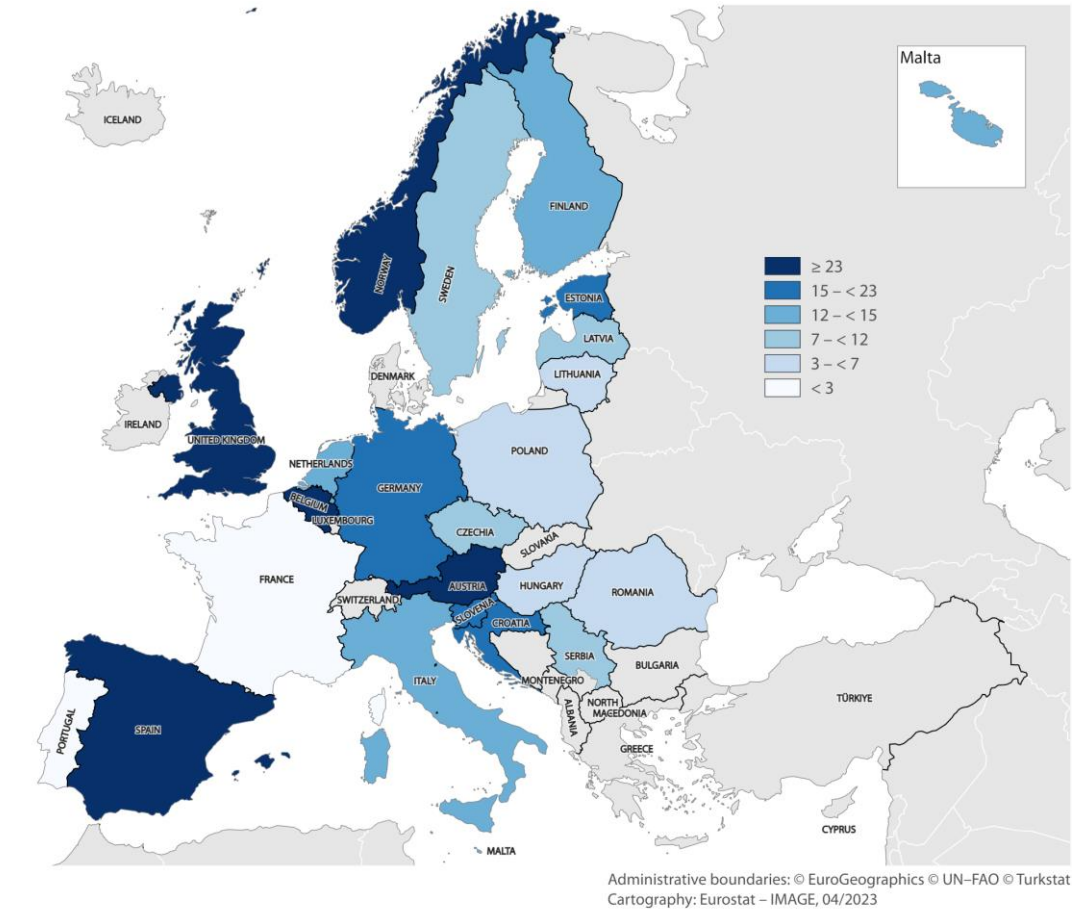
4. Results

4.1. Landscape analysis in a population health-related metadata catalogue

This section presents the results of the analysis of the properties listed above.

On the metadata catalogue of the European HIP it is possible to link metadata records with national nodes (the country where the data source is based). The number of metadata records registered by or linked to each national node was counted, and is demonstrated in the map below (Figure 1). This map of Europe represents in colour-coding the amount of metadata records registered on the metadata catalogue of the European HIP per country. European-level data infrastructures or data collections are not represented on the map as they are not associated with a national node.

Figure 1: Map of Europe showing geographical coverage of data collections in the EU HIP metadata catalogue



A. Geographical coverage of the data within the datasets described in the HIP

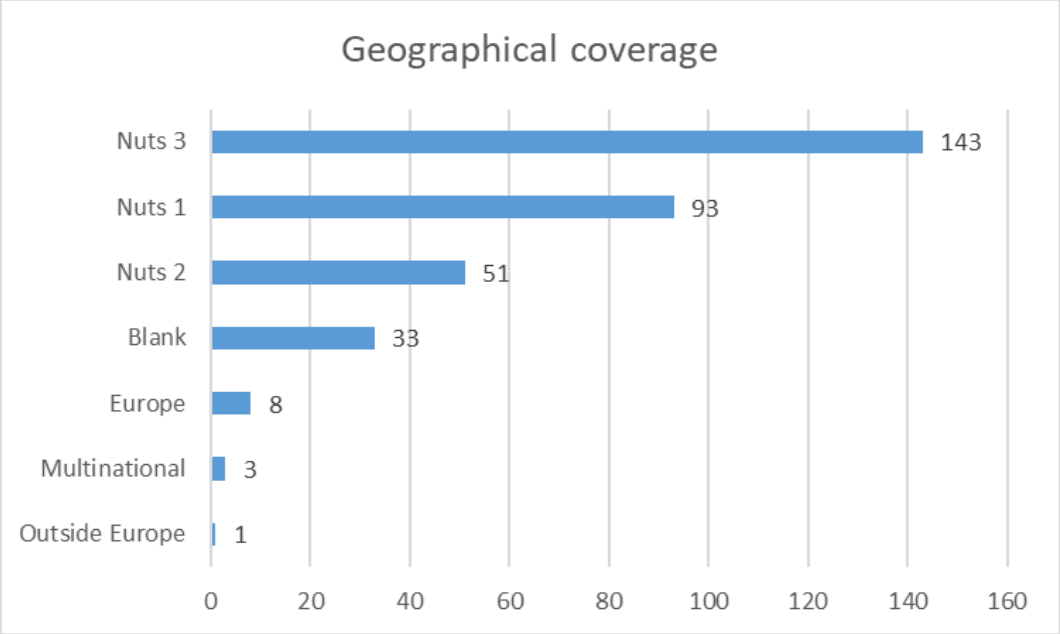
The metadata template of the HIP contains a property on the geographical coverage of the data collection, based on the NUTS classification for national data collections. The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU and the UK for the purpose of collection, development and harmonisation of European regional statistics:

- NUTS 1: major socioeconomic regions
- NUTS 2: basic regions for the application of regional policies
- NUTS 3: small regions for specific diagnoses

The HIP metadata template also includes the option for data collections to note if they have multi-national, European, or outside Europe coverage. There is the possibility to select only one option from: NUTS 1, NUTS 2, NUTS 3, Europe, Multinational, Outside

Europe. The graph below depicts the distribution of geographical coverage of the data collections registered in the HIP.

Figure 2: Geographical coverage of data collections within the HIP



In general, the vast majority of data collections (87%) have national-level coverage (NUTS 1 to 3). The largest proportion (43%) have coverage at the NUTS 3 level. The high proportion of data collections with national coverage and low level of European-level data collections is explained by the fact that the methodology for adding metadata records into the HIP has mainly been through mobilisation of national contacts using the network of National Nodes[1] created within PHIRI and the preceding Joint Action of Health Information (InfAct). Four of the eight European-level records were added by HealthyCloud partners in the development of this deliverable, demonstrating the value of the collaboration to expand the level of coverage of the HIP.

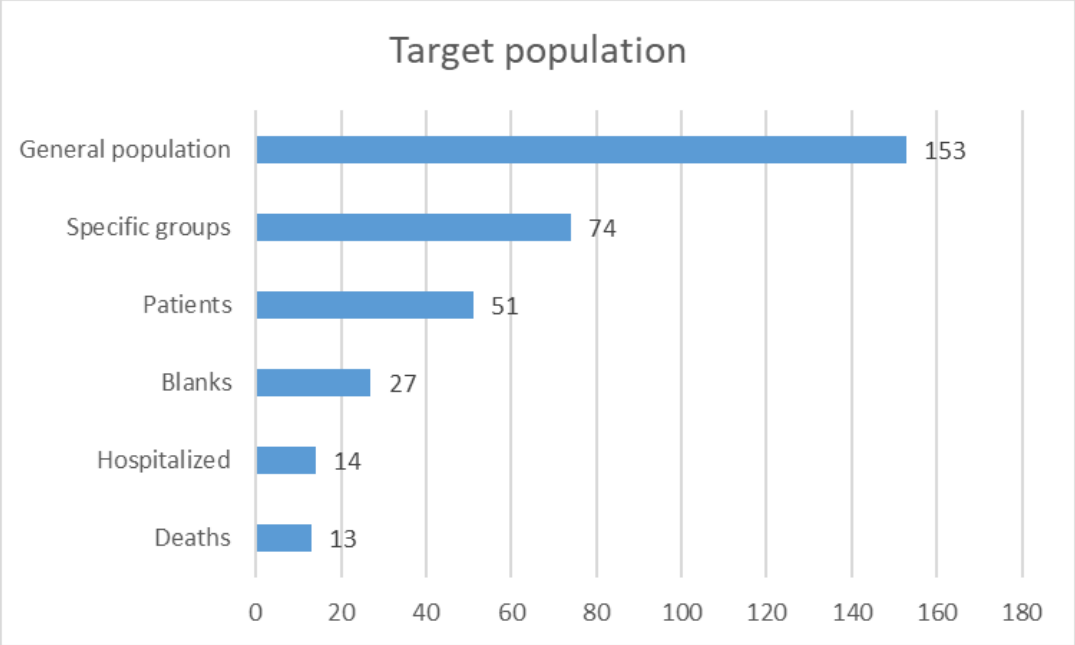
B. Target population

Under 'target population', data collections have to indicate the population group from which their data originates. They are able to select one of the following options:

[1] PHIRI National Node: A National Node (NN) is an organisational entity, often linked to a national institution or governmental unit that functions as a national liaison and brings together relevant national stakeholders in the country in a systematic way. The relevant stakeholders may include, for example, the national statistical office, the national public health institutes, representatives from ministries of health, research and/or science, and others. In addition, the NN may function as a discussion and advisory forum in matters of health data and information both for national or international matters. Examples include aspects of the governance of data, indicators and health reporting at the international level and health information stakeholders at national level.

General population, specific groups, patients, hospitalised, or deaths. The graph below depicts the distribution of target population in metadata records in the HIP.

Figure 3: Target population of data collections recorded in the HIP

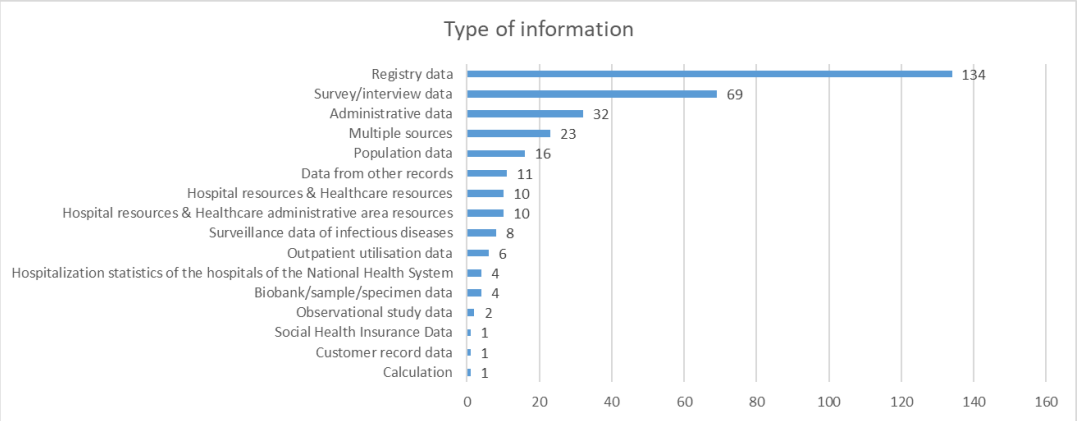


Almost half of records (46%) indicate that the data collection has data from the general population. This finding is consistent with the fact that the HIP is a product of PHIRI, the Population Health Information Research Infrastructure, and the aim is to make population health data sources findable for researchers. Of the eight new records added by HealthyCloud partners, three indicated that the target population was patients, and five indicated that it is the general population, in line with the general pattern noted in the HIP.

C. Type of information

The HIP metadata template includes a property on the type of information included in the data collection. This refers mainly to the type of source from which the data originates. It is important for data users to know what type of information a data collection or infrastructure contains, to be able to know whether the source is adapted to their research question and will be able to provide answers to it.

Figure 4: Type of information



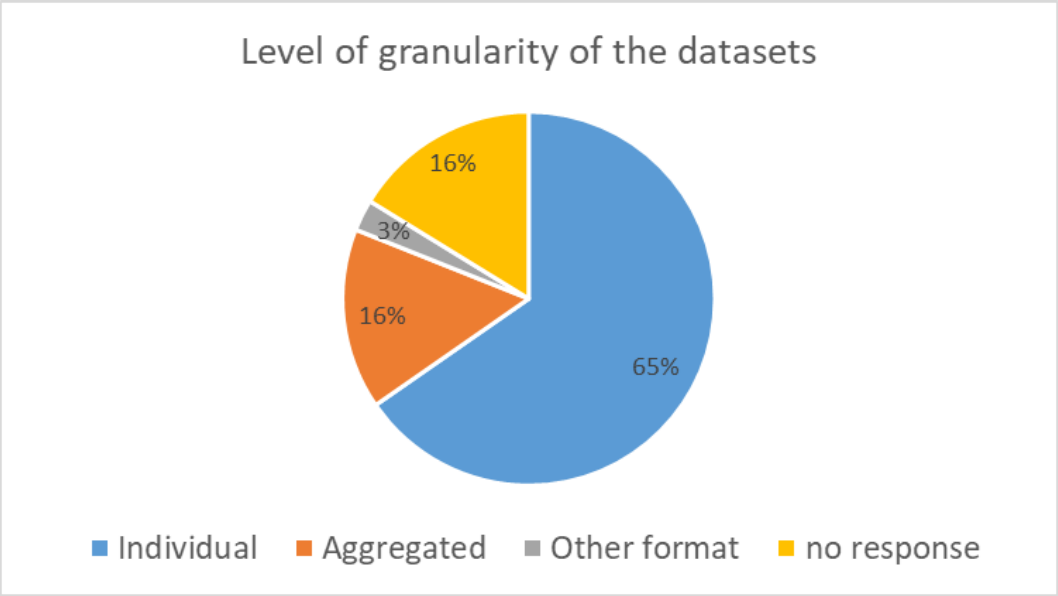
The largest proportion of metadata records indicate that the type of information is from registry data (40%), followed by survey/interview data (21%) and administrative data (10%). Other types of information or data sources have low representation. This distribution is fitting given the HIP's focus on population health information, which commonly uses registries and surveys/interviews as a data source. Of the eight records added by HealthyCloud partners, there were two data collections respectively for each of the following types of information: multiple sources, biobank/sample/specimen data, data from other records, and population data. This different distribution to the general HIP is due to the fact that HealthyCloud partners are from a variety of different fields. It also demonstrates the benefit of this collaboration, which allowed the HIP to be enriched with metadata about different types of data that is useful for public health research.

D. Dataset granularity

In terms of the level of granularity of the data stored in the data infrastructures (i.e., aggregated or individual), 65% have individual level data, 16% have only aggregated level data and 16% did not respond to this question.

This is an important property to know about the data collection or infrastructure as it reflects the possibility of this dataset to be used in research studies requiring linkage with other datasets.

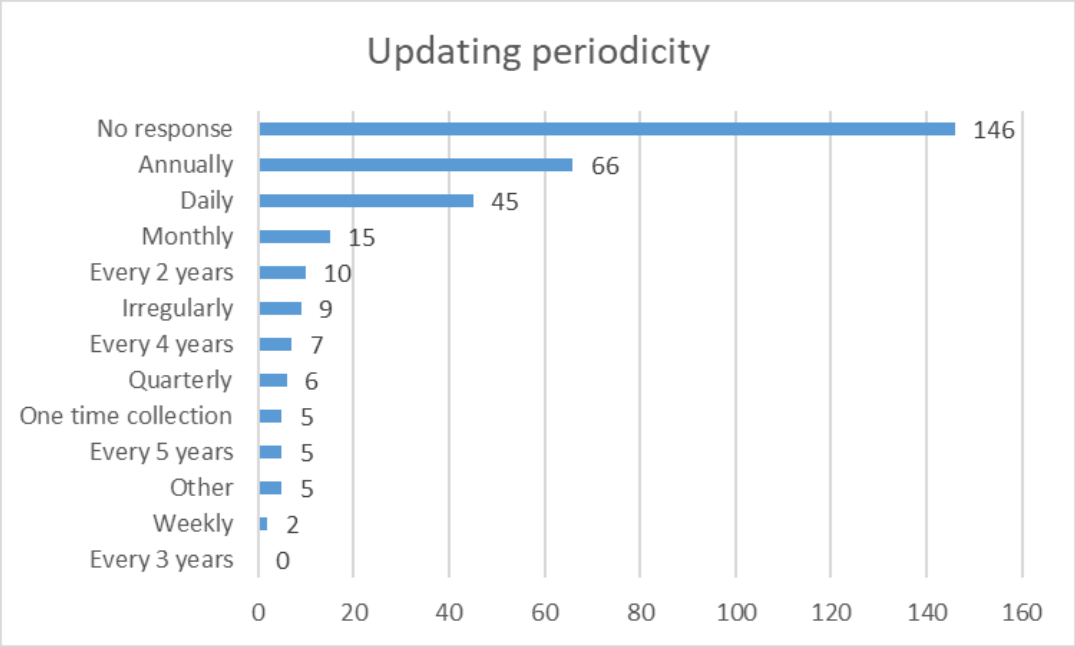
Figure 5: Level of granularity of the datasets recorded in the HIP



E. Updating periodicity

Updating periodicity refers to how often the data contained in the data collection is updated. It is important for data users to know as it indicates how timely the information is. In the HIP metadata template, data collections must select one option. The graph below shows the distribution of updating periodicity indicated in the metadata records in the HIP. It is not mandatory to provide a response for this property.

Figure 6: Updating periodicity indicated in HIP records

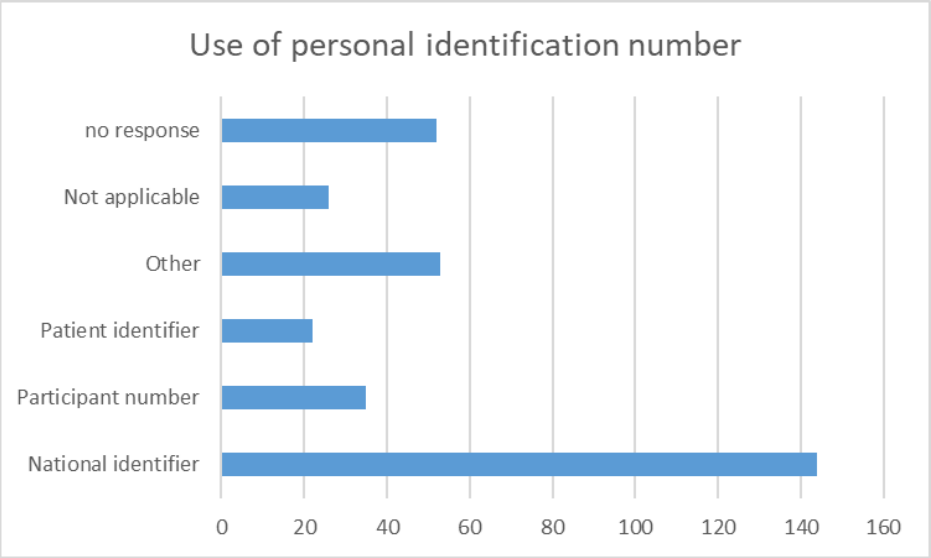


Almost half of records in the HIP do not indicate the updating periodicity (44%). 20% indicate that the data in their data collection is updated on an annual basis. 14% indicate that it is updated daily and less than 1% on a weekly basis.

F. Personal identifier

The use of a unique identifier is a key property to improve linkability. As depicted in the graph below, over 40% of data collections use a national identifier, increasing the possibility to link individual level data with other data collections in the country. Approximately 20% use either a participant number or patient identifier as their personal identifier.

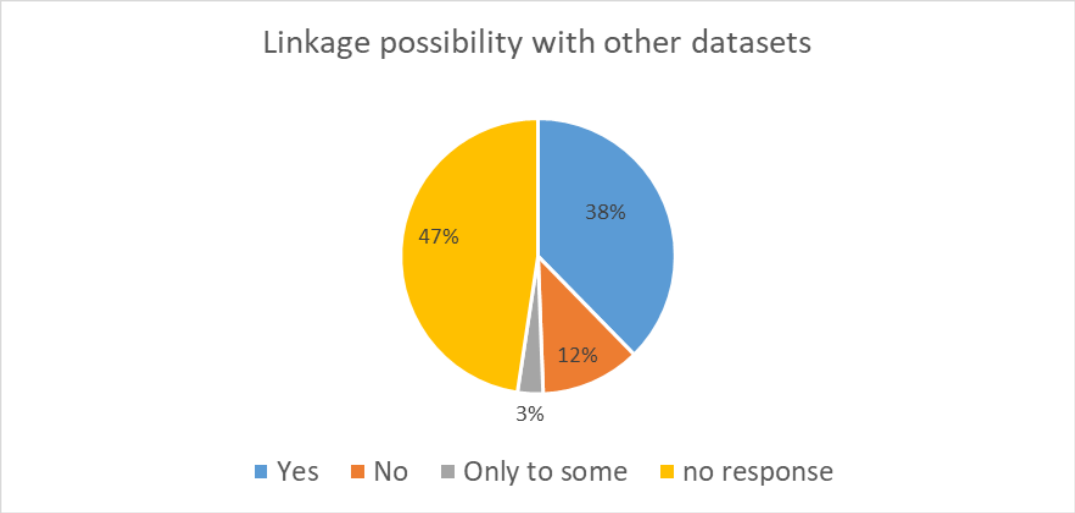
Figure 7: Use of personal identification number



G. Linkage possibility

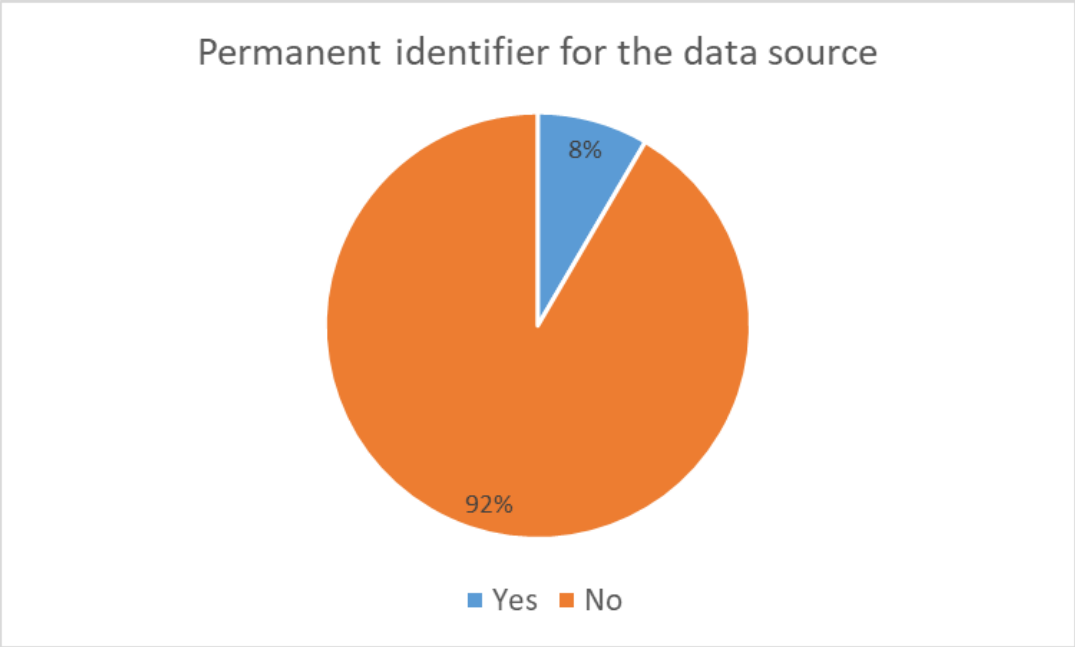
Regarding the possibility to link, only a small proportion (38%) indicate that linkage is possible. This field in the metadata template is not mandatory. Thus, a large proportion (47%) did not provide a response. It can be assumed that those data collections that did not respond do not have the possibility to link with other datasets, or at least not via a simple and well-known process. A small proportion (3%) indicate that linkage is possible, but only to certain other datasets.

Figure 8: Linkage possibility



H. Permanent identifier for the data collection

The HIP metadata template includes a field to indicate the permanent identifier of the data source or collection. Only 8% of records in the HIP indicated a permanent identifier for their data source. A permanent identifier for the data source increases its findability. It is important to note that this question is not obligatory in the HIP metadata template, which likely contributed to the low number of data collections that indicate one. In addition, it is possible that differences in terminology also contributed to the low response rate (e.g., persistent versus permanent identifier).



I. Available access information

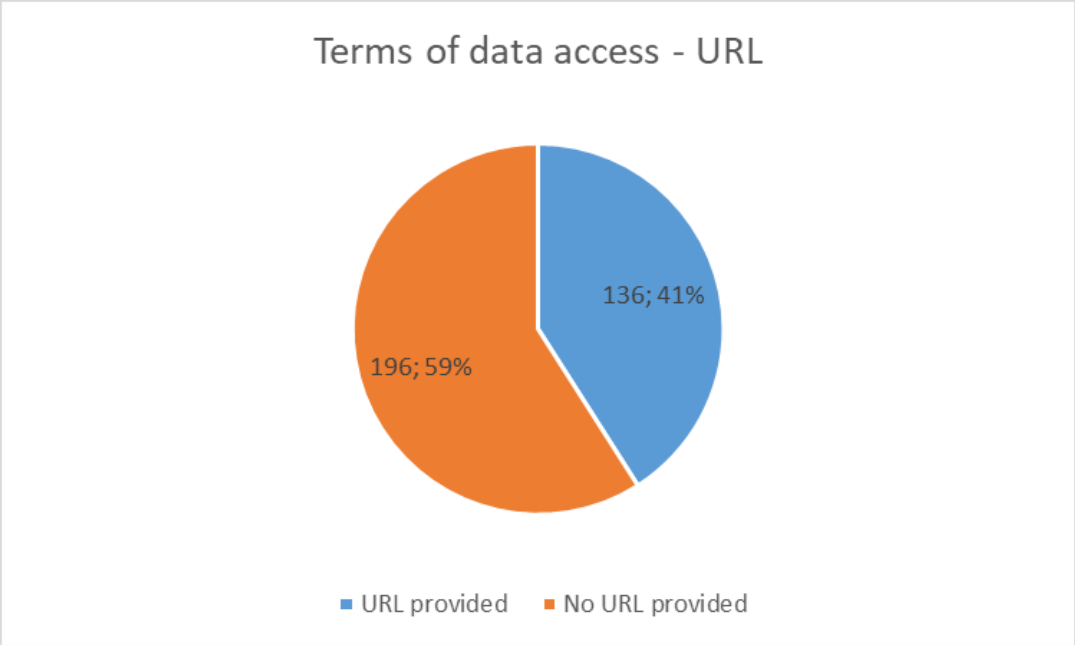
Providing access information is key to having FAIR data, and is one of the key properties for data users accessing a metadata catalogue such as the HIP. The format

of the HIP metadata template means that access information is provided in several different properties:

- Access information (free text)
- Terms of data access (URL)
- Terms of data access (free text)
- Regulations for data sharing (free text)

Given the large number of records for analysis, it was not feasible to perform an analysis of the access information provided by each record in free text format. However, the fact of providing a URL to a web page containing the access information is already a good indicator. The figure below shows the proportion of metadata records that have provided a URL with the terms of data access.

Figure 10: HIP records that have provided a URL with the access information



As depicted in the pie chart above, the majority of records do not have a URL with the access information provided.

4.2. Additional properties

The six additional properties identified as important to add to the HIP metadata template could only be analysed for the eight new data collections added by HealthyCloud partners, as they have not yet been added to the metadata form.

A. Type of data

When searching for data, it is important for researchers to know what type of data is available in a certain data collection, as this defines whether it will be adapted to their research question. The following table demonstrates the results received from HealthyCloud partners regarding the type of data in their data infrastructure.

Table 3: Type of data

Specify the type of data collected	**BBMRI-ERIC Directory**	**BIGAN**	**CRC Cohort**	**EGA**	**Electronic health record of the Andalusian Health System**	**Erasmus Glioma Database**	**SHIP**	**UK Biobank**
Images	X	X	X		X	X	X	X
Text	X	X	X		X		X	X
Numbers	X	X	X		X		X	X
Files	X	X	X		X		X	X
Tissue samples	X		X					
Sounds								
Other (please specify)				X Genetic sequencing and phenotypic information				

Only one of the data infrastructures responded that they have one data type alone: the Erasmus Glioma Database stores only imaging data. All of the other data infrastructures have multiple data types. Five of the data infrastructures (62.5%) have a combination of images, text, numbers and files. The EGA has genetic sequencing and phenotypic information.

B. Anonymisation

The question of whether data is stored in an anonymised or identifiable form is key to determining the usability of data for certain research purposes. If data is anonymised already at the point of storage, this reduces linkability to other datasets.

Table 6: Anonymisation

Is the data stored within the data infrastructure anonymised or identifiable?	**BBMRI-ERIC Directory**	**BIGAN**	**CRC Cohort**	**EGA**	**Electronic health record of the Andalusian Health System**	**Erasmus Glioma Database**	**SHIP**	**UK Biobank**
Anonymised	X					X	X	X
Identifiable		X Pseudonymised	X Pseudonymised	X Pseudonymised	X			
I don't know								
This question does not apply to my data infrastructure								

Of the eight HealthyCloud partners, four (50%) reported that data is stored in their data infrastructure in an anonymised format. Three data infrastructures, BIGAN, CRC Cohort and the European Genome-phenome Archive (EGA) indicated that data is stored in a pseudonymised format. Finally, the Electronic Health Record of the Andalusian Health System stores identifiable data.

C. Standards used for metadata and data

One of the most important factors to link individual level data or datasets across different member states is interoperability. This can be affected by the format in which datasets have been stored in, the semantic interoperability standards used, such as ICD-11 or SNOMED CT, the common data model used to describe them, such as OMOP, or the standard used to transfer data, such as HL7 FHIR.

Data can be structured semantically according to internationally recognised standards, such as SNOMED-CT, LOINC, ICD-10 and ICD-11.

In order to use a dataset to answer a research question, it is important to have prior knowledge on the way this dataset is structured for interoperability reasons, either with the analysis script or if needed to link this dataset with another dataset.

Therefore, this is another one of the properties that we believe would be essential to add in the metadata record template of the European HIP.

Table 4 below shows the responses received when the HealthyCloud partners were asked which community-recognised standards are used in their datasets.

Table 4: Community standards used to structure the data

Which community standard do you use to structure your dataset?	**BBMRI-ERIC Directory**	**BIGAN**	**CRC cohort BBMRI**	**EGA**	**Electronic health record of the Andalusian Health System**	**Erasmus Glioma Database**	**SHIP**	**UK biobank**
SNOMED CT		X			X			
LOINC		X			X			
ICD-10	X	X	X		X		X	X
ICD-11							X	X
Other	X	MIABIS	MIABIS		ATC, ICD-9	None		
I don't know								
This question doesn't apply to my data infrastructure				X				

The table above presents the standards used by the different data collections to structure their data. 75% of the data collections use the same ICD-10 semantic interoperability standard to structure their data. However, in 50% of the metadata records we observe the use of ICD-10 in addition with other standards, such as ICD-11 and SNOMED-CT.

D. Data format for exchange

Formats for health-related data exchange can be: csv, xml, json, ld-json, pdf, R and SAS.

Half of the data infrastructures (50%) distribute data in csv files. Data is also distributed in JSON and XML file formats. This lack of interoperability observed between these data infrastructures might cause a challenge to a research project that aims at linking individual level data between these data collections.

Table 5: Data format for distribution

In what format is your data distributed?	**BBMRI-ERIC Directory**	**BIGAN**	**CRC cohort BBMRI**	**EGA**	**Electronic health record of the Andalusian Health System**	**Erasmus Glioma Database**	**SHIP**	**UK biobank**
csv		X			X		X	X
xml							X	X
json					X			
ld-json								
pdf								
R								
SAS								

Other	X		Data can be provided in different formats depending on the request and the type of data released	X	HL7 FHIR HL7 V2.5	X	X	X
I don't know								
This doesn't apply to my data infrastructure								

E. Metadata record

How many of these data collections already have their dataset described with a metadata record in a metadata catalogue publicly available, prior to this deliverable?

Table 7: Metadata record

Do you have a metadata record API endpoint (m2m) in place?	**BBMRI-ERIC Directory**	**BIGAN**	**CRC cohort BBMRI**	**EGA**	**Electronic health record of the Andalusian Health System**	**Erasmus Glioma Database**	**SHIP**	**UK biobank**
Yes	X		X	X			X	X
No		X			X	X		

From the responses received we can conclude that 62.5% of the data collections have a metadata record with an API endpoint in place. The data collections without another

metadata record had low findability prior to this deliverable. Their addition to the HIP has increased their findability.

F. Unique identifier for metadata

It was considered important to add a question on whether the data infrastructure has a unique identifier for their metadata as it demonstrates that the data collection has been catalogued in a metadata catalogue, and is findable via the unique identifier.

Table 8: Unique identifier for metadata

Do you have a unique identifier for your metadata (e.g., uuid)?	**BBMRI-ERIC Directory**	**BIGAN**	**CRC Cohort**	**EGA**	**Electronic health record of the Andalusian Health System**	**Erasmus Glioma Database**	**SHIP**	**UK Biobank**
Yes	X Persistent identifier (PID)		X	X A numeric code			X	X
No		X				X		
I don't know								
This doesn't apply to my data infrastructure					X			

Five of the data infrastructures (62.5%) reported that they have a unique identifier for their metadata, which is consistent with the previous question on whether they already had a record in a metadata catalogue. Two responded that they do not. One (the Electronic Health Record of the Andalusian Health Service) reported that the question does not apply to their data infrastructure.

4.3. FAIRness evaluation

As explained in the methodology section above, a FAIRness evaluation was carried out on the new data collections added by HealthyCloud partners, using the FAIRness evaluation tool developed by WP3. The FAIRness evaluation tool, including the questions it includes and the corresponding point allocations, can be found on Zenodo [6].

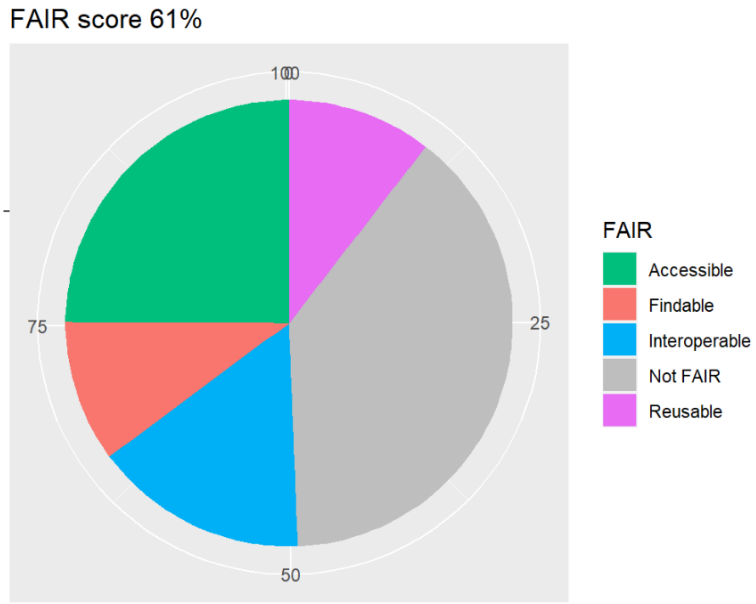
The HealthyCloud FAIRness evaluation tool produces a percentage score and a pie chart for each data collection. The evaluations were carried out by partners in WP3. It was ensured that the evaluation was carried out by someone external to the data collection/infrastructure, to ensure the evaluation was objective and reflected as closely as possible the FAIRness of the data collection. The WP3 partners were asked to provide feedback on their experience carrying out the FAIRness evaluation.

The results of the FAIRness evaluation for the new records added by HealthyCloud partners are shown below. The evaluation could not be carried out for the already existing records in the HIP as all the necessary properties for the evaluation were not covered in the metadata template.

The questions and corresponding points of the HealthyCloud FAIRness assessment tool have been included in Annex 2, to aid with the interpretation of the findings presented below.

1. BIGAN

Findability	Accessibility	Interoperability	Re-usability
41%	100%	62%	42%

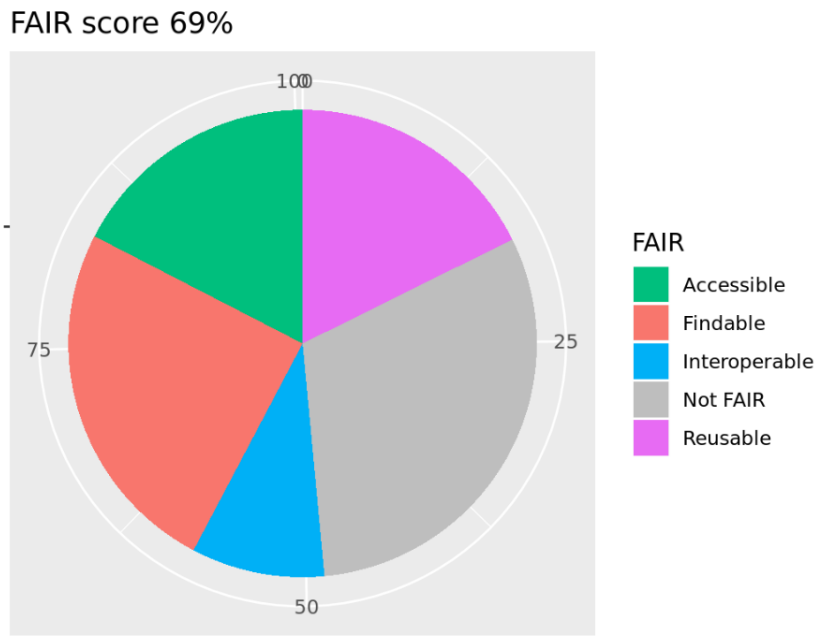


BIGAN had an overall FAIRness score of 61%, with 100% in accessibility and lower scores in findability, interoperability and reusability. The lower findability score was due to the fact that it was indicated in the BIGAN responses to the six additional fields that it does not have a unique identifier for the metadata, nor does it have a public metadata catalogue service. Interoperability was slightly reduced as it uses standardised vocabularies/ontologies/schemas without global identifiers, giving a score of 1 on that question as opposed to the maximum of 2. The re-usability score was lowered as it is not possible to access data and re-use it for more than one

project/purpose, and the assessor did not know if the information about the data collection was available in an open access repository.

2. *CRC Cohort*

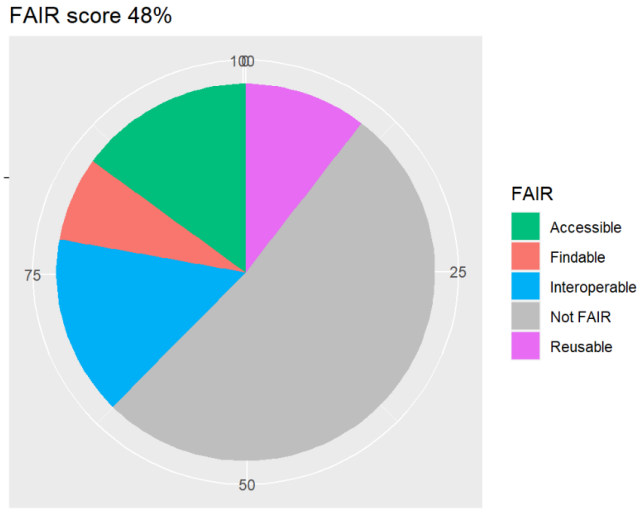
Findability	Accessibility	Interoperability	Re-usability
100%	70%	37%	71%



The CRC Cohort received an overall FAIRness score of 69%. The findability score was 100% as there is a unique identifier for both the dataset and metadata, metadata is produced and provided in a machine readable format, provided in publicly available metadata catalogue. Accessibility and re-usability were both slightly lower at 70% and 71% respectively. The accessibility score was slightly lower as it appears that data is not provided in a secure processing environment, as the method for data provision is decided by the individual data holders [9]. This accessibility requirement to use a secure processing environment is part of the questions in the FAIRness evaluation assessment tool but it is not essential. The re-usability score is slightly lower as it appears that researchers cannot apply for data and re-use it for more than one project. Finally, interoperability scored lowest at 37% as the format for distributing data was noted as 'Other' in the excel for additional fields, reducing the interoperability score.

3. *Electronic Healthcare Record of the Andalusian Public Healthcare System*

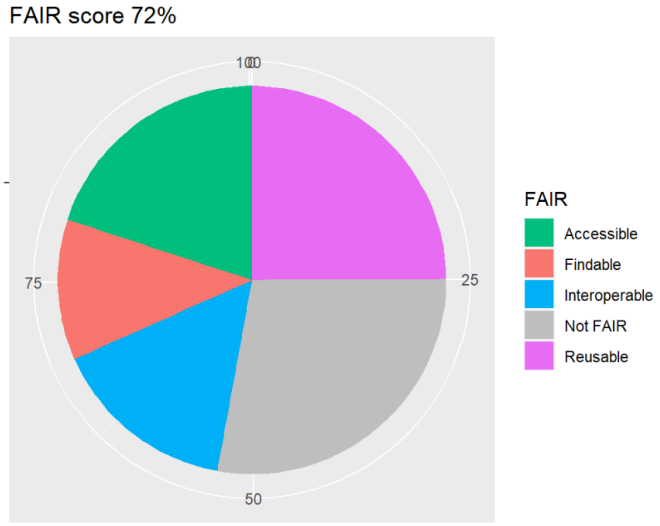
Findability	Accessibility	Interoperability	Re-usability
29%	60%	62%	42%



The Electronic Health Record of the Andalusian Public Healthcare System scored an overall score of 48% for FAIRness. The findability score was reduced by the fact that metadata is not produced for the data and therefore metadata is not publicly available. The accessibility score was reduced because there is no openly available published protocol for data provision for external researchers (e.g., data cannot be downloaded or analysed by external researchers if there is no collaboration agreement through a research project with relevant ethics committee approvals), giving a score of 0 on that question. Interoperability was slightly reduced as it uses standardised vocabularies/ontologies/schemas without global identifiers, giving a score of 1 on that question as opposed to the maximum of 2 (for vocabularies that are open and universal using resolvable global identifiers). In addition, a score of 0 was received for the question on an API endpoint, as this does not apply to this data infrastructure. The re-usability score was reduced by the fact that, based on the information that the assessor could find, they indicated it is not possible for users to access data and re-use it for more than one project, and the information about the data collection had not been placed in an open access repository. It is interesting to note that when the FAIRness evaluation was performed by a person internal to the Andalusian Public Healthcare System the FAIRness score was increased, with an overall FAIRness score of 53% (findability 29%, accessibility 50%, interoperability 62%, reusability 71%). This is discussed in the discussion section below.

4. Erasmus Glioma database

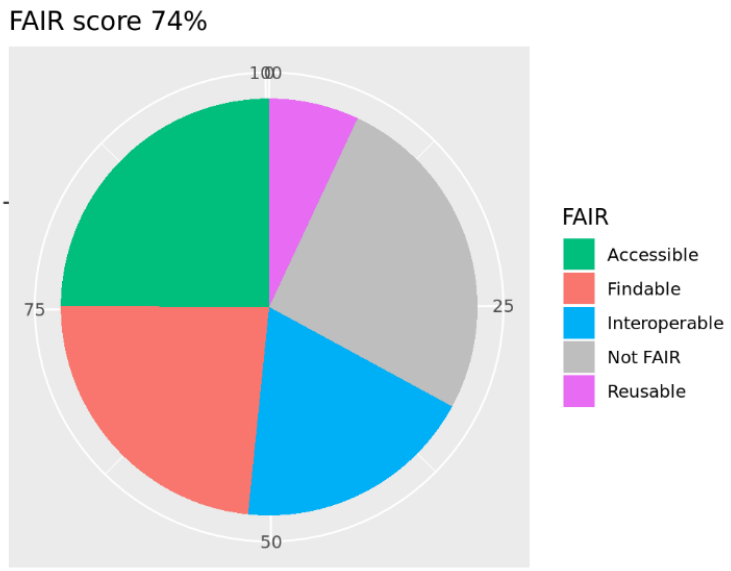
Findability	Accessibility	Interoperability	Re-usability
47%	80%	62%	100%



The Erasmus Glioma Database had an overall FAIRness score that was relatively high at 72%. The lowest score was in findability (47%) as it does not have a unique identifier for the metadata, metadata is collected in a text-based format non-standard format (as opposed to using a machine-readable format) and there is no public metadata catalogue service. Interoperability was lower at 62% as community recognised standards are not used. The accessibility score is 80%, slightly reduced by the fact that it was indicated that access is provided by downloading a file from an online location, as opposed to in a secure processing environment or via an API. The re-usability score is 100% as it was indicated that there is a clear procedure for third party users to request data for re-use, the information about the data collections has been placed in an open access repository, and it is possible for third party users to access the data and re-use it for more than one purpose/project.

5. European Genome Phenome Archive

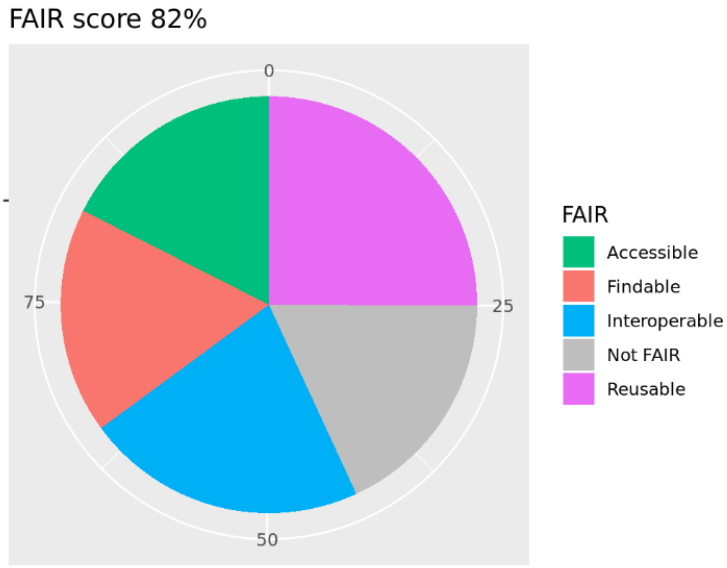
Findability	Accessibility	Interoperability	Re-usability
94%	100%	75%	28%



The European Genome-phenome Archive (EGA) had an overall FAIRness score of 74%. The lowest score was on re-usability as the assessor did not have the information on the questions 'Is there a clear procedure for third party users to request (the licence) for data re-use?' and 'Have you placed in an open access repository this information about your data collections?'. The slightly lower score for interoperability is due to the fact that the EGA had indicated 'This does not apply to my data infrastructure' for the question on community-recognised standards in use, providing a score of 0 for that question.

6. SHIP

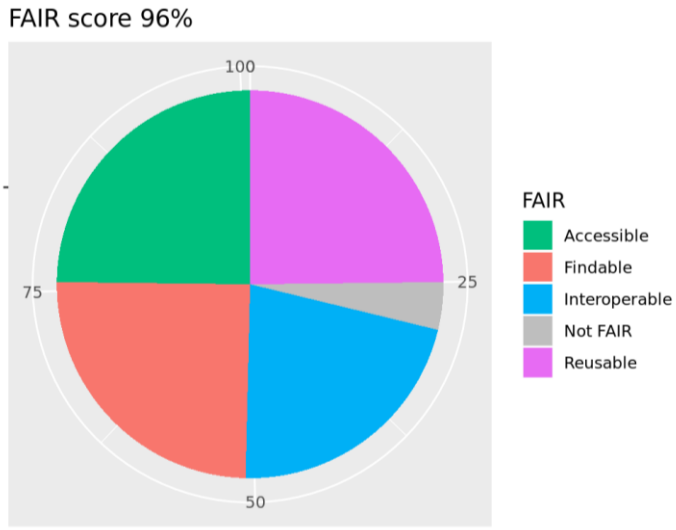
Findability	Accessibility	Interoperability	Re-usability
70%	70%	87%	100%



The Study of Health in Pomerania (SHIP) scored a relatively high overall score of 82%. The findability score was slightly reduced as the assessor did not know whether there was a unique identifier for the data sheet. Accessibility was reduced as the process for access provision is decided by individual arrangement (providing a score of 1 vs the highest score of 4 if access is provided in an SPE or standard web service API).

7. UK Biobank

Findability	Accessibility	Interoperability	Re-usability
100%	100%	87%	100%



The UK Biobank scored very highly in overall FAIRness, with 100% across findability, accessibility and re-usability. For findability, it has a unique identifier for both data and metadata, and produces publicly available metadata using an internationally it recognised metadata schema. For accessibility, it provides access to individual and

aggregated data to third party users, allows access in a secure processing environment, the conditions for access are published, and third party users must register to gain access. For re-usability, there is a clear procedure for third party users to request (the license) for data re-use, the information about the data collection is available in an open access repository, and it is possible for third party users to access the data and re-use it for mote than one purpose/project. It scored slightly lower in interoperability with a score of 87%, as it was indicated that it uses standardised vocabularies/ontologies/schemas without global identifiers (e.g., ICD-10, ICD-11), providing a score of 1 on that question as opposed to the maximum score of 2.

5. Discussion

Co-beneficial process

Overall, the collaboration between HealthyCloud and PHIRI was beneficial for both projects. For HealthyCloud, it allowed the completion of the landscape analysis of health-related data collections in Europe, including over 330 data collections relevant to population health research. The metadata records for these data collections can be viewed on the European Health Information Portal (HIP).

In terms of the benefits for PHIRI, the collaboration allowed the addition of new data collections to the HIP, and extended it to new data types, adding several European level data collections. The collaboration will also facilitate the improvement of the HIP metadata template through the addition of properties identified by HealthyCloud, which will add key information to the metadata records once added to the metadata template.

Limitations

Some limitations can be identified to this study. Firstly, several of the properties in the HIP do not use controlled vocabularies. The terms are not always based on existing definitions. The use of controlled vocabularies is important to ensure interoperability of metadata catalogues.

Another limitation is that currently in the HIP there is no place to add all pan-European metadata records, as all metadata records currently need to be associated with a national node, a country. Therefore, some European metadata records from European research infrastructures for example may have been omitted. In total there are five European metadata records on the HIP portal that do not correspond to a national node. This has been fed back to the PHIRI project.

There was only one metadata record added by HealthyCloud partners from an electronic health record (EHR) from a healthcare system. It was noted by WP3 partners that this should be taken into account when interpreting the respective FAIRness evaluation score, as this cannot be compared to the FAIRness assessment results with other data infrastructures that are aiming to facilitate re-use of health

data. The objective of the data collections are different and EHRs have as a primary purpose healthcare provision services rather than the re-use of the data collected.

On the FAIRness evaluation, several comments were received from WP3 partners. It was noted that given that the evaluation was being carried out by someone external to the data infrastructure, the accuracy of the evaluation depends on the information at hand and it is possible the FAIRness evaluation may be inaccurate due to a lack of information. The accuracy could be increased by cross-checking the responses with someone responsible for the data collection, as a person responsible for the data collection could provide explanations for the fields with lower scores from the external evaluators. For this deliverable, such a comparison was available for the Andalusian Health Service, as the evaluation was carried out by both an internal and external evaluator. It would have been good to perform the same exercise for the other data collections, however this was not possible for the current deliverable due to time constraints.

Conversely, it can also be said that performing the FAIRness evaluation by an independent assessor can be seen to provide a more realistic FAIRness score as it indicates how publicly available the information is from the user perspective. “From the perspective of a researcher, the information only ‘exists’ if it is publicly documented and findable with a reasonable effort. Or seen from the other end: a data collection is only as good as it sells itself.”

6. Conclusion

In conclusion, this document presents the extended landscape analysis, beyond the one introduced in D3.1, of health-related data infrastructures in Europe. The aim of this extended landscape analysis is to analyse the European health data collections available for research purposes, evaluate their FAIRness level and determine the feasibility to perform individual level data linkages. This landscape analysis was achieved by a collaboration between HealthyCloud WP3 and PHIRI, making use of the Health Information Portal’s metadata catalogue, which includes over 330 records.

Overall, the landscape analysis shows that there is a vast amount of health-related data collections in Europe, with different features. Many more exist which have not yet been recorded in the HIP. The results show that most data collections in this landscape analysis are at national or regional level. Only a small proportion of data collections have multi-country or European level data. Most have data from the general population, and the most common data sources are registries and surveys. These findings are in line with PHIRI’s focus on population health as well as the methodology used to add data collections to the HIP, which was previously mainly carried out through national node networks. The collaboration with HealthyCloud allowed the addition of several European level data infrastructures. It also raised the question about how data hubs can be accurately represented in the European Health Information Portal. This question is being looked into by the PHIRI team.

Regarding the feasibility to perform individual level data linkage, the majority of data collections recorded in the HIP have individual level data and use a national identifier, demonstrating good potential for linkability. In addition, only 12% indicated that linkage is not possible with other datasets.

The FAIRness analysis using the HealthyCloud FAIRness evaluation tool could only be performed on the new data collections added by HealthyCloud partners, as some of the responses necessary were covered only in the six additional fields, which have not yet been added to the HIP metadata template. The calculated FAIRness levels of the data collections varied widely, as the data collections added by HealthyCloud partners varied between those whose main purpose is re-use for research and some whose main purpose is primary use for healthcare. In addition, it was noted by WP3 partners that lower scores may be attributable to the fact that they were performed by people external to the organisation.

As noted previously, this collaboration has benefited both projects. It allowed the inclusion of over 330 records in the HealthyCloud landscape analysis. For PHIRI, it has extended the HIP to include new data types, several European level data collections, and has identified areas to improve the HIP metadata template and adapt data hubs, increasing its coverage of health-related data collections and infrastructures in Europe.

3. References

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

[2] HealthyCloud WP3 and WP4 survey. Online tool. Available at: https://bsc3.typeform.com/to/zY1FNgSQ?typeform-source=www.google.com

[3] Cosgrove S., Derycke P., Kesisoglou I. et al. HealthyCloud Deliverable 3.1 Landscape analysis of FAIRness levels of health-related data using catalogue matrix. (2022). Available at: https://healthycloud.eu/wp-content/uploads/2022/11/D3.1.pdf

[4] Tolonen et al. Documentation and user guide for the Health Information Portal: Metadata description. (2022). https://doi.org/10.5281/zenodo.6413408

[5] Australian Research Data Commons. FAIR Data Self Assessment Tool. Last updated 12 May 2022. Available at: https://ardc.edu.au/resource/fair-data-self-assessment-tool/

[6] HealthyCloud FAIRness Assessment Tool Binder. Available at: https://ovh.mybinder.org/v2/gh/PderyckeSciensano/HEALTHYCLOUD/main?urlpath=rstudio

[7] Derycke P., Kesisoglou I., Cosgrove S. HealthyCloud FAIRness Assessment Tool. (2022). Available at: https://doi.org/10.5281/zenodo.7038397

[8] HealthyCloud FAIRness Assessment Tool GitHub. Available at: https://github.com/PderyckeSciensano/HEALTHYCLOUD/

[9] BBMRI BBMRI-ERIC Policy for Access to and Sharing of Biological Samples and Data. Available at: https://www.bbmri-eric.eu/wp-content/uploads/AoM_10_8_Access-Policy_FINAL_EU.pdf

Annex 1: EU HIP metadata template properties

Table 1: Properties asked by the EU HIP metadata record template

Title
Alternative title
Acronym
Type of information
URL of the data source
Description
Governance and legal framework
Funding
Topics
Free keywords
GEO coverage
Country(ies)
Target Population
Age range (from)
Age range (to)
Sample size
Sex

Access information
Data Collection Period
Language
Updating Periodicity
Personal Identifier
Please specify:
Level of aggregation
Linkage possible
Permanent identifier of the data source
Terms of data access
Terms of data access - URL
Data Owner(s)
Institution
Regulations for data sharing
Contact name
Contact e-mail
Contact info (address)
Contact phone number

Table 2: 6 additional properties added from the HealthyCloud survey

INDICATORS	Description of the indicator (example)	Format of the input
Type of data	Specify the type of data collected.	Multiple choices possible: / Images / Text / Numbers / Files / Tissue samples / Sounds / Other (please specify)
	If other, please specify	Free text
Anonymisation	Is the data stored within the data infrastructure anonymised or identifiable?	Select a single response: / Anonymised / Identifiable / I don't know / This question doesn't apply to this data infrastructure
Unique identifier for metadata	Do you have a unique identifier for your metadata (ex: uuid)?	Select a single response: / Yes / No / I don't know / This question doesn't apply to this data infrastructure
	If yes, what type of unique identifier (example: uuid)?	Free text

Standards used for metadata and data	Which community-recognised vocabularies, standards or methodologies are used for data to facilitate interoperability? Conforms to (dct property)	Multiple options possible: / SNOMED CT / LOINC / ICD-10 / ICD-11 / Other / I don't know / This doesn't apply to this data infrastructure
	If other, please specify	Free text
Data format for exchange	What is the format(s) for distributing data?	Multiple options possible: / csv / xml / json / ld-json / pdf / R / SAS / Other / I don't know / This doesn't apply to this data infrastructure
	If other, please specify	Free text
Metadata record	Do you have a metadata record API endpoint (m2m) in place?	Select a single response: / Yes / No / I don't know / This question doesn't apply to this data infrastructure

Annex 2: HealthyCloud FAIRness assessment tool questions

Table 1: HealthyCloud FAIRness assessment tool questions and corresponding points

Question	Response	Points
Findable		
Do you have a unique identifier for your datasets?	Yes	5
	No	0
	This question doesn't apply to my data infrastructure	0
	I don't know	0
Do you have a unique identifier for the metadata?	Yes	4
	No	0
	I don't know	0
Do you produce or collect metadata for all your data (e.g. handbook, guide for users, description, keywords, timestamp, spatial coverage etc.)?	Comprehensively, using a recognised formal machine-readable metadata schema	4
	Comprehensively, but in a text-based, non-standard format	3
	Brief title and description	2
	The data is not described	0
	I don't know	0
Do you have a public metadata catalogue service?	Yes	4
	No	0
	This question doesn't apply to my data infrastructure	0
	I don't know	0
Accessible		

Do you provide access to individual and/or aggregated data (for third party users)?	Provides access to individual or aggregated data	4
	Publicly accessible aggregated	3
	This question doesn't apply to my data infrastructure	0
	I don't know	0
Is it possible to extract the data from the data infrastructure (e.g. download) or do they have to stay in the data infrastructure?	It is possible to access data for analysis in a remote secure processing environment	4
	Standard web service API (e.g., OGC)	4
	Non-standard web service (e.g., openAPI/Swagger/informal API)	3
	File download from online location	2
	By individual arrangement	1
	It is not possible to extract the data	0
	I don't know	0
Do third party users have to register to the data infrastructure and have an account in order to access the data?	Yes	1
	No	0
	This question doesn't apply to my data infrastructure	0
	I don't know	0
Are the conditions of access published?	Yes	1
	No	0
	I don't know	0
Interoperable		
What is the format(s) for distributing data?	In a structured, open standard, machine-readable format (csv, xml, JSON, R…)	4

	In a structured, open standard, non-machine-readable format (e.g., pdf)	2
	Mostly in a proprietary format (e.g., SAS)	0
	Other	0
	This question doesn't apply to my data infrastructure	0
	I don't know	0
Which community-recognised vocabularies, standards or methodologies are used for metadata and data to facilitate interoperability?	Standardised open and universal using resolvable global identifiers linking to explanations (Open ex: …)	2
	Standardised vocabularies/ontologies/schema without global identifiers (e.g., HL7 FHIR, SNOMED CT, LOINC, ICD-10…)	1
	Other	1
	Data elements not described	0
	This question doesn't apply to my data infrastructure	0
	I don't know	0
Do you have a metadata record API endpoint (m2m) in place?	Yes	2
	No	0
	This question doesn't apply to my data infrastructure	0
	I don't know	0
Reusable		
Is there a clear procedure for third party users to request (the license) for data re-use?	Yes	3
	No	0

	This question doesn't apply to my data infrastructure	0
	I don't know	0
Have you placed the metadata related to your data infrastructure (that is, the above information provided in this survey) in another available source already?	Yes	2
	No	0
	This question doesn't apply to my data infrastructure	0
	I don't know	0
Is it possible for third party users to access the data and re-use it for more than one purpose/project?	Yes	2
	No	0
	This question doesn't apply to my data infrastructure	0
	I don't know	0