



Project acronym: HealthData@PT

Project full title: Setting up a Health Data Access Body in Portugal and laying the national foundations for EHDS2

WP6 - National Dataset Catalogue for Health Data

D6.1 - National dataset catalogue requirements and specifications

Project Coordinator	SPMS – Serviços Partilhados do Ministério da Saúde, E.P.E. (Portugal)		
Start date of the project	2023-11-01	Duration	48 months
Work Package	WP6 - National Dataset Catalogue for Health Data		
Status	In Progress		
Document Version	1.0		
Dissemination Level	PU – Public		
Main Author(s)	Nuno Cruz, André Fontes, João Santos, Andreia Serrano, Rúben Jerónimo, Vanessa Lima, Sofia Ferreira, Sara Marques		
Due Date	2027-10-31		
Deliverable Date	2025-08-18		
Deliverable Type	R — Document, report		

This project is co-funded by the European Union's EU4Health (2021-2027) programme under the Grant Agreement Nr. 101128332. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA, the 'granting authority'). Neither the European Union nor the granting authority can be held responsible for them.

Document Revision History

Version	Date	Summary of main changes	Author(s)
0.1	2023-12-01	Document creation and Table of Content drafting	Sara Marques
0.2	2024-04-01	Document structure; Introduction; HealthDCAT-AP	Nuno Cruz
0.3	2024-07-30	Background: datasets catalogue, vocabulary, national datasets catalogue – requirements and recommendations; healthdcat-ap: dcat vs dcat-ap, rdf, healthdcat-ap; portuguese context; Use cases	Nuno Cruz, André Fontes, João Santos, Andreia Serrano
0.4	2025-01-28	Use cases; Requirements and Specifications final draft	Nuno Cruz, André Fontes, João Santos, Andreia Serrano, Vanessa Lima, Sofia Ferreira, Sara Marques
0.5	2025-05-15	Conclusion of the 1 st draft of the portuguese national datasets catalogue requirements and specifications final draft (section 4); conclusion of the “Building up the Portuguese national datasets catalogue” section (section 5)	Nuno Cruz, João Santos, Andreia Serrano, Sara Marques
0.6	2025-06-02	Revision of the first draft of the portuguese national datasets catalogue requirements and specifications with technical teams.	Nuno Cruz, João Santos, Andreia Serrano, Sara Marques, Vanessa Lima, Rúben Jerónimo
0.7	2025-06-20	Preparation of the final draft of the portuguese national datasets catalogue requirements and specifications.	Nuno Cruz, João Santos, Sara Marques, Rúben Jerónimo
0.8	2025-08-08	Final revision and corrections	Sara Marques
1.0	2025-08-18	Final version for submission	

Executive Summary

This document constitutes the deliverable D6.1 “National dataset catalogue requirements and specifications” part of the work package 6 “National Dataset Catalogue for Health Data” of the HealthData@PT project.

To facilitate cross-border collaboration and promote the effective (re)use of health data across Europe, the European Health Data Space (EHDS) was established as a unified framework for health data. This initiative addresses key challenges such as the fragmentation of data repositories, the lack of harmonized international standards, and the inconsistent methods for locating and requesting access to health data.

Under this regulation, Member States are required to establish a Health Data Access Body which, among other responsibilities, must provide information about the available datasets for secondary use purposes and their key characteristics. This ensures that potential health data users are informed of essential details and can find and assess the relevance of the datasets to their needs. This machine-readable dataset catalogue refers to a systematically organized collection of dataset descriptions in the form of metadata, including a public, user-oriented section through which information on individual dataset parameters is electronically accessible via an online and public national portal.

In D6.1, the outlining of the functional and technical specifications and requirements for the HealthData@PT’s datasets catalogue are defined based on several key sources: the “Requirements Catalogue for Scale-up Version (D02.03)” provided by DG SANTÉ, the requirements of the EHDS Regulation, and a landscape analysis of catalogue-related recommendations resulting from TEHDAS, HealthyCloud, the EHDS2 Pilot, and TEHDAS2 (particularly TEHDAS2 WP5 deliverables). Additionally, insights from dataset catalogues from Sweden, Norway, Finland, and France have been considered, as well as the extensive work on interoperability and on the standardization of electronic health data within the Portuguese context, which pave the way for the groundwork necessary towards the development of the Portuguese national metadata catalogue.

The resulting framework is presented in Section 4 of this deliverable, in which the defined requirements and specifications for the HealthData@PT dataset catalogue are outlined guaranteeing alignment with the EHDS requirements. Furthermore, details on the first datasets under construction and their metadata specifications, as well as screenshots of the preliminary version of the catalogue under development by August 2025 can be consulted in sections 5.1, 5.2 and 5.3 of this document, respectively.

The framework outlined in sections 4 and 5 will govern the national datasets design and implementation throughout HealthData@PT.

Table of Contents

Executive Summary	ii
Abbreviations and Acronyms	viii
1. Introduction.....	1
1.1. Background	1
1.2. National Datasets Catalogue	4
2. Scope of this Deliverable	6
3. Landscape analysis on requirements and specifications	7
3.1. Gathered recommendations from European initiatives	7
3.1.1. TEHDAS	7
3.1.2. Healthy Cloud	8
3.1.3. EHDS2 Pilot	9
3.1.4. TEHDAS2 and European Commission's documents.....	9
3.2. The European Standard for metadata: HealthDCAT-AP	10
3.3. Examples of operational National Dataset Catalogues	18
3.3.1. Swedish catalogue	19
3.3.2. Norwegian catalogue	22
3.3.3. Finnish (Findata) catalogue.....	26
3.3.4. French catalogue	28
3.4. Fair Data Points	31
3.5. Current Portuguese Context	32
3.5.1. The Portuguese NHS Transparency Portal	32
3.5.2. Other national developments.....	34
4. Requirements and Specifications for the Portuguese National Datasets Catalogue ..	36
4.1. A public national portal holding the national datasets catalogue	36
4.2. Datasets description – HealthDCAT-AP standard and variables.....	39
4.3. Datasets creation, maintenance, storage, review and auditability.....	39
4.4. Catalogue Backend.....	40
4.5. API for the data access application/data request form portal	41
4.6. Metadata ingestion	41
4.7. Catalogue API for synchronisation with HealthData@EU datasets catalogue upon CRUD operations	42
4.8. Non-functional requirements	42
5. Building up the national datasets catalogue.....	44
5.1. Dataset Specifications	44
5.2. Metadata specifications	46
5.3. nHDsC Mock-up.....	47

5.4. National Implementation..... 48

6. Concluding Remarks..... 49

7. Bibliography..... 50

Annexes 52

Annex 1. HealthyCloud WP3 recommendations 53

Annex 2. Mandatory Properties for the HealthDCAT-AP Standard 54

List of Figures

Figure 1. An example of a datasets catalogue (retrieved from EUCAIM project). On the left of the image, expandable lists of different filters are available, as well as a free-text search bar. ...	3
Figure 2. Data users' journey in the EHDS, sourced from TEHDAS (Graphic by Topias Dean, SITRA). [9]	5
Figure 3. Representation on how the different standards expand on each other: RDF working as the core foundation of the DCAT standard, further extensions increase their specificity.	10
Figure 4. HealthDCAT-AP mandatory properties for the "Dataset" class. Excerpt from original in [19].	12
Figure 5. Example of a dataset in the EHDS2 Pilot Sandbox. In column at the left, the first rectangle represents the current DCAT-AP property type (M - Mandatory, R-Recommended, O-Optional) when applicable, and the second rectangle represents the HealthDCAT-AP property type.	15
Figure 6. HealthDCAT-AP editor, available at http://ehds.healthdataportal.eu/editor2/ (screenshot from the version at 20/03/2025).....	16
Figure 7. Current version of the HealthDCAT-AP Records Sandbox, now acting as the online backup for all dataset descriptions published via the Editor.....	17
Figure 8. Example of a dataset description available in the HealthDCAT-AP Record Manager. ..	17
Figure 9. Example of the information depicted to the user when filtering by Registry name.	20
Figure 10. Example of the information page for a register (U-CAN) in this Swedish catalogue..	21
Figure 11. Example of the sub-register comparison tool for temporal coverage in the Swedish catalogue. Three sub-registers were selected, with temporal coverage overlap existing only between two of them. Consequently, a message is displayed to the user indicating that no contemporary data is available.	22
Figure 12. First page of the Norwegian catalogue, divided into two sections featuring tools for exploring data sources (a) and data access guidelines (b).	23
Figure 13. Example of the page layout for a dataset in the Norwegian catalogue.	24
Figure 14. Example of the variable search page, enabling search using free text or applying different filters (presented on the left). The middle tab is automatically updated according to the chosen filtering type and provides more information about the selected items. On the right, the list of variables, refined according to the applied filters.....	24
Figure 15. An example of a variable when searching for 'diabetes,' when a data source and a data collection has been chosen. A variable group is selected (but not yet marked) which expands the information in the tab designated for this purpose. The list of variables corresponding to the selected filtering is refined accordingly.	25
Figure 16. Example of a search for "smoking" in the query for variables within the Finnish national catalogue.	26
Figure 17. Example of a search for data resources in the Finnish catalogue.	27
Figure 18. Example of the page layout for the data source "Finish Cancer Registry".	27
Figure 19. Example of a page layout for a dataset from the Cancer Registry in the Finnish national catalogue, highlighting the option to download variables in CSV format.....	28

Figure 20. Metadata workflow for metadata synchronization between national catalogues and the nHDsC.....	29
Figure 21. French Data Catalogue search functions. Existing filters are "medical domain" and "data category".	30
Figure 22. French Data Catalogue, example of a dataset registry. Information includes data controller and scientific manager, population of interest, dataset size, temporal coverage, provenance, database documentation, etc.....	31
Figure 23. Portuguese NHS Transparency Portal Catalogue.....	32
Figure 24. Portuguese NHS Transparency Portal Catalogue - Antibiotics dataset.....	33
Figure 25. Portuguese NHS Transparency Portal - User generated map with a dataset overlay for antibiotic consumption.	33
Figure 26. Portuguese NHS Transparency Portal Catalogue Datasets' metadata generation flowchart.	34
Figure 27. Some of the Clinical Terminologies Centre's Catalogues.....	35
Figure 28. Architecture of the metadata flow, with particular emphasis on the role of the National Datasets Catalogue, from data holders to the European Datasets Catalogue.	36
Figure 29. First development version of the Portuguese national datasets catalogue.	47
Figure 30. Example of a dataset record in the first development version of the Portuguese national datasets catalogue.	48

List of Tables

Table 1. National datasets computational and storage infrastructure options.....	8
Table 2. Main entities (Classes) that make up the HealthDCAT-AP standard [19].....	13
Table 3. Some property differences between HealthDCAT-AP and EUCAIMDCAT for the "Dataset" class. DCAT-AP comparison provided as baseline.	18
Table 4. Sample of 8 datasets described with the HealthDCAT-AP standard, for the nHDsC MVP.	44
Table 5. HealthDCAT-AP Recommended Properties that will be mandatory in the PT nHDsC metadata.	46
Table 6. HealthyCloud WP3 recommendations to uphold FAIR data principles in Health Data that are related with datasets' catalogues.....	53
Table 7. Mandatory properties for the main classes of the HealthDCAT-AP Standard (for non-public data). Classes without mandatory properties are instead hyperlinked to their definition [19].	54

Abbreviations and Acronyms

CDM	Common Data Model
CPAL	Portuguese National Catalogue for Clinical Analysis
CRUD	Create, Read, Update, Delete
CSIRT	Computer Security Incident Response Team
D	Deliverable
DAAMS	Data Access Application Management System
DBC	Digital Business Capability
DCAT	Data Catalogue
DCAT-AP	Data Catalogue Application Profile
DOI	Digital Object Identifier
EC	European Commission
EHDS	European Health Data Space
EHDS2	European Health Data Space for Secondary Use
EU	European Union
EUCAIM	European Federation for Cancer Images
FAIR	Findable, Accessible, Interoperable, Reusable
FDP	FAIR Data Point
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
HDAB	Health Data Access Body
HDH	French Health Data Hub
HealthDCAT-AP	Health Data Catalogue Application Profile
HTTPS	Hyper Text Transfer Protocol Secure
ICPC	International Classification of Primary Care
MVP	Minimum Viable Product
nHDsC	National Health Dataset Catalogue
NHS	National Healthcare System
OMOP CDM	Observational Medical Outcomes Partnership Common Data Model
PRR	Recovery and Resilience Plan
RDF	Resource Description Framework
SEMIC	Semantic Interoperability Community
SIEM	Security Information and Event Management
SIMSNS	NHS Information and Monitoring System
SPE	Secure Processing Environment
SPMS	SPMS – Serviços Partilhados do Ministério da Saúde, E.P.E. – Portugal (<i>Shared Services of the Ministry of Health</i>)
SSO/IAM	Single Sign-on/Identity Access Management

TEHDAS	Towards the European Health Data Space
TEHDAS2	Second Joint Action Towards the European Health Data Space
URL	Uniform Resource Locator
WP	Work Package

1. Introduction

Across Europe, the health data landscape is fractured: the fragmentation of data repositories, following of different international standards and differences in how to find and request access to health data greatly hinder cross-border collaboration and health data (re)use. To combat this, a common European environment for health data was created: the European Health Data Space (EHDS).

To establish the EHDS, Member-States are required to design and implement multiple digital business capabilities (DBC), one of which being a national health datasets catalogue that contains the corresponding dataset descriptions (metadata), to inform data users on the datasets available for secondary use purposes, improving data discovery. Currently, Portugal lacks such a catalogue. As such, it is necessary to scope the requirements and specifications of such a solution for a successful implementation during the HealthData@PT project.

This document will describe these requirements and specifications for the Portuguese catalogue, based on an analysis of the current European health data environment and analogous solutions already available. It will also leverage from the main outputs of relevant international projects, such as any key standards that are to be implemented. The outputs from this document will then be used to guide the design and implementation of the national health datasets catalogue.

1.1. Background

According to the Article 57 of the EHDS Regulation, it is an obligation of the national health data access bodies (HDABs) across the European Union (EU) to provide and maintain a publicly available dataset catalogue *“that shall include details about the source and nature of electronic health data, in accordance with Articles 77 and 78, and the conditions for making electronic health data available.”* [1] Through this description, it can be concluded that this datasets catalogue will be a **repository** of descriptive information about the available datasets for secondary use purposes, a repository of datasets’ metadata.

To deconstruct these concepts, it is necessary to define what is a dataset catalogue, as well as to distinguish data from metadata.

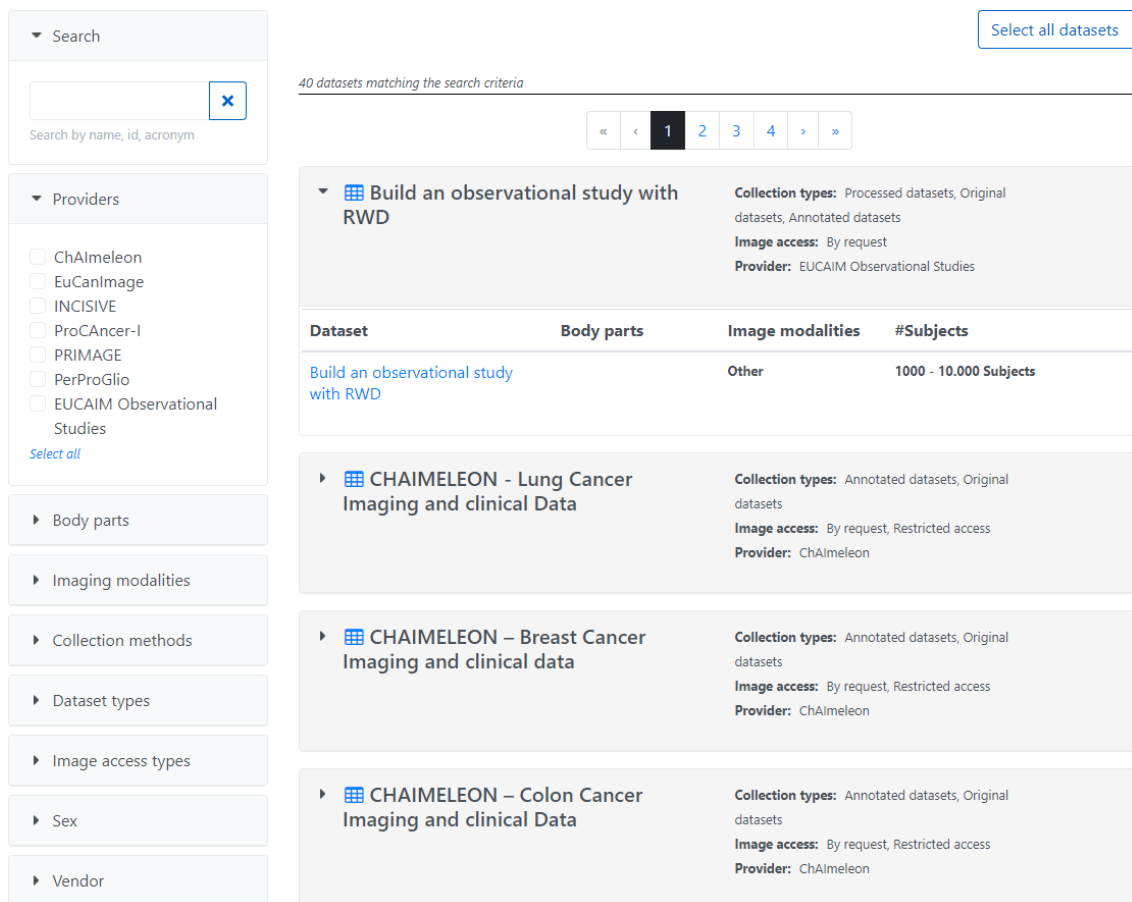
A datasets catalogue is a library, containing descriptive information regarding those datasets (metadata). [2,3] It functions as a key tool for data discoverability, allowing potential users to more easily search and find datasets relevant to their work, not only because it works as a repository of useful information, but also via the inclusion of standard search tools (plain text search, filters, information regarding source and type of data, etc). Article 77 of the EHDS legislation states the required characteristics of the **national datasets catalogue**: [1,4]

- Publicly available;
- Standardized, machine-readable format;
- Provides metadata on the available datasets and their characteristics;

- Source, scope, nature of electronic health data and conditions for data availability;
- Information on the quality and utility of the data, and maturity of the data holder, in accordance with Article 78.
- Available in, at least, one official language of the Union:
 - Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish and Swedish;
- Follow the specifications on minimum elements for datasets and their characteristics set in the respective EHDS implementing acts.

As an example, Figure 1, featuring the European Federation for Cancer Images (EUCAIM)¹ datasets catalogue, shows what such a datasets catalogue may look like.

¹ The EUCAIM project aims at deploying a pan-european federated and interoperable digital infrastructure of anonymised cancer imaging data, including data from existing cancer image repositories and from various sources, such as clinical imaging, pathology, molecular and laboratory data. For more information, consult <https://cancerimage.eu>



▼ Search

Search by name, id, acronym

▼ Providers

- ☐ ChAlmeleon
- ☐ EuCanImage
- ☐ INCISIVE
- ☐ ProCancer-I
- ☐ PRIMAGE
- ☐ PerProGlio
- ☐ EUCAIM Observational Studies

Select all

► Body parts

► Imaging modalities

► Collection methods

► Dataset types

► Image access types

► Sex

► Vendor

40 datasets matching the search criteria

« < 1 2 3 4 > »

▼ Build an observational study with RWD

Collection types: Processed datasets, Original datasets, Annotated datasets
Image access: By request
Provider: EUCAIM Observational Studies

Dataset	Body parts	Image modalities	#Subjects
Build an observational study with RWD		Other	1000 - 10.000 Subjects

► CHAIMELEON - Lung Cancer Imaging and clinical Data

Collection types: Annotated datasets, Original datasets
Image access: By request, Restricted access
Provider: ChAlmeleon

► CHAIMELEON - Breast Cancer Imaging and clinical data

Collection types: Annotated datasets, Original datasets
Image access: By request, Restricted access
Provider: ChAlmeleon

► CHAIMELEON - Colon Cancer Imaging and clinical Data

Collection types: Annotated datasets, Original datasets
Image access: By request, Restricted access
Provider: ChAlmeleon

Figure 1. An example of a datasets catalogue (retrieved from EUCAIM project). On the left of the image, expandable lists of different filters are available, as well as a free-text search bar.

Therefore, a dataset catalogue is a systematic way to **structure metadata in a discoverable, standardized way**. But how do distinguish between data and metadata? What are their defining characteristics? Data refers to facts, concrete information (e.g. health records), whereas metadata is “data about data”. Metadata contains descriptive information about the content of a dataset, not the content itself. For a book, this could represent information normally featured in a bibliographic reference: the authors, title, date of publication, editor, etc.

Metadata is important as an aid for data users to find relevant datasets for their work and learn more about the characteristics of the dataset. Furthermore, metadata is useful to organize, document and archive purposes. Given this, it comes as no surprise that metadata was commonly used in libraries, traditionally in the form of card catalogues before the transition to digital databases [5].

When it comes to the types of metadata, they follow different categories, with different expectations for their goal and content. Following the bibliographical example given earlier, metadata that is useful to find and understand the data it refers to is called *descriptive metadata*. Metadata that depicts the relationships between objects is *structural metadata*. *Administrative metadata* is useful for resource management, and might include information on long-term data management, or intellectual property. Finally, there are *markup languages*, which integrate metadata and content: in a textual context, this could mean marking paragraphs, providing additional semantic context to words or expressions, etc. [6]

Armed with the knowledge on how data and metadata differ, and for what metadata can be used for, it becomes important to understand how this “data about data” can be organized. *Metadata repositories* are databases where this metainformation is stored and given context so it can be understood as metadata. For the end-user, it empowers them to better find and retrieve relevant data for their work, by providing an agile way to query for datasets containing the necessary information. In short, these repositories are a tool for efficient database query, through organized storage of metadata pertaining to datasets. [7]

For these repositories to function in a standardized manner, the terms utilized, and their meanings and relationships need to be considered. For this, **vocabularies** are employed, providing consistency on the terms to be used and on their meaning, while **ontologies** define the classes, properties and relationships utilized to characterize metadata (and data). In the domain of computer science, this structure of terms and relationships - an Ontology, is the backbone of standards such the Data Catalogue (DCAT) standard. [8]

Also key in the management of datasets and their metadata is the notion of **interoperability**, that is, the ability to exchange data between different systems. This interoperability can be broken down in three main aspects: technical (how the data is exchanged), semantic (the meaning of the data) and syntactic (form the data takes, “language”). Having a structured, standardized catalogue based on a metadata standard is a way to increase interoperability.

In short, **dataset catalogues are a user-oriented metadata repository** focused on the **agile discovery of information** based on the desired dataset characteristics and that contains information on how to access the datasets themselves when possible. The types of information present in the datasets are defined by the vocabulary utilized, which can be specialized to facilitate access to certain information.

1.2. National Datasets Catalogue

The national datasets catalogue is a **mandatory digital business capability** for the secondary use of health data, as is through this catalogue that data can be discovered for secondary use (step 1 of the user journey, Figure 2). The national datasets catalogue will display the datasets available for secondary use purposed at national level, and is meant to interface with the HealthData@EU Datasets Catalogue, which feeds from each Member State’s national catalogues.

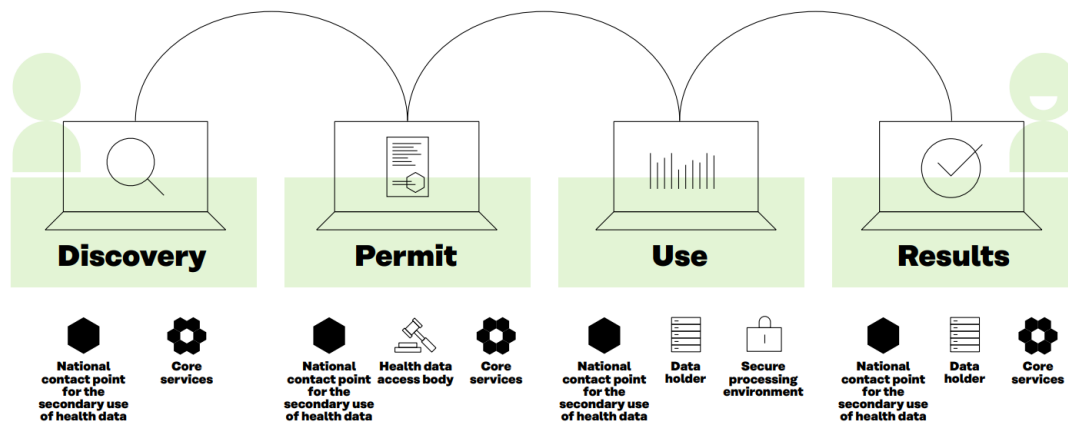


Figure 2. Data users' journey in the EHDS, sourced from TEHDAS (Graphic by Topias Dean, SITRA). [9]

Further, the EHDS regulation proposal states in its recitals that “(...) *HealthData@EU* should therefore enable the secondary use of different categories of electronic health data, including linking of the health data with data from other data spaces (...)”, as this interoperability has the possibility to create deeper insight into “health determinants”.

2. Scope of this Deliverable

This deliverable aims to describe the requirements and specifications for the Portuguese National Dataset Catalogue. Being under the purview of Task 6.1 - national Health Dataset Catalogue (nHDsC): Analysis and design of HealthData@PT, this document includes an analysis of the prerequisites for the national health datasets' catalogue, as well as sections on the use of structured data, international standards, dataset definitions, metadata requirements, dataset sources (data holders), and corresponding variables, as well as a harmonised technical sheet. This requirements and specifications will guide the development and implementation of the national datasets catalogue.

3. Landscape analysis on requirements and specifications

A landscape analysis was conducted to leverage the recommendations issued by European initiatives and the best practices developed both nationally and internationally to develop the requirements and specifications for the nHDS, which are present in Section 4.

This landscape analysis was divided into four main parts: the analysis of relevant recommendations for catalogue construction produced by European initiatives, a deep-dive on the HealthDCAT-AP standard, an overview of operational catalogues from other Member States and a brief overview of the Portuguese context regarding data cataloguing.

3.1. Gathered recommendations from European initiatives

To inform the specifications and requirements of the Portuguese datasets, recommendations from various European initiatives were consulted. A brief overview on these initiatives and on their relevant outcomes in the context of the construction of the present deliverable are provided next.

3.1.1. TEHDAS

TEHDAS Joint Action issued recommendations for the technical implementation of the EHDS2, namely options for the implementation of information systems to manage the national metadata catalogues [9]. Important recommendations regarding datasets' catalogue were developed during this joint action, building the rationale for subsequent work, namely for the development of the HealthData@EU central services, for the definition of the metadata standard to use on the EHDS2 datasets' catalogue, to name a few.

Likewise, recommendations regarding national datasets infrastructure options were also provided, as summarized in Table 1.

Additionally, important considerations about data safety also resulted from TEHDAS2, in accordance with the Data Governance Act: despite the metadata information in the catalogues will not include personal data, **there should be sufficient data security measures in place to prevent malicious acts, such as subversion of links to redirect to malicious websites.**

Quality metadata is key in a dataset catalogue to increase data findability and reuse, and thus the subject of metadata standards and semantic interoperability was broached in TEHDAS guidelines. A Data Catalogue Application Profile (DCAT-AP) based system was recommended due to its purpose-built design to support interoperability between data portals and catalogues in the public sector. The creation of an extension of this system, a HealthDCAT-AP, designed to address the specificity of health data was recommended. This standard was later adopted as the metadata standard for the EHDS, and will be explained in further detail in Section 3.2.

Table 1. National datasets computational and storage infrastructure options.

Infrastructure	Solution	Observations
Computational	Server (16 cores, 32GB RAM, 1TB backup)	Server specifications dependent on traffic, solution needs to be scalable to fit demand.
	Cloud server	Security standards of outsourced server need to comply with Secure Processing Environment (SPE) standards.
Storage	Data Lakes (Massive and centralised repository of raw data)	Size of the solution allows integration of the catalogue into the data lake. Federated architecture is possible. More complex and infrastructurally demanding than a Data Warehouse.
	Data Warehouse (Long-term storage solution for data)	Less demanding than a Data Lake, but normally institution based. Decentralization requires communication between warehouses and HDAB to satisfy data queries. Still requires large-scale storage solutions.
	Data Marts (Lower-scale long-term storage solution)	Designed to satisfy individual data requests to the HDAB, providing long-term storage of requested data. Highly decentralized.

Finally, it was also concluded that to facilitate adoption of the systems and standards necessary for the function of the EHDS2, it is necessary to assure data experts, be data users, holders or HDABs have their training and capacity building needs met. The role of a multi-approach strategy for education, dissemination and training to foster informed, data-aware individuals with the knowledge and skills to find, interpret and use datasets for their work was also highlighted in TEHDAS guidelines.

3.1.2. Healthy Cloud

HealthyCloud is a project dedicated to outputting various recommendations and specifications for distributed health research across Europe, compliant with the EHDS regulation. As part of that work, WP3 has conducted a landscape analysis of European health data collections available for research purposes, evaluating their FAIRness (Findable, Accessible, Interoperable, Reusable) levels. HealthyCloud 's findings point towards inadequacies within the health data systems, cultural barriers, outdated policies and inadequate incentives as some of the main deterrents to the generation and management of FAIR data. [10]

The WP3 of this project performed a landscape analysis to capture the European health data collections available for research purposes, and to evaluate their FAIRness level, and determine the feasibility to perform individual level data linkages. Over 300 data collections belonging to the European Health Information Portal (<https://www.healthinformationportal.eu>) were surveyed, and these surveys found that most data collections were at the regional or national level, with a small percentage (approx. 3%) of data collections having multi-country or European

level data. Over 60% of these collections either do not periodically update their data, or update it annually, and less than 40% allow for linkage with other datasets. [11]

Furthermore, it was found that the application of FAIR data principles to metadata was key to sustainable, efficient and transparent research, increasing the amount of information about datasets (metadata) and facilitating the discovery and request of relevant datasets for research activities. In “Guidelines to standardise metadata templates and assessment of FAIRness maturity levels”, the HealthyCloud project issues recommendations to each part of the FAIR principles [10] (Annex 1).

All of these recommendations were considered in the requirements and specifications of the Portuguese datasets catalogue presented in Section 4.

3.1.3. EHDS2 Pilot

The HealthData@EU EHDS2 Pilot project (<https://ehds2pilot.eu>), led by the French Health Data Hub (HDH) was launched in July 2022 so as to facilitate the implementation of the EHDS for the secondary use of data (EHDS2). In its two-year runtime, the project piloted the EHDS2 by: i) connecting data platforms in a network infrastructure and developing services to support the user journey for research projects using health data from various EU Member States, and ii) providing guidelines for data standards, data quality, data security and data transfer, supporting the EHDS2 infrastructure. To this aim, the project brought 17 European partners together, ensuring representation of health data access bodies, European agencies and health data sharing infrastructures.

One of the key outputs of this project, crucial for the implementation of the national dataset catalogue and EU central catalogue, is the development of a standard for metadata in the HealthData@EU: HealthDCAT-AP. This standard was designed as an extension of the Data Catalogue (DCAT) metadata standard and was purpose-built for the requirements for (re)use of data in the EHDS. In Section 3.2, the details of this standard will be elaborated on.

3.1.4. TEHDAS2 and European Commission’s documents

Building on the recommendations issued by TEHDAS and by the groundwork made in the EHDS2 Pilot project, TEHDAS2 provided specifications and requirements for implementing and maintaining a national metadata catalogue for datasets [2,3] and guidance on how data should be described using HealthDCAT-AP as common metadata model [12]. These will guide the forthcoming implementing acts on datasets catalogues, and were therefore fully considered for the outlining of the requirements and specifications of the Portuguese datasets catalogue.

On further, a comprehensive requirements and specification catalogue for scale-up document setting out the European Commission’s strategic vision for the digital business capabilities essential to the implementation of HealthData@EU, was officially published and endorsed by DG SANTE in December 2024. [13] This document aimed to assist Member States in the execution of their HDAB implementation activities. This document presents a detailed and structured compilation of system requirements across various domains, accompanied by pertinent references to the applicable regulatory frameworks.

The first module of this document, “Dataset Catalogues Requirements”, is structured into three core components: Dataset Creation, Dataset Publication and Update, and Dataset Search Functionality. The analysis of this module played a key role in the development of Section 4 of the present report.

3.2. The European Standard for metadata: HealthDCAT-AP

HealthDCAT-AP was decided as the metadata standard for the EHDS at large, and as such its core (DCAT), and how HealthDCAT-AP expands upon this foundation will be discussed. DCAT is a Resource Description Framework (RDF) vocabulary designed specifically for dataset description, facilitating interoperability between catalogues and decentralized publishing of datasets. First standardized in 2014, DCAT incorporates terms from pre-existing vocabularies whenever their definitions were considered “stable”, further broadening its compatibility.

As a data model, RDF allows one to express information about “resources”, an abstract term that can refer to anything from places, to people, to objects, to concepts, etc. It provides this information in a machine-readable way, to facilitate reading by applications and interconnectivity. One of the key features of this data model is the utilization of semantic triples to construct statements on resources. These triples follow a “subject, predicate, object” structure, any one of these positions can be tied to an international resource identifier such as a Uniform Resource Locator (URL). A vocabulary, such as DCAT, can then be associated to this data model to provide semantic context to each “resource”. [14,15]

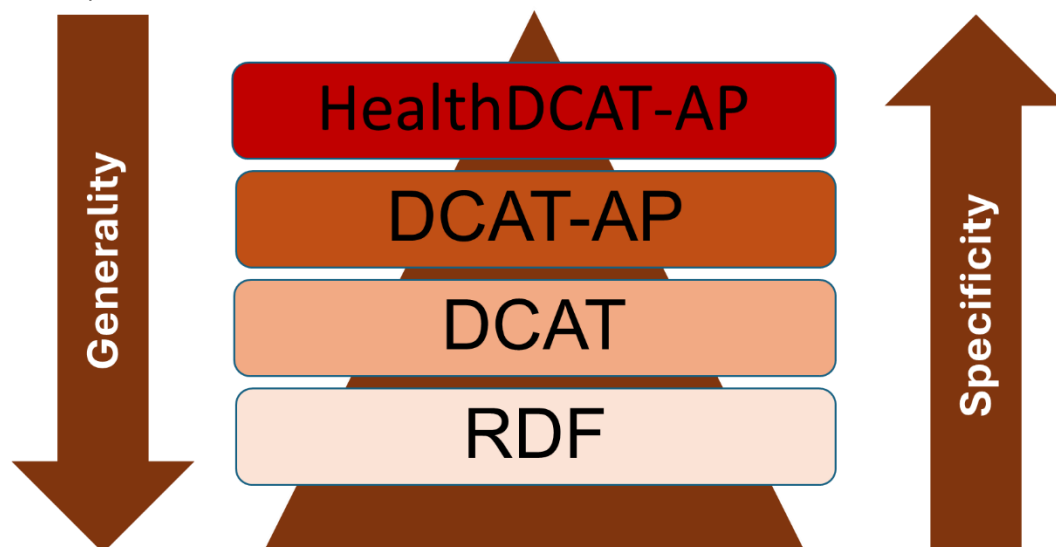


Figure 3. Representation on how the different standards expand on each other: RDF working as the core foundation of the DCAT standard, further extensions increase their specificity.

The DCAT Application Profile (DCAT-AP) is an extension of DCAT, specialized to address catalogue information sharing in Europe. Called Profiles or Application Profiles, extensions like DCAT-AP add constraints to the base DCAT vocabulary, be it the addition of mandatory classes and properties, definition of a minimum set of metadata fields, and so forth. DCAT-AP seeks to increase metadata availability and reuse by standardizing metadata agreements for sharing dataset descriptions between data portals. A key feature of this standard is its expandability, and

extensions have been created to address specific metadata needs, such as GeoDCAT-AP for geospatial data, StatDCAT for statistical data, and HealthDCAT-AP for health data. [16,17]

Creating such profiles entails complying with certain requirements, according to the Semantic Interoperability Community (SEMIC)[18]:

- Respect the minimum conformance requirements as defined in section 4 of the specification of DCAT-AP v3;
- Extensions must not widen but may only narrow down the usage;
- Extensions may add classes and properties that are not specified for and are not similar to existing DCAT-AP classes;
- Mandatory properties must remain mandatory: Recommended and Optional properties may be re-classified or removed;
- Extensions must include all the DCAT-AP mandatory controlled vocabularies present in Section 10.2 of DCAT-AP v3;
- Extensions may add mandatory controlled vocabularies.

With these requirements, one can see how a catalogue compliant with DCAT-AP would be compliant with DCAT, but the reverse is not necessarily true. It is then useful to think about DCAT-AP as a refinement of DCAT that considers documentation classes, properties and other requirements that are focused on metadata availability and reuse. From the section on DCAT-AP's scope, here are some of the aspects considered essential, and as such, are considered in this profile [17]:

- Understanding the data or service structure, and how to get access to the data;
- Information on scope or purpose of the data;
- Legal information;
- Knowledge on data publishers, and any other agents involved;
- Knowledge of data availability and change policies.

After building up the understanding on how these profiles build upon the DCAT vocabulary, now follows an exploration on HealthDCAT-AP as an extension of DCAT, the problems it tries to tackle, and the essential aspects of this profile.

DCAT-AP was chosen as the standard to enhance semantic interoperability within the overall European dataspace and defined as the baseline specification for metadata records in the European Union. As an extension of DCAT-AP, **HealthDCAT-AP** reutilizes its main classes (Catalogue, Dataset, Distribution, etc.), but further refines the model to deal specifically with cataloguing electronic health data. One of the standout features of this vocabulary is the

utilization of triples to allow for metadata enrichment without compromising established structures [19]. HealthDCAT-AP has had its first proposal published in December of 2023, as a response to the proposal for a regulation on the European Health Data Space, and the subsequent need to develop an RDF vocabulary ready to accommodate the unique requirements of metadata about electronic health data.

HealthDCAT-AP was designed from the ground-up to directly tackle the broad scope of EHDS: for example, the inclusion of wellness data from health tracking applications, or statistically and policymaking relevant data such as population-wide health registries come into play in the EHDS, and as such require a viable metadata approach [12].

A GitHub for this standard was deployed, containing a visual representation of the main classes and their main supporting classes, wherein one can see the new and updated properties (in green), and the DCAT-AP properties utilized as is (in red), as present in the excerpt of the mandatory properties of the “Dataset” class in Figure 4.

dcat:Dataset	
"mandatory"	
dct:description:	rdfs:Literal [1..n]
dct:title:	rdfs:Literal [1..n]
dct:identifier:	rdfs:Literal xsd:anyURI [1..n]
dct:spatial:	dct:Location [1..n]
dct:publisher:	foaf:Agent [1..1]
healthdcatap:hdab:	foaf:Agent [1..1]
dcatap:applicableLegislation:	rdfs:Resource [1..n]
dcat:theme (dct:subject):	skos:Concept [1..n]
healthdcatap:healthCategory: (dct:subject):	skos:Concept [1..n]
dct:accessRights:	dct:RightsStatement [1..1]
dct:provenance:	dct:ProvenanceStatement [1..n]
dcat:distribution:	dcat:Distribution [1..n]
adms: sample:	dcat:Distribution [1..n]
healthdcatap:populationCoverage:	rdfs:Literal [1..n]
healthdcatap:minTypicalAge:	rdfs:nonNegativeInteger [1..1]
healthdcatap:maxTypicalAge:	rdfs:nonNegativeInteger [1..1]
healthdcatap:numberofRecords:	rdfs:nonNegativeInteger 1..1]
healthdcatap:numberofUniqueIndividuals:	rdfs:nonNegativeInteger 1..1]
dpv:hasLegalBasis:	dpv:LegalBasis [1..n]
dpv:hasPurpose:	dpv:Purpose [1..n]
dqv:hasQualityAnnotation:	dqv:QualityCertificate [1..n]
dct:type:	skos:Concept [1..1]

Figure 4. HealthDCAT-AP mandatory properties for the "Dataset" class. Excerpt from original in [19].

The properties with the “healthdcatap” affix are novel additions, and cover important information, such as the HDAB supporting the access to data, population coverage as a measurement of representativity, age ranges to facilitate search, number of records as a measurement of dataset size, to name a few. These properties are relevant considering that the health data in these datasets is meant to be archived and searched for in catalogues, so creating mandatory properties that can be research-relevant is key for high quality metadata and to align with FAIR principles. Table 2 contains the main entities (Classes) that comprise the HealthDCAT-AP standard, with short descriptions and hyperlink to their source. In Annex 2, the mandatory

properties that populate these classes (for non-public data, the main data type managed by SPMS) are shown in full.

Table 2. Main entities (Classes) that make up the HealthDCAT-AP standard [19].

Class	Description
Agent	Class of objects that perform actions regarding the “Core” entities (link)
Catalogue	A repository that stores Datasets and supplies other data services. (link)
Catalogue Record	A record of a Dataset’s entry in the Catalogue (link)
Catalogued Resource	Resource catalogued by an agent. (link)
Concept	A “unit of thought” that is the subject of a Catalogue, Dataset or Service (link)
Dataset	Represents information published or curated by an agent (link)
Distribution	How the dataset is made available: its manifestation (link)
Location	Geographical position, a named place (link)
Relationship	Class that defines the connection between DCAT Classes (link)
Rights Statement	A declaration of the intellectual property rights (IPR) associated with or over a resource, an official legal document granting authorisation to use the resource, or an information regarding access rights. (link)

The additions and changes to DCAT-AP made in HealthDCAT-AP can be further framed by the EHDS legislation, where, according to Article 51: “Data holders shall make the following categories of electronic data available for secondary use (...):

- (a) electronic health data from EHRs;
- (b) data on factors impacting on health, including socioeconomic, environmental and behavioural determinants of health;
- (c) aggregated data on healthcare needs, resources allocated to healthcare, the provision of and access to healthcare, healthcare expenditure and financing;
- (d) data on pathogens that impact human health;
- (e) healthcare-related administrative data, including on dispensations, reimbursement claims and reimbursements;
- (f) human genetic, epigenomic and genomic data;
- (g) other human molecular data such as proteomic, transcriptomic, metabolomic, lipidomic and other omic data;
- (h) personal electronic health data automatically generated through medical devices;
- (i) data from wellness applications;

- *(j) data on professional status, and on the specialisation and institution of health professionals involved in the treatment of a natural person;*
- *(k) data from population-based health data registries such as public health registries*
- *(l) data from medical registries and mortality registries;*
- *(m) data from clinical trials, clinical studies, clinical investigations and performance studies subject to Regulation (EU) No 536/2014, Regulation (EU) 2024/1938 of the European Parliament and of the Council(35), Regulation (EU) 2017/745 and Regulation (EU) 2017/746;*
- *(n) other health data from medical devices;*
- *(o) data from registries for medicinal products and medical devices;*
- *(p) data from research cohorts, questionnaires and surveys related to health, after the first publication of the related results;*
- *(q) health data from biobanks and associated databases.*

This legal text leaves plenty of room for interpretation. One of the key challenges identified by TEHDAS2 during their exercise in clarifying these categories to obtain clear, full-text definitions, examples and guidelines to map health data to these categories was precisely the non-triviality of this mapping. Indeed, making the vast wealth of unstructured health data comply with FAIR principles requires not only a processing of that data to make it fit for secondary use, but also a paradigm shift in how approach data production, organization and curation [12].

Besides this effort in classification and clarification of these categories made in TEHDAS2, the EHDS2 Pilot has developed a Sandbox as an example of the HealthDCAT-AP extension in action for a catalogue (Figure 5), containing the ability to search different datasets from the project consortium, although not featuring any of the robust search capabilities seen in the EUCAIM catalogue (Figure 1).

Catalogue: [Cancer Registry examples](#)

['Lithuanian Cancer Registry']

[Acronym:](#) []

[Data type:](#)

['http://publications.europa.eu/resource/authority/dataset-type/STATISTICAL']

[Access rights:](#)

[Applicable legislations:](#) ['https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/6d39efd163b611e688d29c6e5ef0deee?jfwid=18117lpp3']

[Personal data:](#)

[Identifier:](#)

Global resolvable permanent identifiers for metadata records are crucial for effective metadata management as it increases dataset discoverability across multiple catalogues on the Web, and ensures proper citation. (Comment: If the same metadata record is harvested from several sources, dct:identifier helps avoid duplicates, even if the record has been modified by a HDAB.)

[Health Data Access Body:](#)

Name: _g_L49C4526

Homepage:

Email: mailto:_g_L49C4526

[Contact points:](#)

[Creator:](#)

Information provided (19):

dct:title ; dct:type ; dcat:applicableLegislation ; dct:identifier ; healthhealthap:hdab ;
dct:publisher ; health:publisherType ; health:publisherNote ; dct:spatial ;
dct:description ; dpv:hasPurpose ; healthdcatap:minTypicalAge ;
healthdcatap:maxTypicalAge ; dcat:keywords ; dcat:theme ;
healthdcatap:healthCategory ; dct:conformsTo ; healthdcatap:hasCodeValues ;
dct:temporal ;

Mandatory information not provided (4):

dcat:accessRights ; dcat:contactPoint ; dct:provenance ; healthdcatap:healthTheme ;

Recommended information not provided (18):

healthdcatap:populationCoverage ;
p:numberOfRecords ;
individuals ; dct:language ;
dct:accrualPeriodicity ; dcat:temporalResolution ;
source ; dct:isReferencedBy ; dct:relation ;
dcat:landingPage ; dcat:distribution ; adms:sample ; healthhealthap:analytics ;

Optional information not provided (13):

dct:alternative ; dct:creator ; dcat:version ; adms:versionNotes ; dct:issued ;
dct:modified ; healthdcatap:retentionPeriod ; dcat:spatialResolutionInMeters ;
adms:identifier ; dcat:qualifiedRelation ; prov:qualifiedAttribution ;
prov:wasGeneratedBy ; dcat:hasVersion ;

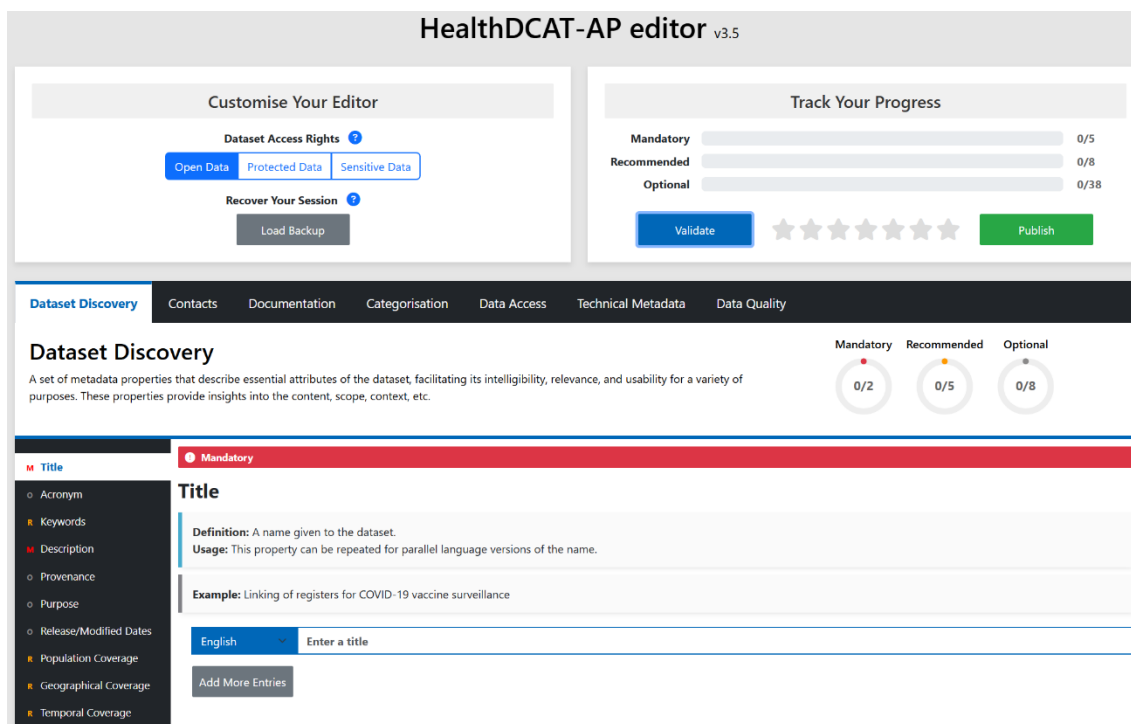
[Metadata issued:](#) ['2024-07-09T13:55:57.040Z']

[Metadata modified:](#) ['2024-07-09T14:06:42.164Z']

Figure 5. Example of a dataset in the EHDS2 Pilot Sandbox. In column at the left, the first rectangle represents the current DCAT-AP property type (M - Mandatory, R-Recommended, O-Optional) when applicable, and the second rectangle represents the HealthDCAT-AP property type.

The preliminary version of this sandbox (Figure 5) presented an exhaustive list on the properties allowing a comparison with the DCAT-AP standard directly (in the rectangles at the left of each property). Hovering over the second rectangle, which features the initial of the property (Optional, Recommended, Mandatory) and an arrow signifying whether the property had been moved up in demand, a tooltip will appear to justify the property's position on the Optional to Mandatory scale.

Recently, a [HealthDCAT-AP editor and validator](#) (Figure 6) were provided to better guide data holders when using this standard for their metadata descriptions.



The screenshot shows the HealthDCAT-AP editor v3.5 interface. At the top, there's a header with the title 'HealthDCAT-AP editor v3.5'. Below this, the interface is divided into two main sections: 'Customise Your Editor' and 'Track Your Progress'.

Customise Your Editor: This section includes 'Dataset Access Rights' with buttons for 'Open Data', 'Protected Data', and 'Sensitive Data'. Below these are 'Recover Your Session' and 'Load Backup' buttons.

Track Your Progress: This section shows progress bars for 'Mandatory' (0/5), 'Recommended' (0/8), and 'Optional' (0/38) properties. It includes a 'Validate' button, a star rating, and a 'Publish' button.

Below these sections is a navigation bar with tabs: 'Dataset Discovery', 'Contacts', 'Documentation', 'Categorisation', 'Data Access', 'Technical Metadata', and 'Data Quality'. The 'Dataset Discovery' tab is active.

Dataset Discovery: This section provides a description of metadata properties and includes progress indicators for 'Mandatory' (0/2), 'Recommended' (0/5), and 'Optional' (0/8) properties.

On the left, there's a sidebar with a list of metadata properties: Title, Acronym, Keywords, Description, Provenance, Purpose, Release/Modified Dates, Population Coverage, Geographical Coverage, and Temporal Coverage. The 'Title' property is selected.

Title: This section shows the definition and usage of the 'Title' property. It includes an example: 'Linking of registers for COVID-19 vaccine surveillance'. There's a text input field with a dropdown menu set to 'English' and a placeholder 'Enter a title'. Below the input field is an 'Add More Entries' button.

Figure 6. HealthDCAT-AP editor, available at <http://ehds.healthdataportal.eu/editor2/> (screenshot from the version at 20/03/2025).

This tool provides key information regarding the process of describing the dataset with HealthDCAT-AP metadata, allowing for data holders to not only track the completion of this task live, but also have examples for each property to fill. Depending on the type of dataset being described (open, protected or sensitive), the Mandatory, Recommended and Optional properties change, and the Editor adapts in real time to this, making sure the data holder is always aware of what fields need to be completed for an adequate description of their datasets.

The [current version of the sandbox](#) (Figure 7) acts as an online backup for all HealthDCAT-AP compliant dataset descriptions published via the Editor tool, no longer featuring the detailed descriptions of properties or the exhaustive list of properties filled and not filled. Instead, the HealthDCAT-AP Record Manager now simply allows to find datasets and exposes all available metadata for them (Figure 8).

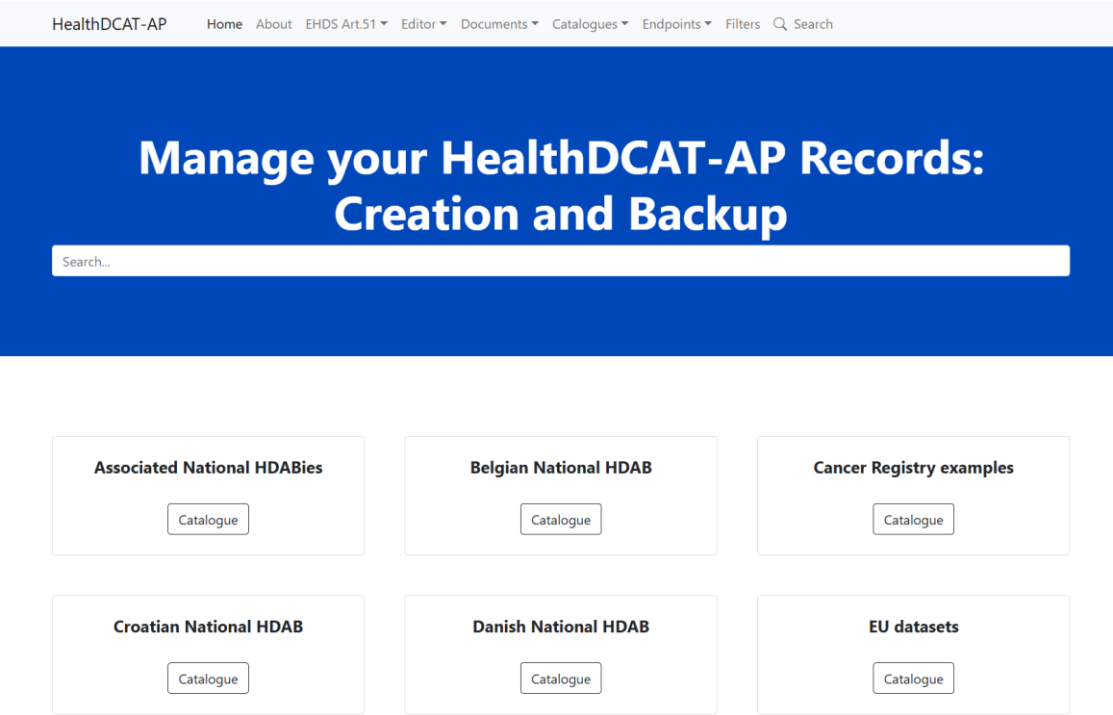


Figure 7. Current version of the HealthDCAT-AP Records Sandbox, now acting as the online backup for all dataset descriptions published via the Editor.

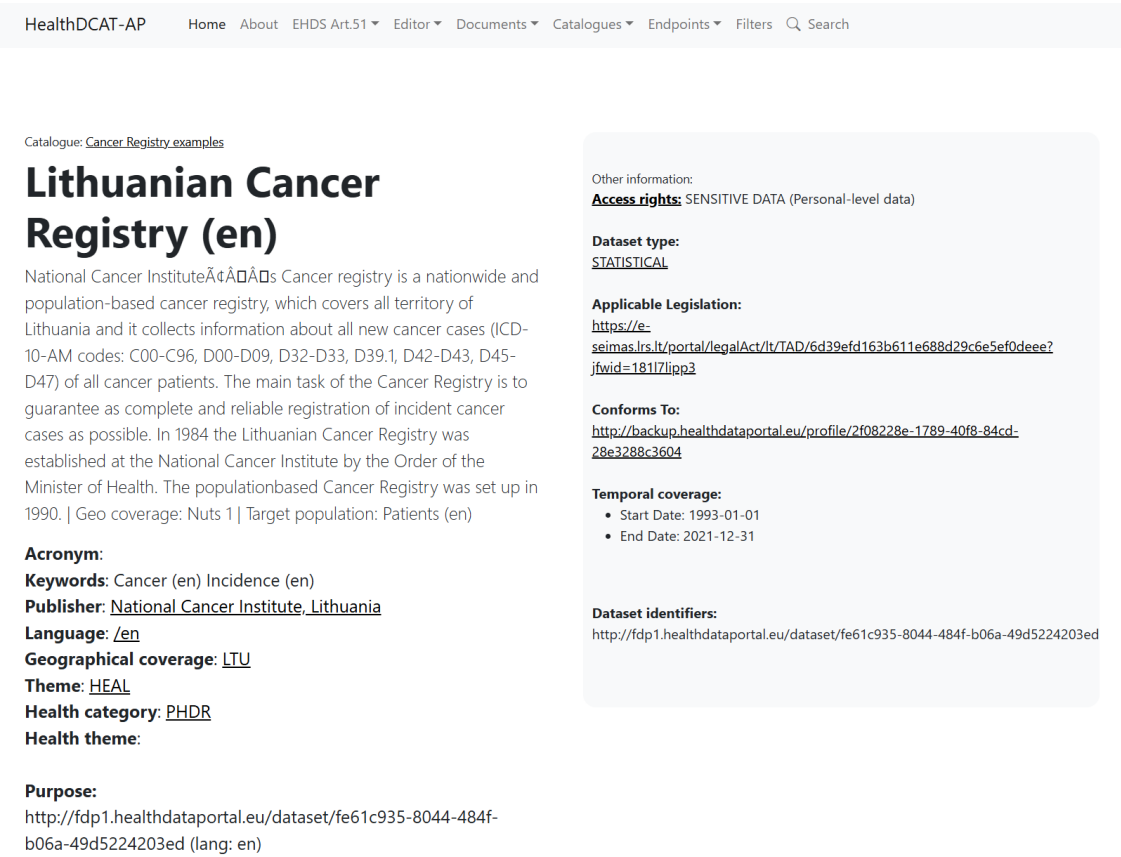


Figure 8. Example of a dataset description available in the HealthDCAT-AP Record Manager.

In a different example of a DCAT-AP extension for health, EUCAIM has been developing its DCAT application profile built upon DCAT-AP and HealthDCAT-AP, to satisfy the metadata describing needs inherent to clinical imaging data. The EUCAIM project is designing a Common Data Model (CDM) and Hyper-ontology to standardize protocols and interoperability standards for federated learning applications in the domain of cancer imaging. Both the CDM and Hyper-ontology rest upon DCAT and DCAT-AP metadata standards, but the EUCAIM Catalogue extends the scope to cover the intricacies of clinical imaging data. In concrete terms, this extension consists of the determination of obligatory metadata for medical images that are part of the EUCAIM public catalogue. EUCAIM made a [mapping of DCAT-AP, HealthDCAT-AP and EUCAIMDCAT-AP classes and properties](#), and allows an easy comparison of the three standards. [4,20] Table 3 contains an excerpt of this comparison, focusing on the differences between HealthDCAT-AP and EUCAIMDCAT for the “Dataset” class, in the same vein as Figure 4.

Table 3. Some property differences between HealthDCAT-AP and EUCAIMDCAT for the "Dataset" class. DCAT-AP comparison provided as baseline.

Property	Property type		
	HealthDCAT-AP	EUCAIMDCAT	DCAT-AP
Contact point	Recommended	Mandatory	Recommended
Geographical coverage	Mandatory	Recommended	Recommended
Modification date	Optional	Recommended	Optional
Sample (distribution)	Mandatory/Recommended	Recommended	Optional
Type	Recommended	Mandatory	Optional
Version	Optional	Mandatory	Optional
Number of series	N/A	Recommended	N/A
Acronym	Optional	Recommended	N/A
HDAB	Mandatory	N/A	N/A

By consulting Table 3, one can observe that EUCAIMDCAT isn't a direct expansion of HealthDCAT-AP, as there are properties with a lower Type in EUCAIMDCAT than in HealthDCAT-AP (Geographical coverage), and properties that exist in HealthDCAT-AP (HDAB) that don't exist in EUCAIMDCAT, and vice-versa. For instance, EUCAIMDCAT includes properties directly related to imaging: modality, vendor, body part, etc. This is yet another example of how the DCAT standard can adapt to the specific metadata needs of a project via the construction of these different application profiles.

3.3. Examples of operational National Dataset Catalogues

Besides taking into consideration the recommendations from the initiatives pointed out in 3.1 and the specifications of the metadata standard adopted for the EHDS, HealthDCAT-AP (explained in 3.2), the outlining of the Portuguese datasets catalogue's (that will be described in sections 4 and 5) was also influenced by some operational datasets catalogues, such the ones described next.

3.3.1. Swedish catalogue

Sweden's metadata catalogue is available [here](#). The homepage of this platform features two key functionalities, which will be described as follows: the data catalogue and an advanced search tool referred to as the Metadata Tool.

Metadata Catalogue

The Metadata Catalogue provides an overview of the content of registers and sub-registers present in the catalogue. Within this catalogue, users can apply filters and compare sub-registers in terms of the temporal coverage of the respective data.

In this catalogue, the list of registers is automatically updated as the filter settings are adjusted. It allows the user to filter the list of records by (Figure 9):

- register name, with the selection being updated dynamically as each letter is typed (in the search bar);
- list of organizations and register maintenance authorities;
- type of register: biobank, researcher-generated data, government register, or national quality register;
- timeframe of the register: Users can input a start date and an end date;
- Option to select whether the register is included in the metadata tool (yes/no).

After applying these filters, the records that meet the selected criteria are depicted. Key information, including the register title, type of register, the temporal coverage of the data within the register, the organization responsible for maintaining the registers, whether it is included in the metadata tool (yes/no), a descriptive summary outlining the register content and main focus, and the titles of all sub-registers contained in it are depicted in the output results of the applied filters (Figure 9).

For example, when selecting a register related to cancer (U-CAN), the user can see that it is classified as a Biobank register, held by the Uppsala Region, with data available from 01-04-2010 until further notice (i.e., ongoing), and that it is included in the metadata tool. The description clarifies that U-CAN is a research program aimed at cancer research, collecting medical and research records, biomolecules, tumour samples, and blood samples from cancer patients. The sub-registers associated with it, such as Breast Cancer, Gynaecological Cancer, Haematological Malignancies, among others, are also listed (Figure 9).

Filter the register list

The list of records is automatically updated when you change the filter settings.

Registry name

U-CAN

Type of register

☐ Biobank
☐ Researcher-generated data
☐ Government register
☐ National quality register

Record-keeping organization

Search by organization

Data in the register

Start date
End date

In the Metadata Tool

☐ Yes

Showing 330 of 330 records.

Reset

U-CAN

Uppsala/Umeå Comprehensive Cancer Consortium

Type of register

Record-keeping organization

Data in the register

Biobank

Uppsala Region

2010-04-01 - until further notice

In the Metadata Tool

Yes

The U-CAN research program is an infrastructure for cancer research. U-CAN collects and structures medical records, questionnaire responses, biomolecules, tumor and blood samples from cancer patients. Information and samples are collected longitudinally (before, during and after cancer treatment) within the following diagnostic areas: colorectal cancer, hematological malignancies and lymphomas, brain tumors, neuroendocrine tumors, prostate cancer, breast cancer, gynecological cancer, lung cancer, esophageal and gastric cancer, cancer in organ transplant recipients, liver and pancreatic cancer, and ear, nose and throat cancer. U-CAN's biobank sample collection is described in RUT.

Sub-register

U-CAN Gynecological cancer - Uppsala Biobank

U-CAN breast cancer - Uppsala Biobank

U-CAN hematological malignancies - Uppsala Biobank

U-CAN brain tumors - Uppsala Biobank

U-CAN lung cancer - Uppsala Biobank

U-CAN neuroendocrine tumors - Uppsala Biobank

U-CAN colorectal cancer - Uppsala Biobank

Figure 9. Example of the information depicted to the user when filtering by Registry name.

Upon selecting this register, users are provided with comprehensive and detailed information. Beyond the previously described, the page includes a direct link to the U-CAN’S official website and features additional sections, such as the primary objective and a detailed breakdown of the associated sub-registers (Figure 10). These sections incorporate interactive elements, such as clickable buttons, enabling users to explore explanatory content about the topics.

Continuing with the previous example, Figure 10 depicts the additional information provided for U-CAN, such as the brief descriptions of the associated sub-registers (titles) and information on the secrecy, which refers to the law or directive governing data confidentiality in the sub-register (Figure 10).



Co-funded by
the European Union

Copyright©2025 HealthData@PT. All rights reserved.

20

U-CAN

Uppsala/Umeå Comprehensive Cancer Consortium

The U-CAN research program is an infrastructure for cancer research. U-CAN collects and structures medical records, questionnaire responses, biomolecules, tumor and blood samples from cancer patients. Information and samples are collected longitudinally (before, during and after cancer treatment) within the following diagnostic areas: colorectal cancer, hematological malignancies and lymphomas, brain tumors, neuroendocrine tumors, prostate cancer, breast cancer, gynecological cancer, lung cancer, esophageal and gastric cancer, cancer in organ transplant recipients, liver and pancreatic cancer, and ear, nose and throat cancer. U-CAN's biobank sample collection is described in RUT.

RECORD-KEEPING ORGANIZATION ⓘ
Uppsala Region

WEBSITE
<https://www.u-can.uu.se/> ⓘ

PRIMARY PURPOSE ⓘ
The purpose of U-CAN is to make samples and data available to researchers and companies with the aim of developing new diagnostic methods, enabling early detection of cancer, and predicting the most effective treatment at an individual level.

TYPE OF REGISTER
Biobank ⓘ

DATA IN THE REGISTER ⓘ
2010-04-01 - until further notice

IN THE METADATA TOOL ⓘ
Yes

Sub-register ⓘ

U-CAN has the following sub-registers:

U-CAN Gynecological cancer - Uppsala Biobank

Prospective longitudinal collection of blood, tissue and information from adult patients with gynecological cancer (ICD C51-58)



Reference period ⓘ
2012-02-01 - until further notice

Secrecy ⓘ
Act (2002:297) on biobanks in healthcare etc.

Figure 10. Example of the information page for a register (U-CAN) in this Swedish catalogue.

Furthermore, on this page, the user has the possibility to select the “Compare Sub-registers” button. When selecting this, the specified sub-registers are systematically added to a comparison cart, enabling the user to subsequently assess whether the selected sub-registers exhibit temporal coverage overlap (i.e., whether they include data from congruent time periods) or if there is no common temporal coverage. In the comparison page, users can also add more sub-registers directly via a search bar (Figure 11).

Register data in time periods

Here you can see if there is collected data within the same time periods for the sub-registers you have selected.

You can add more sub-registers to the list by selecting more sub-registers in the control below or by going back to [the Metadata Catalog](#) page and selecting more there.

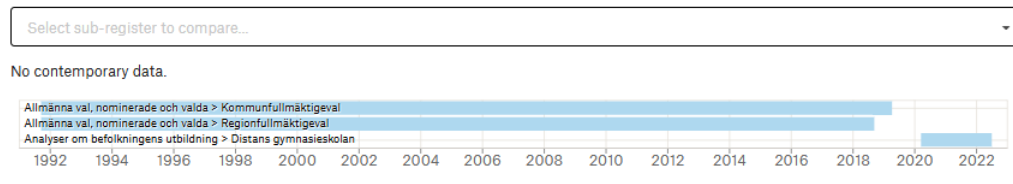


Figure 11. Example of the sub-register comparison tool for temporal coverage in the Swedish catalogue. Three sub-registers were selected, with temporal coverage overlap existing only between two of them. Consequently, a message is displayed to the user indicating that no contemporary data is available.

Metadata tool description

Another feature of the Swedish catalogue is the Metadata Tool. The metadata tool provides a structured overview of the data available in Swedish registers and biobank sample collections. The content of the registers is described using standardized, detailed metadata, allowing for advanced searches and comparisons amongst the various variables. This tool includes information on the significance and value ranges of register variables, among other details, enabling researchers to assess whether the variables meet their needs. Thus, the metadata tool facilitates the identification and evaluation of registers and variables relevant to the user.

Moreover, this tool allows the creation of customized lists of variables from various sub-registers, offering users a direct way to compare only the variables of interest.

Access to the metadata tool requires users to possess an account from the Swedish Research Council. The service is free of charge but requires the email address to be affiliated with a research provider, authority, or register-maintaining organization.

There is also a user manual available for this tool, in Swedish ([Link](#)), to facilitate its use.

3.3.2. Norwegian catalogue

Norwegian's metadata catalogue is available [here](#). The catalogue is available in two languages: Norwegian and English.

The homepage of the catalogue is strategically divided into two sections. The first section is dedicated to enabling users to explore datasets, while the second facilitates requests for data (Figure 12).

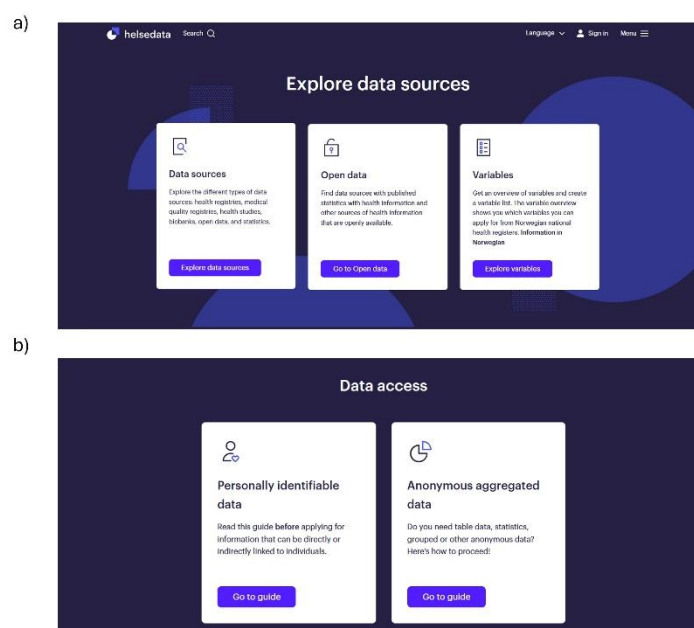


Figure 12. First page of the Norwegian catalogue, divided into two sections featuring tools for exploring data sources (a) and data access guidelines (b).

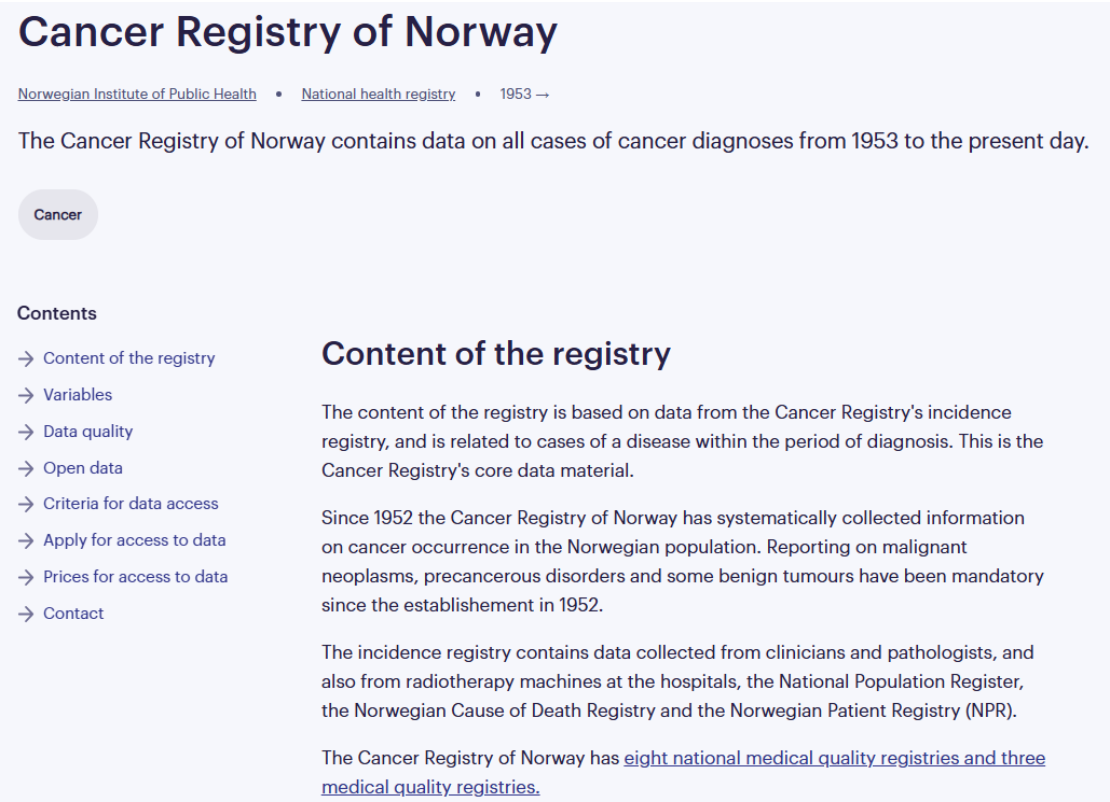
Upon selecting the "Data Sources" option (Figure 12), users are directed to an overview page showcasing various types of data sources, including health records, medical quality registries, biobanks, statistics, among other. This page features a free-text search for term-based location of relevant data sources, supplemented by four key filters:

1. **Type of Data Level:** Identifying whether datasets are open access or require application-based access.
2. **Type of Data Sources:** Differentiating among national medical quality registries, national health registries, studies, statistics, etc.
3. **Categories:** Enabling selection based on specific categories, such as the medical conditions (e.g. autoimmune diseases, cancer, and others), body systems (e.g. gastrointestinal system, respiratory system, musculoskeletal system), socio-economic data, etc.
4. **Data holder:** Filtering by institutions responsible for managing and maintaining the datasets.

Once the filters are applied, the list of data sources on the right is refined accordingly. In these results' list, each result is accompanied by a brief description, name of the data holder and dataset's temporal coverage. Users can click directly on the entries in the results' list to obtain more detailed information.

Within each data source specific page (upon selection), clickable links that direct automatically the user to specific sections are provided. In this page, information about the content of the registry, datasets variables, and data quality are provided, along with information related to the process of applying to data access. In the variables section, there is a clickable button that allows the creation of variables lists and access to detailed metadata. In the data quality section,

information on data quality indicators, such as completeness, validity, comparability, timeliness, and reliability are provided (Figure 13).



Cancer Registry of Norway

Norwegian Institute of Public Health • National health registry • 1953 →

The Cancer Registry of Norway contains data on all cases of cancer diagnoses from 1953 to the present day.

Cancer

Contents

- Content of the registry
- Variables
- Data quality
- Open data
- Criteria for data access
- Apply for access to data
- Prices for access to data
- Contact

Content of the registry

The content of the registry is based on data from the Cancer Registry's incidence registry, and is related to cases of a disease within the period of diagnosis. This is the Cancer Registry's core data material.

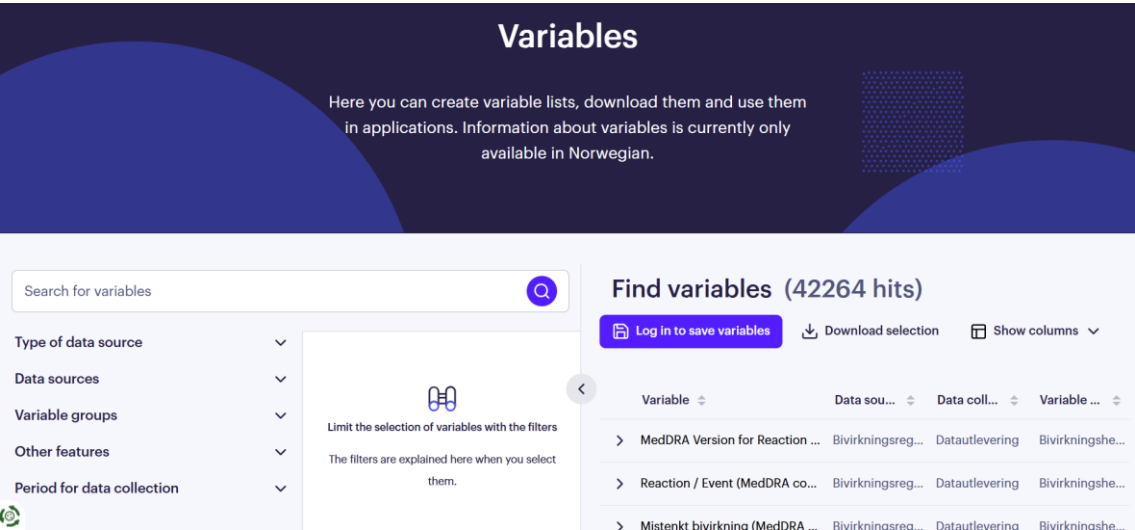
Since 1952 the Cancer Registry of Norway has systematically collected information on cancer occurrence in the Norwegian population. Reporting on malignant neoplasms, precancerous disorders and some benign tumours have been mandatory since the establishment in 1952.

The incidence registry contains data collected from clinicians and pathologists, and also from radiotherapy machines at the hospitals, the National Population Register, the Norwegian Cause of Death Registry and the Norwegian Patient Registry (NPR).

The Cancer Registry of Norway has [eight national medical quality registries and three medical quality registries](#).

Figure 13. Example of the page layout for a dataset in the Norwegian catalogue.

A particularly useful feature is the variable search function, that allows users to search via free text or based on filter's application (e.g. the data source, variable group, etc.) (Figure 14).



Variables

Here you can create variable lists, download them and use them in applications. Information about variables is currently only available in Norwegian.

Search for variables

Type of data source
Data sources
Variable groups
Other features
Period for data collection

Limit the selection of variables with the filters
The filters are explained here when you select them.

Find variables (42264 hits)

Log in to save variables Download selection Show columns

Variable	Data sou...	Data coll...	Variable ...
MedDRA Version for Reaction ...	Bivirkningsreg...	Datautlevering	Bivirkningshe...
Reaction / Event (MedDRA co...	Bivirkningsreg...	Datautlevering	Bivirkningshe...
Mistenkt bivirkning (MedDRA ...	Bivirkningsrea...	Datautlevering	Bivirkningshe...

Figure 14. Example of the variable search page, enabling search using free text or applying different filters (presented on the left). The middle tab is automatically updated according to the chosen filtering type and provides more information about the selected items. On the right, the list of variables, refined according to the applied filters.

For example, when selecting a specific data source, users can explore the related data collections. Each collection provides further information, such as title, description, data source, and other relevant information. There is also detailed content and quality information, including inclusion/exclusion criteria, geographic coverage, completeness and how it's measured, and overall data coverage. Legal and administrative details are also available, including links to relevant legislation and regulations, identification of the data controller and processor, and the level of personal identification involved. Each data collection page includes administrative metadata, such as statistical type, counting unit, and a URL.

Within data collection, users can also explore variable groups (Figure 15), which include titles, descriptions, and metadata update dates.

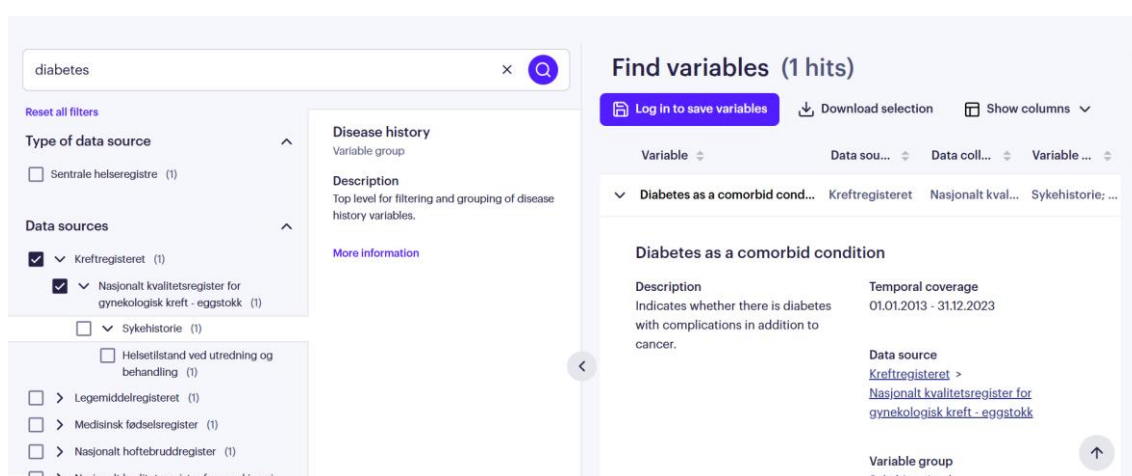


Figure 15. An example of a variable when searching for 'diabetes,' when a data source and a data collection has been chosen. A variable group is selected (but not yet marked) which expands the information in the tab designated for this purpose. The list of variables corresponding to the selected filtering is refined accordingly.

The variable list updates automatically as search criteria are refined, and users can download the final selection in CSV format. The downloaded file includes comprehensive metadata for each variable, such as: URL, Variable ID, Variable Name (in both original language and English), Data Source, Classification Order, Data Collection, Descriptions (original and English), Database Reference, Data Type, Master Code, Desired Data, Measure Type, Unit of Measure, Degree of Identification and temporal coverage.

For each variable, additional details are available, including values and categories, and more specific information such as title, description, origin, content and quality (including notes on data quality, type, and length), and administrative codes and database references.

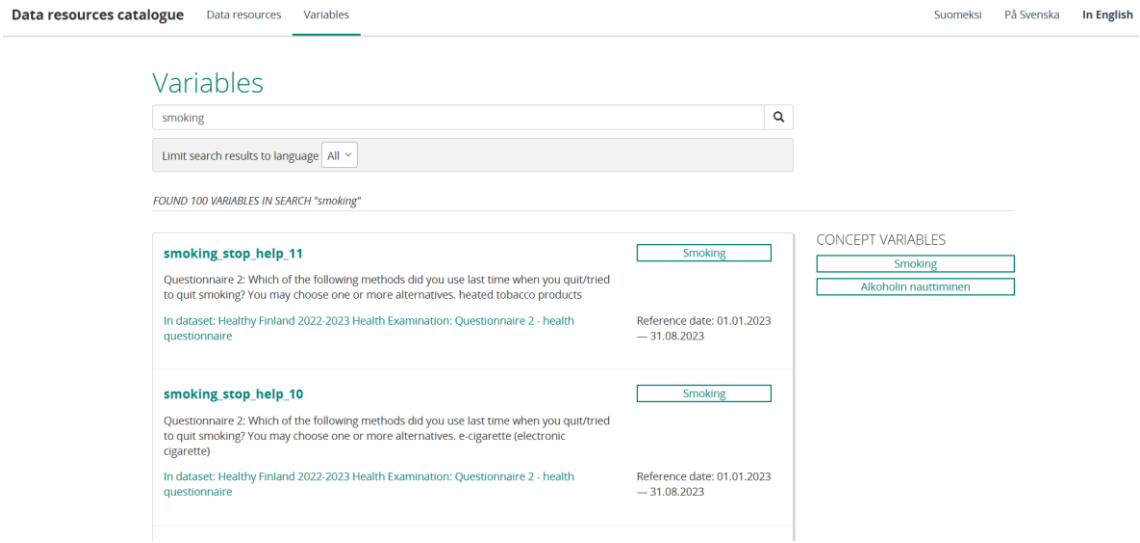
Users also have the option to log in and save selected variables to a personalized list. Additionally, each variable can be shared via a direct link or sent by email.

3.3.3. Finnish (Findata) catalogue

Findata's metadata catalogue is available [here](#). The catalogue is available in three languages: Finnish, Swedish and English.

The homepage of this catalogue is designed with user-centric functionalities, comprising two distinct sections. The first section highlights the most recent data resources, ensuring immediate visibility for updated datasets. The second section focuses on variable search, allowing users to perform queries based on name, description, or keywords.

For example, a search for "smoking" directly navigates users to the variable search page, where more than 100 variables associated with the concept are displayed, including immediate visibility of their corresponding datasets (Figure 16). Furthermore, users can search directly for datasets (Figure 17). By leveraging organizational filters or using the search bar—for example, querying "cancer"—users are redirected to the organization holding the queried dataset (in the case of querying for cancer data resource, for the Finnish Cancer Registry, Figure 17). When accessing the page of the registry, the user is provided with comprehensive information, namely on the content of the registry, its data sources, regional and temporal coverage, keywords, and on the available datasets within the registry.



The screenshot shows the 'Variables' section of the Findata catalogue. At the top, there are navigation links: 'Data resources catalogue', 'Data resources', and 'Variables' (which is highlighted). On the right, there are language options: 'Suomeksi', 'På Svenska', and 'In English'. The main heading is 'Variables'. Below it is a search bar containing the text 'smoking' and a magnifying glass icon. Under the search bar is a dropdown menu labeled 'Limit search results to language' with 'All' selected. Below this, a message states 'FOUND 100 VARIABLES IN SEARCH "smoking"'. The search results are displayed in a list. The first result is 'smoking_stop_help_11', which includes a description: 'Questionnaire 2: Which of the following methods did you use last time when you quit/tried to quit smoking? You may choose one or more alternatives. heated tobacco products'. It also mentions 'In dataset: Healthy Finland 2022-2023 Health Examination: Questionnaire 2 - health questionnaire' and 'Reference date: 01.01.2023 — 31.08.2023'. To the right of the description is a button labeled 'Smoking'. The second result is 'smoking_stop_help_10', with a description: 'Questionnaire 2: Which of the following methods did you use last time when you quit/tried to quit smoking? You may choose one or more alternatives. e-cigarette (electronic cigarette)'. It also mentions 'In dataset: Healthy Finland 2022-2023 Health Examination: Questionnaire 2 - health questionnaire' and 'Reference date: 01.01.2023 — 31.08.2023'. To the right of the description is a button labeled 'Smoking'. On the right side of the search results, there is a section titled 'CONCEPT VARIABLES' with two buttons: 'Smoking' and 'Alkoholin nauttiminen'.

Figure 16. Example of a search for "smoking" in the query for variables within the Finnish national catalogue.

Data resources catalogue Data resources Variables

Suomeksi På Svenska In English

Data resources

Filter by organisation

- Askolan kunta
- Association of Finnish Local and Regional Authorities
- Central Finland Health Care District
- Findata - Social and Health Data Permit Authority
- Finnish Centre for Pensions
- Finnish Education Evaluation Centre
- Finnish Institute for Educational Research
- Finnish Institute for Health and Welfare
- Finnish Medicines Agency Fimea
- Finnish National Agency for Education
- Heinolan kaupunki
- Helsinki Social Services and Health Care Division

Limit search results to language: All

➤ [Finnish Cancer Registry](#)

Figure 17. Example of a search for data resources in the Finnish catalogue.

For example, the Finnish Cancer Registry, contains four datasets, each with a visible count of variables, as can be consulted in its page in the Findata's catalogue (Figure 18). Users can seamlessly click on the dataset of greatest interest, being redirected to the dataset-specific page where information on dataset's populational coverage, observation unit type, dataset's lifecycle phase, and variables can be found.

Finnish Cancer Registry

The Cancer Registry stores information necessary for the to perform its statutory tasks in terms of the recording of cancer, prevention, early detection, treatment and medical rehabilitation, and the use of services. Data are collected for statistical and scientific research purposes.

Persons living in Finland and suffering from cancer and data on the target population of national mass screenings are registered in the Cancer Registry. The data is to be used exclusively for statistical purposes. Cancer information, health insurance information, and information on mass screening for breast and cervical cancer. The information is deleted in the instruction year of national board of health no. 6/1988 (Dns 4/1988) and 15/1988, guidelines and classifications 2004.

Registration of cancers began nationwide in 1963. Since 1961, the submission of cancer reports has been mandatory. The Mass Screening Registry was established in 1963 and has collected service cancer screening data since 1963 and to date cancer screening data since 1963. Cervical cancer screening was introduced nationwide in 2002.

THL is the register of the Cancer Registry. The Finnish Cancer Society is the technical administrator of the Cancer Registry and the provision of patient data.

Statutory register:

- The Act on the National Institute for Health and Welfare (568/2008)
- The Government Decree on the Department of Health and Welfare (675/2008)
- Health Care Act (1303/1992)
- The Government Decree on Screenings (335/2011)
- The Act on the Data Protection (523/2009)
- The Act on the Data Protection (523/2009)
- The Act on the Secondary use of Health and Social Data (552/2019)

Regular sources of information in the register:

- Administrative data on cancer patients and data on migration and immigration are obtained from the Population Information System of the Home and Population Data Service Agency.
- The monitoring data of the Mass Screening Registry and the data of the target population for mass screenings are also obtained from the Population Information System of the Social and Population Data Service Agency.
- Data on socioeconomic status, occupation and education, or well as data on causes of death, are available from Statistics Finland.

Findata's data permission is required to process all the individual level material in the Cancer Registry. Based on the data permit issued by Findata, basic variables, permit specific variables and, at discretion, other variables are provided if the study significantly benefits them.



Finnish population



Historical data:
01.01.1953 –



National coverage

Keywords

cancer, cancerous diseases, registers, screening

Cancer Registry 

Applying for an information permit for research use 

Show less

Datasets

Variables are divided in 4 datasets

Cancer Registry

41
Variable

[GO TO DATASET](#)

Cervical cancer screening

31
Variable

[GO TO DATASET](#)

Breast cancer screening

76
Variable

[GO TO DATASET](#)

Colorectal cancer screening

49
Variable

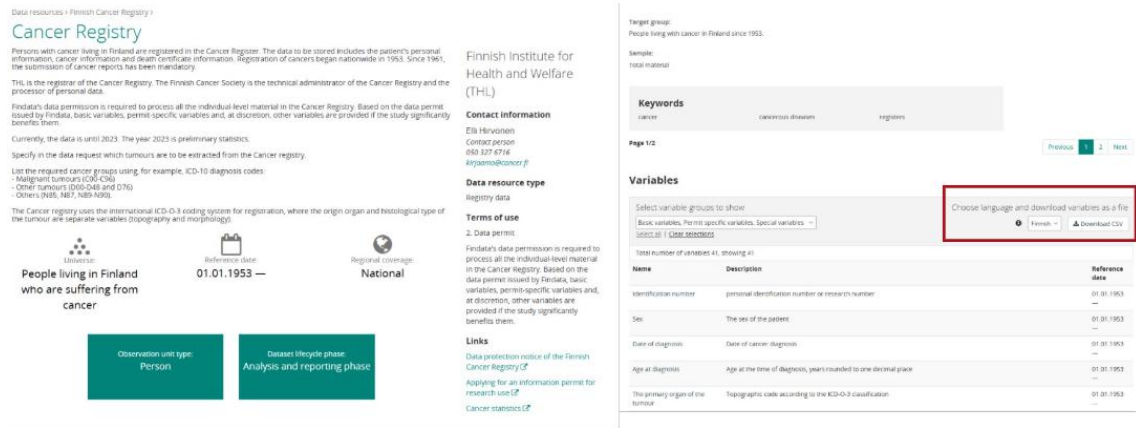
[GO TO DATASET](#)

Figure 18. Example of the page layout for the data source “Finish Cancer Registry”.

Regarding the section with information on the variables, herein variables are categorized into three groups:

1. **Basic Variables:** Such as ID, gender, and diagnosis date, which are accessible without restriction.
2. **Permission-Required Variables:** Including birth date, follow-up details, and patient status.
3. **Advanced Variables:** Requiring specific conditions for access.

Within variables section it is possible to download the information on the variables in a CSV file (Figure 19). Likewise, each variable includes a clickable link that provides detailed metadata on the variable, such as description, and variable data type (e.g. integer) and observation unit type (e.g. person).



The screenshot shows the 'Cancer Registry' page for the Finnish Institute for Health and Welfare (THL). The 'Variables' section is active, displaying a table of variables. A red box highlights the 'Download CSV' button in the top right corner of the variables table.

Name	Description	Reference Data
Identification number	personal identification number or research number	01 31 1953
Sex	The sex of the patient	01 31 1953
Date of diagnosis	Date of cancer diagnosis	01 31 1953
Age at diagnosis	Age at the time of diagnosis, years rounded to one decimal place	01 31 1953
The primary origin of the tumour	Topographic code according to the ICD-O-3 classification	01 31 1953

Figure 19. Example of a page layout for a dataset from the Cancer Registry in the Finnish national catalogue, highlighting the option to download variables in CSV format.

The possibility to download the CSV file, listing the variables and their features, although not described in Findata's page as such, offers an interest possibility for users to indicate the variables of interest in a dataset (for example, by editing the CSV file leaving only the intended variables, and submitting this file in the data access application form), facilitating the subsequent request for data access. Such possibility will be explored while deploying the Portuguese catalogue.

Notably, the Finnish catalogue operates separately from the portal used for submitting data access requests.

3.3.4. French catalogue

The French Health Data Hub is responsible for developing the French metadata catalogue, as part of their effort to become the national HDAB. To this end, in 2024 they published the first Deliverable for this task, containing the technical architecture definition for this digital business capability. This document contains the description of their technical solutions to develop a service that is EHDS compliant, capable of smoothly connecting to the European catalogue while also meeting the national secondary use of health data needs. [21]

Catalogue infrastructure and meta-data standards

According to the French Health Data Hub Deliverable 5.1 "nHDsC: Requirements/Specifications" [21], the French metadata catalogue rests on five main guiding principles:

- **White-label solution:** a generic software solution that can be reutilized and rebranded by other institutions without revealing the underlying Health Data Hub's details;

- **Role management:** capability to assign and manage different administrative roles for access control and governance;
- **Compliance with standards:** follow the HealthDCAT-AP metadata format, for interoperability with European and international metadata standards;
- **Scalable metadata management:** uploading and management of metadata at the catalogue, dataset and database levels;
- **Personalization:** an extension of the white-label solution principle, the solution will be able to be personalized by other institutions to fit their own brand identities.

Regarding the **technical architecture** of this DBC, it is divided into four components: **front-office**, **back-office**, **database**, and **user management**, all of them Dockerized. Both the front-office and back-office have a React webapp front-end, with a Python based back-end (Fast API). Meanwhile, the database was built with the **open-source** solution Fuseki (Apache), chosen for its support for RDF data (key for HealthDCAT-AP compatibility, which is an RDF vocabulary). It also allows for querying via SPARQL, the ability to query the database being a key functionality for the metadata catalogue. Finally, user management uses Keycloak, an open-source solution like Fuseki, to create a Single Sign-On/Identity Access Management (SSO/IAM) system in a GDPR-compliant way.

To **synchronize with national catalogues**, it is planned to employ a solution that makes the local metadata catalogues available through **HTTPS** (Hyper Text Transfer Protocol Secure) and the institution's own SSO or the white-label solution. The white-label solution will possess tools for **pulling** (allowing for its data to be retrieved for synchronization) and **pushing** (using the address of a target catalogue – such as the HDAB's – to send the data for synchronization automatically) the data (in this case, metadata). The workflow is represented in Figure 20.



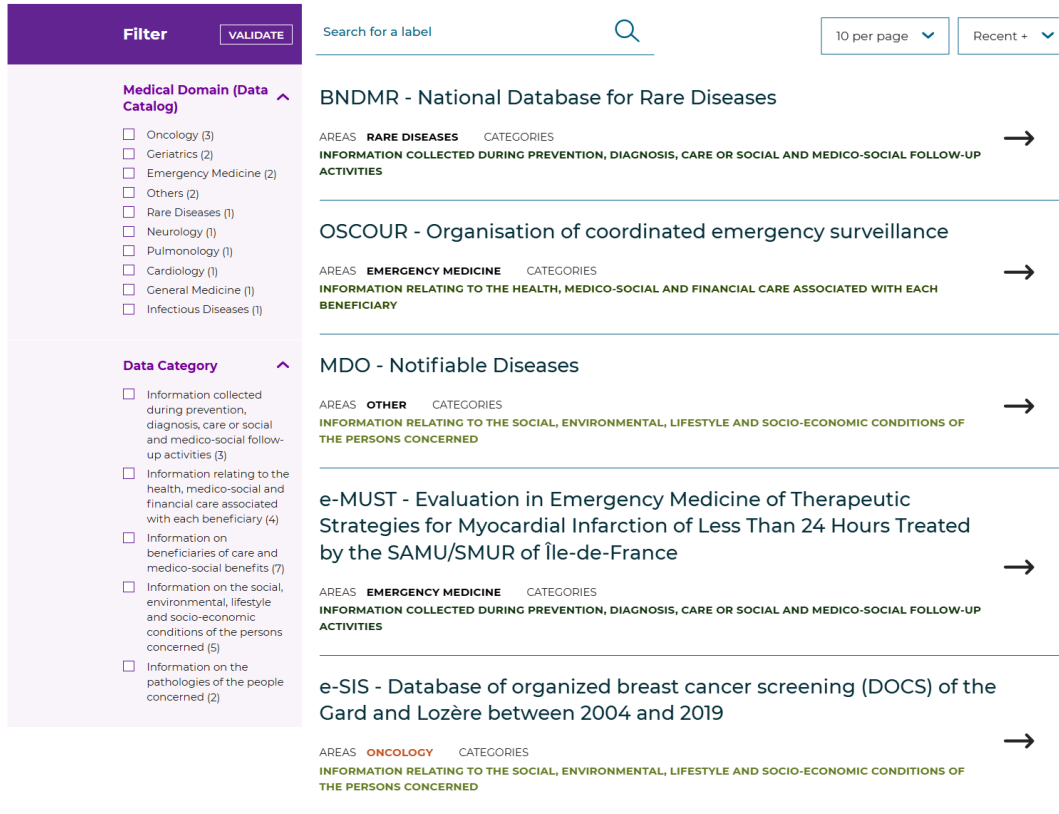
Figure 20. Metadata workflow for metadata synchronization between national catalogues and the nHDsC.

Synchronization with the European catalogue is planned to involve a similar technology, implemented via a Cross Border Engine, pushing the data to the health data dispatcher which is responsible for communication with the cross-border gateway, the component for bridging connection with the HealthData@EU central services catalogue.

Examples from the French Catalogue

France currently has a Data Catalogue (<https://www.health-data-hub.fr/catalog-de-donnees>) that covers datasets from National Health Data System (SNDS), encompassing hospital data,

health insurance data and other sources directly related with the French NHS. Examples of their data catalogue can be seen in Figures 21 and 22.



The screenshot displays the French Data Catalogue search interface. On the left, there are two filter sections: 'Medical Domain (Data Catalog)' and 'Data Category'. The 'Medical Domain' filter includes checkboxes for Oncology (3), Geriatrics (2), Emergency Medicine (2), Others (2), Rare Diseases (1), Neurology (1), Pulmonology (1), Cardiology (1), General Medicine (1), and Infectious Diseases (1). The 'Data Category' filter includes checkboxes for information collected during prevention, diagnosis, care or social and medico-social follow-up activities (3); information relating to the health, medico-social and financial care associated with each beneficiary (4); information on beneficiaries of care and medico-social benefits (7); information on the social, environmental, lifestyle and socio-economic conditions of the persons concerned (5); and information on the pathologies of the people concerned (2). The main search area has a search bar with the placeholder 'Search for a label', a magnifying glass icon, and buttons for '10 per page' and 'Recent +'. Below the search bar, there are five search results, each with a title, a description, and a right-pointing arrow. The results are: BNDMR - National Database for Rare Diseases (AREAS: RARE DISEASES, CATEGORIES: INFORMATION COLLECTED DURING PREVENTION, DIAGNOSIS, CARE OR SOCIAL AND MEDICO-SOCIAL FOLLOW-UP ACTIVITIES); OSCOUR - Organisation of coordinated emergency surveillance (AREAS: EMERGENCY MEDICINE, CATEGORIES: INFORMATION RELATING TO THE HEALTH, MEDICO-SOCIAL AND FINANCIAL CARE ASSOCIATED WITH EACH BENEFICIARY); MDO - Notifiable Diseases (AREAS: OTHER, CATEGORIES: INFORMATION RELATING TO THE SOCIAL, ENVIRONMENTAL, LIFESTYLE AND SOCIO-ECONOMIC CONDITIONS OF THE PERSONS CONCERNED); e-MUST - Evaluation in Emergency Medicine of Therapeutic Strategies for Myocardial Infarction of Less Than 24 Hours Treated by the SAMU/SMUR of Île-de-France (AREAS: EMERGENCY MEDICINE, CATEGORIES: INFORMATION COLLECTED DURING PREVENTION, DIAGNOSIS, CARE OR SOCIAL AND MEDICO-SOCIAL FOLLOW-UP ACTIVITIES); and e-SIS - Database of organized breast cancer screening (DOCS) of the Gard and Lozère between 2004 and 2019 (AREAS: ONCOLOGY, CATEGORIES: INFORMATION RELATING TO THE SOCIAL, ENVIRONMENTAL, LIFESTYLE AND SOCIO-ECONOMIC CONDITIONS OF THE PERSONS CONCERNED).

Figure 21. French Data Catalogue search functions. Existing filters are "medical domain" and "data category".

As can be seen in Figure 21, the search functionalities of the French data catalogue are tailored to the limited entries: free-text search and filtering by medical domain and data category are meant to lead the user towards a database, not specific datasets. Despite this, a lot of important fields in for the HealthDCAT-AP standard. Information regarding the provenance, data controller, temporal and geographical coverage, regulatory framework, data retention period and more are present when consulting the registry for each database in the data catalogue. The host website also allows for data users to submit their data access requests.

Data characteristics

Initial base objective

1. The essential
2. **Data characteristics**
3. Basic documentation
4. Quality
5. Data Protection Officer
6. Regulatory

In 2005, the Ministry of Health launched a national Rare Disease plan in order to better structure the management of these diseases, which although affecting a minority of people represent a health risk. This plan led to the creation of expert sectors and centers and the establishment of nomenclature to be able to identify and monitor these rare diseases, as long as patients are followed in the expert centers. The BNDMR is the national health data warehouse for these diseases. In June 2021, it brings together data from more than 800,000 patients suffering from 4,800 diseases and recorded in 80 health establishments.

Medical field

RARE DISEASES

Population of interest

Patients consulting in rare disease centers of expertise certified by the Ministry of Health and Solidarity (including fetuses).

Target workforce

800,000 patients as of June 2021

Geography

People supported in the network certified by the General Directorate of Care Supply for expertise in the management of rare diseases

Figure 22. French Data Catalogue, example of a dataset registry. Information includes data controller and scientific manager, population of interest, dataset size, temporal coverage, provenance, database documentation, etc.

As such, it is possible to see where Health Data Hub, as one of the members of the EHDS2 Pilot consortium, started considering the necessary aspects of a datasets catalogue, and the metadata standards required for such an infrastructure to function.

3.4. Fair Data Points

In each Member State, the national datasets catalogue of the national HDAB will expose and aggregate in a unique catalogue metadata of the datasets available for secondary use purposes in that Member State. [3]

For this purpose, each data holder will publish and expose metadata of the datasets owned via FAIR Data Points (FDPs), which are interoperable services aligned with FAIR principles (Findable, Accessible, Interoperable, Reusable). A FDP is a REST API for creating, storing, and serving metadata with a customizable metadata repository. FDPs provide a uniform method for publishing metadata about datasets, catalogues, and their distributions. This technology addresses key requirements for metadata selection and exposure for ensuring that metadata is both machine and human readable. Reverse Proxy, FAIR Data Point Client, FAIR Data Point Server, and a Triple Store/MongoDB for metadata and user data, respectively, are commonly found in an FDP's architecture. [22,23]

This enables automated metadata information harvesting to the portal in a standardized and scalable manner. In the Section 4, the metadata workflow strategy from different data holders to the national datasets catalogue of the portuguese HDAB will be addressed.

3.5. Current Portuguese Context

Portugal does not yet possess an nHDS: that is the purview of this Direct Grant. However, extensive work has been done in the fields of interoperability and standardization of electronic health data, as well as efforts in creating an Open Data platform for data access.

3.5.1. The Portuguese NHS Transparency Portal

The Portuguese [NHS Transparency Portal](#) is a platform that aims to provide visibility and access to the vast data generated through the NHS' operation, encouraging and facilitating health data reuse. Listing 155 datasets of anonymised and aggregated data sourced from 13 data holders, the datasets can be searched and accessed via the [Transparency Portal Catalogue](#) (Figure 23), which allows for plain text searching and filtering by different categories (themes, keywords, publisher, date of modification. etc.).

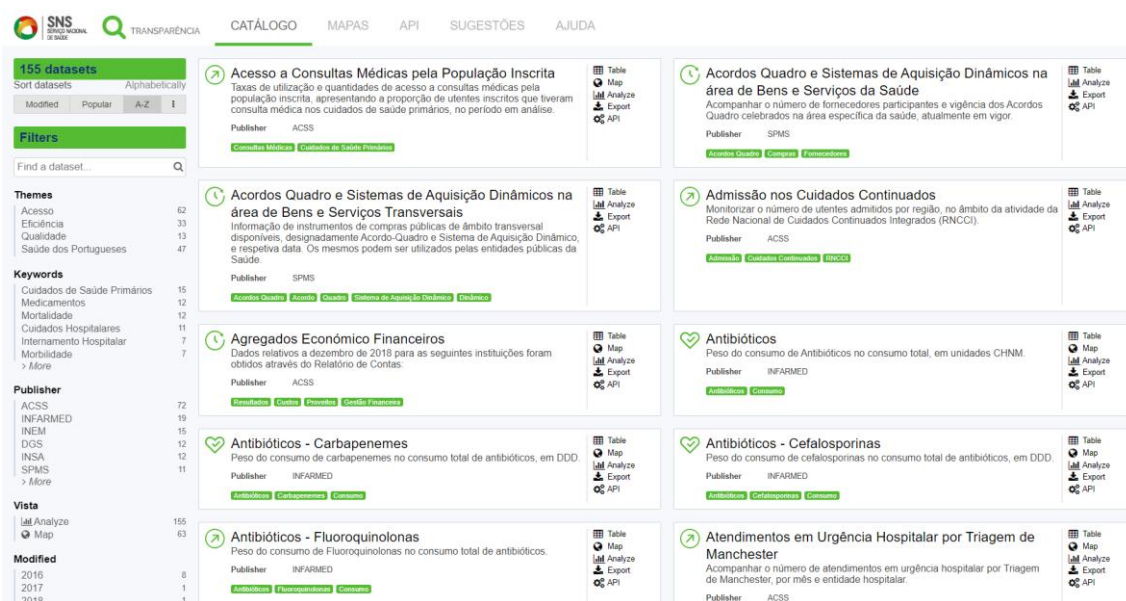


Figure 23. Portuguese NHS Transparency Portal Catalogue.

These datasets, generated by the NHS' different institutions, can be opened to perform various actions (Figure 24 and Figure 25), such as visualising information regarding the dataset, observing the data in tabular form, observe geographical distribution in a map view (only certain datasets), generate and analyse graphs with different parameters, feed the dataset to an API to search and download records according to certain parameters, and download the dataset in different formats. Furthermore, the dataset itself can have filters applied to it, narrowing the dataset by time, hospital and/or hospital group. Thus, **the transparency portal aligns with the EHDS article 60**, on the duty of providing access to non-personal health data through public, open and reliable databases.

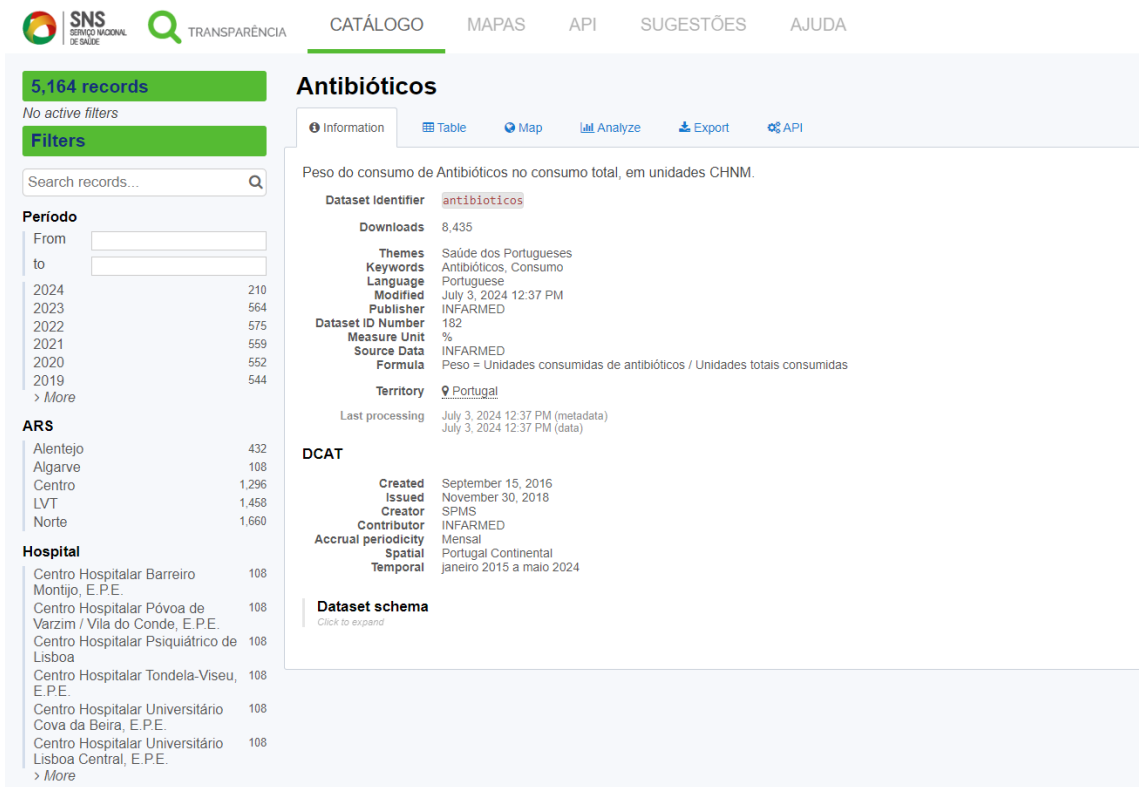


Figure 24. Portuguese NHS Transparency Portal Catalogue - Antibiotics dataset.

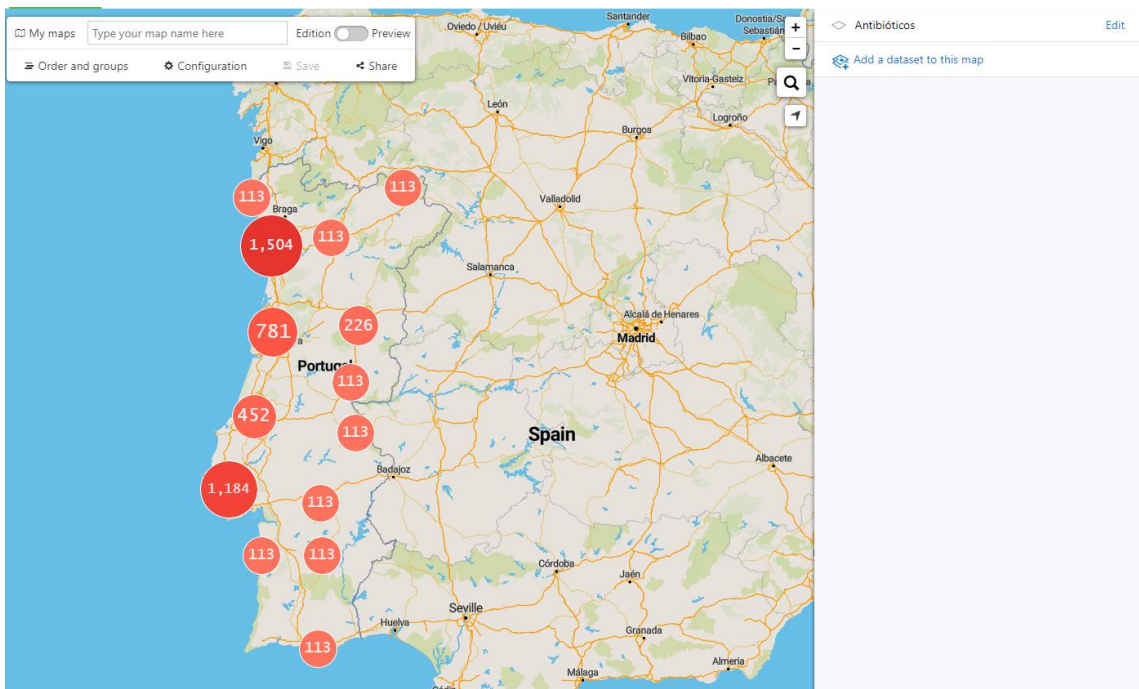


Figure 25. Portuguese NHS Transparency Portal - User generated map with a dataset overlay for antibiotic consumption.

All metadata published in this catalogue utilizes the DCAT standard, following in the steps of its EU-wide adoption, and in preparation for the mandatory adoption of the HealthDCAT-AP metadata standard for electronic health data. To encourage and facilitate submission and updating of the datasets regardless of the data holder's infrastructure, data integration can be

done manually (file upload), semi-automatically (Extract, Transform, Load with manual pre-processing), and automatically (via webservices). Current developments in this strategy seek to fully automate data integration via webservices. Considering that not all datasets will follow the same metadata standards, the system for creating metadata for the NHS Transparency Portal Catalogue involves the data holders in such a way to reinforce their cooperation and advise in the creation of metadata according to DCAT (Figure 26).

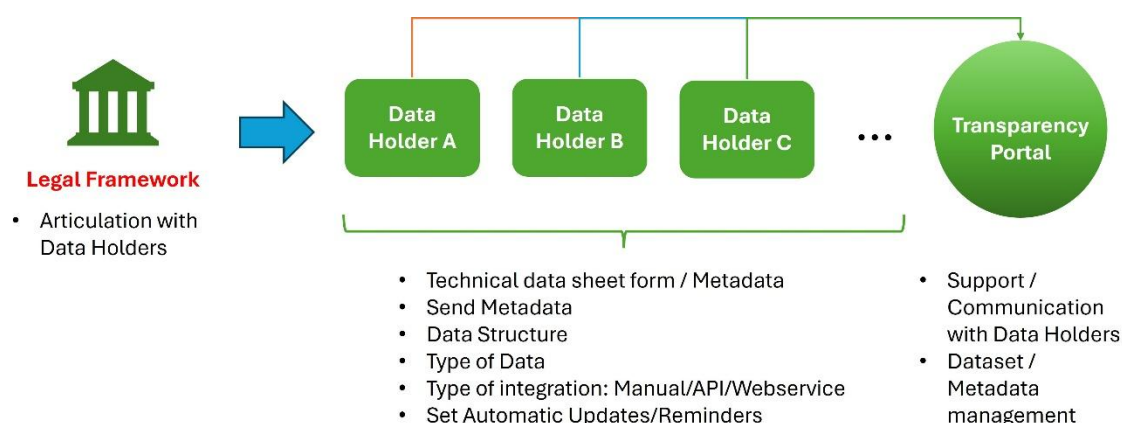


Figure 26. Portuguese NHS Transparency Portal Catalogue Datasets' metadata generation flowchart.

With the work already performed in this infrastructure, a solid foundation upon which the national health dataset catalogue (nHDsC) can be built has been established: containing a system for data ingestion and metadata enrichment following the DCAT standard, user-centric catalogue searching and data analysis tools resting on the Open Data principle, this Portal already facilitates health data reutilization for news generation and research purposes.

3.5.2. Other national developments

Portuguese National Catalogue for Clinical Analysis (CPAL)

The Portuguese National Catalogue for Clinical Analysis (CPAL), first released for public consultation in 2016, is currently in version 7, and provides a normalized structure to register Clinical Pathology laboratory results. It adopts LOINC and SNOMED-CT identifiers to ensure semantic interoperability. While not a datasets catalogue in the fashion observed in previous examples, this catalogue is **an example of ongoing national efforts in health data interoperability**.

The Clinical Terminologies Centre (<https://www.ctc.min-saude.pt>) has published several other catalogues that seek to standardize health data registry (Figure 27) and will be collaborating in this project for the implementation of our national DBCs.

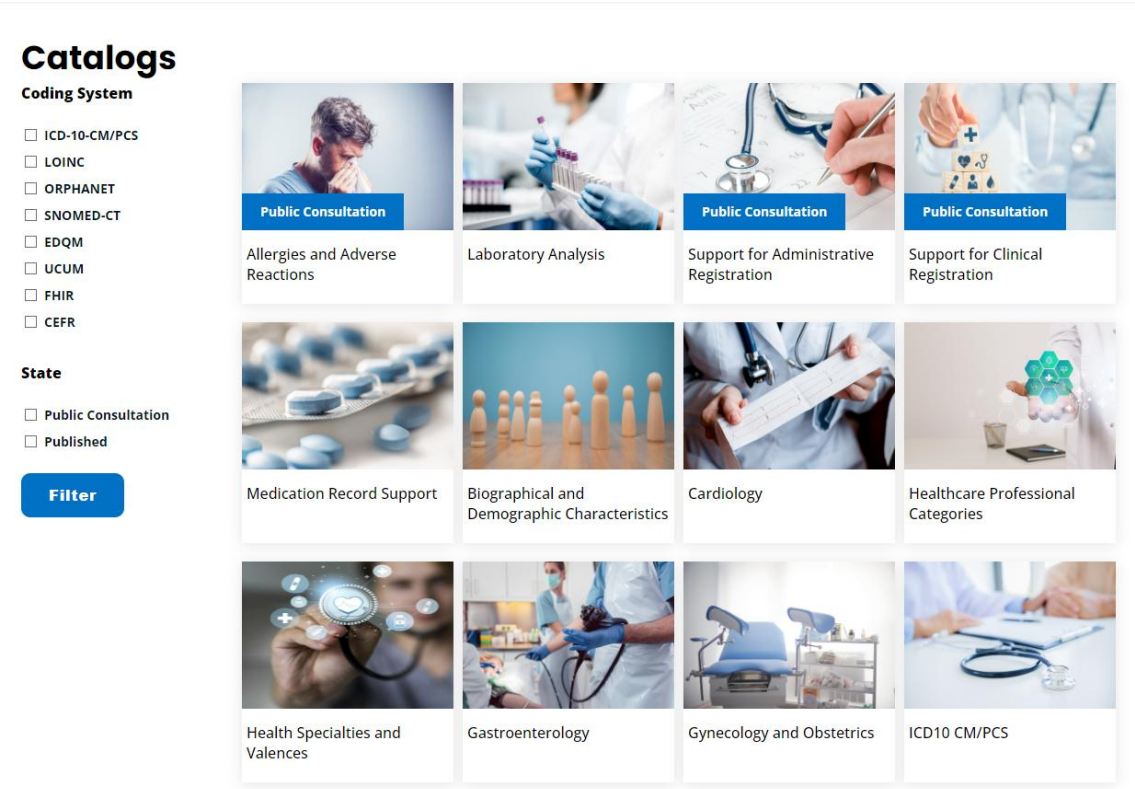


Figure 27. Some of the Clinical Terminologies Centre’s Catalogues.

Efforts made in the field of interoperability have a direct impact in the primary use of data, by making data easily interpretable and compatible across healthcare providers, and also benefit secondary use.

4. Requirements and Specifications for the Portuguese National Datasets Catalogue

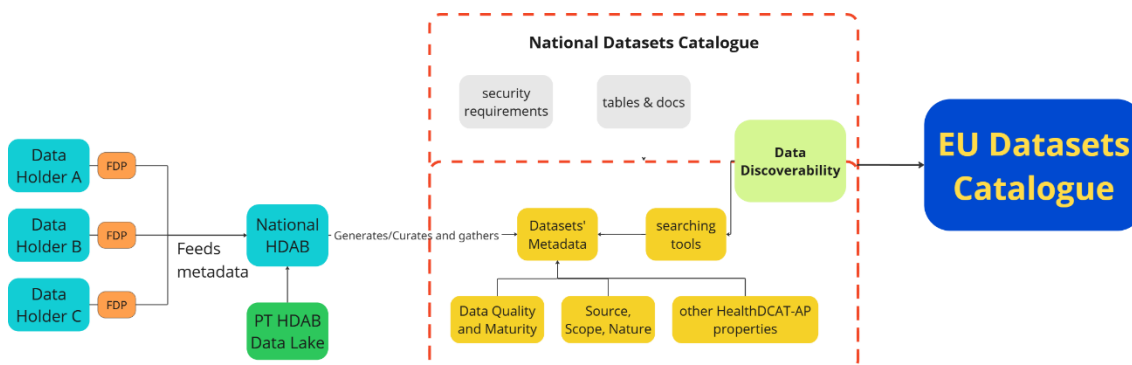


Figure 28. Architecture of the metadata flow, with particular emphasis on the role of the National Datasets Catalogue, from data holders to the European Datasets Catalogue.

In Figure 28, the diagram illustrates the overarching metadata flow architecture from data holders to the National Datasets Catalogue and from this last to the HealthData@EU (EU) Datasets Catalogue. Each data holder publishes and exposes metadata via FAIR Data Points (FDPs). These FDPs feed metadata into the National HDAB, which centralizes and aggregates metadata into the National Datasets Catalogue.

The National Datasets Catalogue includes detailed metadata elements such as those pertaining to dataset quality and utility, source and scope, and other HealthDCAT-AP-compliant properties. It also supports advanced search tools and adheres to defined security requirements and documentation standards. The data discoverability component enables seamless publication of metadata from the national catalogue to the EU Datasets Catalogue, promoting cross-border data findability and accessibility within the HealthData@EU infrastructure.

Next, the outlined requirements and specifications for the datasets catalogue under development during HealthData@PT project are presented.

4.1. A public national portal holding the national datasets catalogue

- A **public metadata catalogue** is a EHDS requirements to facilitate access to health-related datasets intended for secondary use. The Portuguese public catalogue is currently under development (Section 5.3) and will be accessible through a dedicated public portal, designed to provide an intuitive interface for exploring metadata records. This platform will adopt the **CKAN framework**—an open-source solution widely implemented in public data catalogues across Europe—and will comply with all relevant national and European legal and regulatory requirements.
- The catalogue will provide an intuitive user experience, fully aligned with the established design system for the graphical presentation layer. The platform will **support bilingual access** (Portuguese and English) and will ensure full compliance with the Web Content Accessibility Guidelines (WCAG) 2.0, meeting, at a minimum, the Level AA criteria for web content accessibility.

- The catalogue will initially include **25 datasets**, with the first 12 being currently under development (as detailed in Section 5.1). All datasets will be described complying with the **HealthDCAT-AP metadata standard** specification, providing dataset descriptions in a standardised, machine-readable format. This ensures interoperability and alignment with international best practices for metadata documentation. Each dataset will also include detailed descriptions of associated variables, as specified in 4.2 Datasets description – HealthDCAT-AP standard and variables.
- The platform is designed as a **publicly accessible electronic tool to dataset descriptions** (as read only); no login will be required to browse or search within the catalogue.
- Users will be provided with **search functionalities** for browsing and displaying dataset catalogue results. Users will be able to perform free-text searches, as well as to apply **structured filters**, consistent with those available in the HealthData@EU central services Portal. These filters will include age range, format, health themes, data model, and additional parameters found in the [Portuguese Transparency Portal](#), such as year of birth, year of death, temporal coverage, keywords, and provenance.
- A **variable-based search tool** will also be implemented, drawing inspiration from advanced functionalities available in the Norwegian health data catalogue (described in section 3). This will enable users to identify datasets based on specific variables of interest.
- The portal will also include a **contact section** where users can request further information or support
- The catalogue public portal will provide a **"basket" functionality** to allow users to compile/save the searched datasets of interest. This will be useful, for example, for later submissions of data access applications/data requests (in a separate portal and upon authentication²).
- Future enhancements to the platform will include the integration of large language model (LLM) prompt engine and machine learning capabilities, aimed at improving search relevance and user experience. A dedicated tool for comparing the temporal coverage of selected datasets will also be developed, inspired by features available in the Swedish health data portal (described in section 3). This tool will enable users to visually assess and compare the time periods covered by different datasets.
- Although the catalogue will remain publicly accessible for the purposes of browsing and information retrieval, **authenticated access will be mandatory for data holders to submit or edit metadata records**. This process will comply with established security protocols, including the implementation of **multi-factor authentication (MFA)**. The ability for data holders to directly input metadata will be also developed in future versions and will also introduce an accreditation system, integrated with the ORACLE Identity and Access Management (IAM) infrastructure, to guarantee that only duly

² Further details on the portal for data access applications and requests in HealthData@PT D5.1.

authorized institutional representatives may contribute or update dataset information. The technical specifications for this authentication framework are still under development and are expected to incorporate two independent verification methods. During the initial phase of project execution, the role of the data holder will be provisionally assumed by the catalogue manager. This individual will act as an intermediary, responsible for facilitating the submission of metadata on behalf of external data-providing entities. In subsequent phases, this responsibility will be transferred to the data holders themselves; however, **all submitted datasets metadata will be subject to internal validation** prior to publication, to ensure that dataset description include the minimum elements required.

- **For security and compliance purposes, the catalogue will not be directly connected to any data storage systems.** Likewise, it will never expose actual data to allow any description that could lead to the re-identification of individuals. All dataset samples (HealthDCAT-AP mandatory property for non-public data) to be provided for non-public data will be based on synthetic data. The following dedicated security mechanism will be in place to assess the safety of synthetic dataset generation:
 - Perform re-identification tests;
 - Synthetic data will not be generated with high level of detail.
- In catalogue's development, **tamper-proof links and immutable publication controls** will be implemented. To ensure this, the following security measures will be embedded by design:
 - Configure the Content Security Policy (CSP);
 - Configure the HTTP Headers.

These security measures must ensure that **once content is published, it cannot be altered by general users**. These safeguards ensure that no public element of the catalogue can be redirected, modified or exploited for malicious purposes.

- Other **security measures** will be included, such as:
 - Applying strict access control processes, that implement both the concept of segregation of duties and principle of least privilege;
 - All security events (Authentication, Authorization, etc) will be logged and stored for compliance and forensic analysis.

Logs generated will be integrated in a Security Information and Event Management (SIEM), to proactively identify suspicious events that may be cybersecurity incidents. In case of any suspicious events, these will be communicated to the Computer Security Incident Response Team (CSIRT) through this email address: csirt@spms.min-saude.pt.

4.2. Datasets description – HealthDCAT-AP standard and variables

- Each dataset description will be assigned a **unique identifier** and include the minimum elements specified in the HealthDCAT-AP standard (Annex 2). These descriptions will be made findable and discoverable, and will be **persistently stored, versioned, and maintained** to ensure accuracy, completeness, and compliance with EHDS requirements.
- All metadata will follow the HealthDCAT-AP standard, adhering to the standard's mandatory properties (Annex 2), as well as making mandatory any other properties deemed necessary to tackle the national health data reuse needs. Each dataset description encompasses at least the minimum elements that health data holders are required to provide for each dataset, along with the detailed characteristics of those elements.
 - To that end, the catalogue will facilitate the creation of metadata records following the HealthDCAT-AP standard, via access to appropriate tools.
- A data quality and utility label will be included in accordance with EHDS regulation.
 - This label will appear as mandatory metadata for all datasets created with public and/or European funding
- Variables will be included in the form of a data dictionary, following the HealthDCAT-AP Property "Sample Distribution".
 - This variable list will feed the "Personal Data" Property.
 - The data dictionary will also be presented in an easy to visualize format (i.e. an expandable box where the data applicant can click to visualise all dataset variables listed)
- The nationally published dataset descriptions will be **synchronized with the EU dataset catalogue** via national contact point.

4.3. Datasets creation, maintenance, storage, review and auditability

- The catalogue will allow for the creation, reading, updating and deletion (**CRUD**) of dataset descriptions in a **standardised** and machine-readable format, HealthDCAT-AP.
 - The dataset descriptions present in the national catalogue will be **updated yearly at minimum**.
 - The update frequency of the opt-out information will align with that of the remaining metadata. The catalogue will include a disclaimer informing that due to opt-out mechanisms, the dataset metadata at the time of data access application/data request may be slightly different.
 - An automated solution must detect whenever alterations in the national catalogue (creation, update, deletion, restore) occur, towards synchronization with the EU dataset catalogue.
- The catalogue will ensure the **persistent storage** of dataset descriptions.
- A mechanism will be in place for the validation of dataset descriptions through the use of appropriate tools such as the HealthDCAT-AP Editor.

- The catalogue will include versioning tools, so that alterations to the metadata can be traced, for auditability and troubleshooting.
 - All audits will be logged, for traceability and accountability, using the Red Hat Advanced Cluster Security for Kubernetes.

4.4. Catalogue Backend

- In the backend, **relational tables** associated with each dataset will be maintained (excluding the actual data content) and will be accessible together with relevant documentation to authorized and adequately authenticated catalogue (HDAB) staff. This **supporting information** stored in the catalogue backend includes documentation related to data quality and utility labels, in accordance with the QUANTUM framework, and metadata on the location of the source data to build the dataset. A standardized methodology for dataset table construction will be adopted, to ensure consistency, promote traceability, and facilitate interoperability between systems and services.
- Additionally, the backend will include a **management tool** to manage the incoming messages with metadata records (standardised metadata ingestion) from external data holders. For instance, when a message is received containing updated descriptions for a dataset collection, this screening mechanism will push the message and perform an initial validation to verify whether the metadata conforms to the required standards for publication.
- **Catalogue administrator profiles** will be in place and will require authentication (login); **all operations will be logged** to keep track of changes made to dataset descriptions for **auditability** purposes. A catalogue manager will be responsible for the publication, maintenance and quality assurance of datasets within the National Dataset Catalogue. This role includes coordinating with data holders to prepare and validate metadata, ensuring alignment with national and European standards (e.g., Health DCAT-AP) and appropriate data quality and utility labels. The catalogue manager plays a key role in enhancing the discoverability, interoperability and trustworthiness of datasets available for reuse.

Examples of catalogue administrator profiles:

- Data Governance or Catalogue Operations Unit
- Designated curators or technical staff supporting publication workflows
- An interface for managing and reviewing dataset descriptions will be made available to catalogue administrators via the catalogue backend.
- The catalogue will be entirely decoupled from data storage systems and reference tables, ensuring a secure and controlled environment. Only metadata and descriptive information that cannot be used to identify individuals will be exposed in the public catalogue.
- The catalogue will also report **statistical information** related to dataset access and usage as part of their biennial activity reports.

4.5. API for the data access application/data request form portal

While navigating the catalogue, potential applicants will be able to select the datasets they intend to request using a dedicated selection button. As mentioned in 4.1, the datasets of interest will be saved on a basket (similar to an e-commerce shopping cart), and this selection is stored via session cookies. In the infrastructure under development, the data access application/data request forms will be held in another portal that will operate independently from the catalogue's. Nevertheless, both portals are iframe connected. A robust security framework will ensure that the relevant information from the selected datasets will migrate from the catalogue basket to the data access application/ data request platform. At least, migration of the dataset ID will be ensured.

Upon logging into the private area of the portal for requesting access to data/data, the metadata on the user's chosen datasets will be automatically imported via an API into the data access application/data request form, pre-filling the corresponding form fields, streamlining the application process and avoiding errors related with manual inputs.

4.6. Metadata ingestion

- A dedicated workflow will be implemented to manage the ingestion of metadata, receiving dataset descriptions from data holders through FDPs using **standardized formats and interfaces**. Furthermore, **annual maintenance cycles** and an **automated notification mechanism** (e.g., system clock alerts) will be established to prompt timely updates and notifications and to ensure ongoing metadata quality and compliance with EHDS requirements on metadata updates.
- An API will be implemented to facilitate the importing of metadata from external entities, ensuring seamless integration with third-party systems and keeping the catalogue's metadata up-to-date and comprehensive.
- Data holders will always be notified in cases of approval, publication, or revocation of their datasets.
- A secure messaging protocol will be established to allow the exchange of metadata and other relevant information between the data holders and the catalogue administrators. This mechanism will guarantee the integrity and security of the messages, enabling efficient communication and updates.
- As mentioned in 4.3, mechanisms will be in place to validate the compliance of dataset descriptions with HealthDCAT-AP requirements.
- To ensure the legitimacy of the data holders sending messages and metadata, a secure identity verification process will be in place (e.g. using an accredited institutional e-mail).
- Data holders will be notified of the update requirements and guided through a standardized process for submitting the updated metadata, building on existing experience with the Portuguese NHS Transparency Portal described in 3.5.1.

4.7. Catalogue API for synchronisation with HealthData@EU datasets catalogue upon CRUD operations

- The **cross-border engine** (CBE) is one of the national components within the HDAB framework, bridging the national datasets catalogue with the datasets catalogue from the European infrastructure HealthData@EU. The CBE will automatically detect when an alteration occurs in the catalogue (for example, an addition of a new dataset or an update in an existing dataset in the national datasets catalogue) and will subsequently invoke the corresponding health data dispatcher APIs³. The health data dispatcher will then prepare the message communicating these changes, that will be transmitted to the central European platform (HealthData@EU) via eDelivery, through the Cross-border Gateway, for these alterations in the national catalogue to be reflected on the HealthData@EU datasets catalogue.
- After the message informing on the alterations on the national catalogue is received in the HealthData@EU central services, a return message informing about the success of the operation will be sent from the central services to the national infrastructure. The operation status transmitted through the returning message will be stored in the database of the health data dispatcher. The CBE should allow to automatically verify the status of these operations, generating alerts when a failure or lack of acknowledgment occurs.

4.8. Non-functional requirements

As outlined in the D5.3 Technical Specification for the National Metadata Catalogue from the TEHDAS2 project [3], it is also recommended that Member States consider a set of non-functional requirements when developing their national dataset catalogues, to foster a common baseline for performance, availability, reliability, and security across EU countries.

Key non-functional aspects include:

- **Availability:** Targeting a system uptime of at least 99%, excluding planned maintenance.
- **Performance:** Ensuring quick responses to typical metadata queries (ideally under 1 second).
- **Reliability:** Implementing failure detection and recovery to support system stability.
- **Scalability:** Supporting growth in dataset volume and user access without affecting performance.
- **Security and Access Control:** Applying appropriate access controls, authentication, and authorisation—defined nationally—to protect sensitive metadata and maintain system integrity.

³ The cross-border gateway, health data dispatcher and cross-border engine are 3 components that together ensure the transmission of eDelivery messages from the national datasets catalogue to the HealthData@EU datasets catalogue. Further details on this components are out of the scope of this deliverable and can be consulted in HealthData@PT deliverable 8.1 and in the HealthData@EU release 4 related documentation [24].

- **Training** (For HDAB Staff, data users, data holders) on catalogue matters.

These abovementioned non-functional requirements were also considered when drafting the requirements and during the development of HealthData@PT national datasets catalogue. Likewise, and regarding training, two internal workshops for HDAB technical teams were conducted in 2025 (further details in D2.3 of HealthData@PT) and training modules are being constructed for data holders and data users.

5. Building up the national datasets catalogue

The following sections present a comprehensive description of the ongoing work on implementing the HealthData@PT national datasets catalogue.

5.1. Dataset Specifications

The minimum viable product (MVP) of the HealthData@PT nHDsC will contain 25 datasets, fully described using the HealthDCAT-AP metadata standard. To that end, the definition and cataloguing of datasets are currently in progress, using data from the NHS Information and Monitoring System (SIMSNS) database and following Article 51 of the EHDS regulation. The datasets selected for cataloguing are based on categories for which data is available in SIMSNS and aligned with current key topics, such as major diseases identified by the WHO. Another important factor that was considered was the interest of national researchers in specific data, based on the insight gathered through requests previously received.

Table 4 exhibits some features of the first datasets under preparation for the Portuguese HealthData@PT datasets catalogue, providing the dataset's name, a short description of its contents, the EHDS Regulation Art 51 health data category it belongs to, and examples of contained variables and use cases are presented. These datasets are being described following the metadata standard HealthDCAT-AP.

Table 4. Sample of 8 datasets described with the HealthDCAT-AP standard, for the nHDsC MVP.

Dataset	Description	Art. 51 category ^a	Example of Variables	Example use-case
Reason for consultation in Primary Health Care Units	Registration of the reason for consultation, by ICPC-2 code (International Classification of Primary Care).	I	Gender, age, date of consultation, ICPC code.	During the year 2020, how many people were diagnosed with tuberculosis in primary health care?
Vaccines	Record of vaccine administration according to the vaccination plan.	C	Gender, year of birth, death (yes/no), year of death, date of administration, location, vaccine ID (following Portuguese DGS classification), vaccine batch.	Monitor the administration of the tetanus vaccine in the Portuguese population.
COVID-19 vaccines	Record of the administration of COVID-19 vaccines.	C	Gender, year of birth, death (yes/no), year of death, date of administration, location, vaccine ID, vaccine batch.	Monitor the COVID-19 vaccination in 2023.

Dataset	Description	Art. 51 category ^a	Example of Variables	Example use-case
Diagnosis and monitoring of Hypertension	Monitoring of the program for the follow-up of patients with hypertension. Includes data since 2015, for all patients diagnosed with hypertension.	I	Records of weight, height, blood pressure measurements, and other patient data collected during follow-up appointments.	Find out how many people in the population have high blood pressure, how many are women, and how many patients are overweight.
Medication for patients with Hypertension	Identification of chronic medication taken by patients diagnosed with hypertension, with and without associated issues, followed up in Primary Health Care units.	O	Prescription date, end date of medication, active ingredients, dosage, defined daily dose, medication form and quantity, medication description.	Check the most common medication used in the population diagnosed with hypertension.
Diagnosis and monitoring of diabetes	Monitoring of the follow-up program for patients with diabetes (types I, II, and gestational). Data since 2015, for all patients identified with diabetes.	I	Weight, height, abdominal circumference, fasting blood glucose, postprandial blood glucose, glycated hemoglobin.	Study of weight variations in patients with type II diabetes.
Medication for patients with diabetes	Identification of chronic medication taken by patients with diabetes (types I, II, and gestational), followed up in Primary Health Care units.	O	Prescription date, end date of medication, active ingredients, dosage, defined daily dose, medication form and quantity, medication description.	Check the most common medication in the population diagnosed with diabetes.
Electronic prescription of medications.	Monitor the electronic prescription of medications, also considering the number of packages prescribed according to the International Nonproprietary Name.	O	Period (month/year), region, total number of prescriptions; number of packages prescribed by brand (by exception A, B, C).	Check the number of prescriptions in a specific region of Portugal during the period of 2020 and 2021.

5.2. Metadata specifications

Beyond the mandatory HealthDCAT-AP properties present in Annex 2, the metadata of the datasets present in the HealthData@PT nHDsC will also feature some HealthDCAT-AP recommended properties (Table 5). These were chosen based both on their utility as possible catalogue search filters and on their general utility to the data user (for example, the properties “Number of Records” and “Number of unique individuals” often are exclusion factors for researchers that require large sample size for cohort studies).

Table 5. HealthDCAT-AP Recommended Properties that will be mandatory in the PT nHDsC metadata.

Property	Description
Frequency	The frequency at which the dataset is updated.
Landing Page	A web page that provides access to the dataset, its distributions and/or additional information.
Legal Basis	The legal basis used to justify processing of personal data
Language	Language of the Dataset
Max Age	Maximum typical age of the population within the dataset
Min Age	Minimum typical age of the population within the dataset
Number of Records	Size of the dataset in terms of the number of records
Number of unique individuals	Number of records for unique individuals
Personal Data	Key elements that represent an individual in the dataset
Population Coverage	A definition of the population within the dataset
Temporal Coverage	A temporal period that the dataset covers
Temporal Resolution	The minimum period resolvable in the dataset

Additionally, a **data dictionary** including all of the datasets’ variables will be included in with its respective sample (i.e. mock-up, anonymized or synthetic data), to better inform data users on datasets’ characteristics and structure. The dataset sample will always be constituted by synthetic data whenever possible, and anonymised data otherwise, to minimize reidentification risk.

Following the EHDS Regulation, a data quality and utility label will also be included as metadata for all datasets financed with public funds (European and/or National). This label will follow the QUANTUM (<https://quantumproject.eu>) requirements and specifications [25,26].

5.3. nHDsC Mock-up

The first version of the nHDsC has begun construction, using CKAN. To start off with catalogue development, datasets present in the HealthData@EU central catalogue were imported, to streamline the development process of the user interface with datasets already available using the HealthDCAT-AP standard. The first development version (Figure 29) exhibits some of the basic search functionalities provided, including a free-text search function and basic filtering capabilities by keywords.

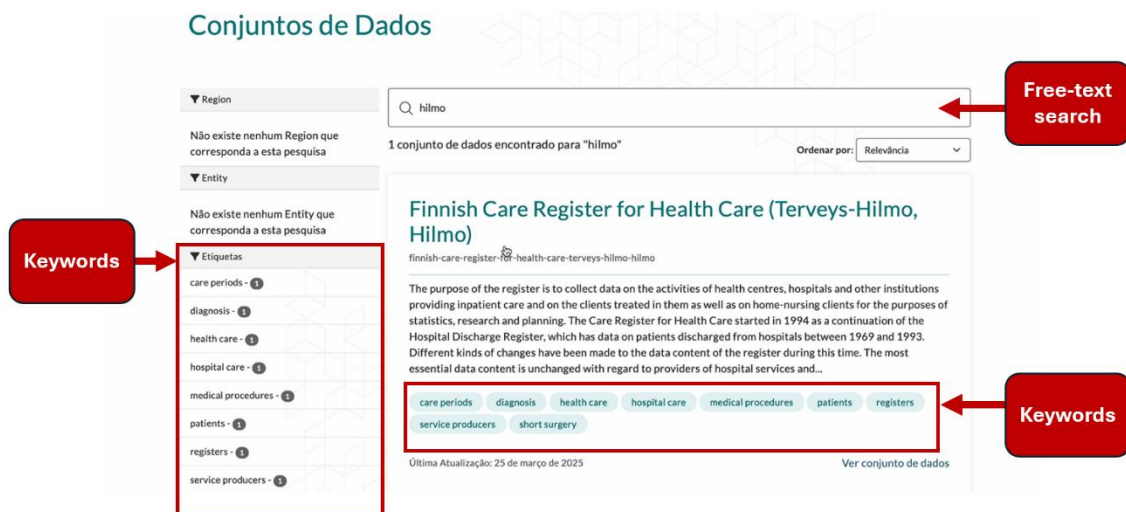
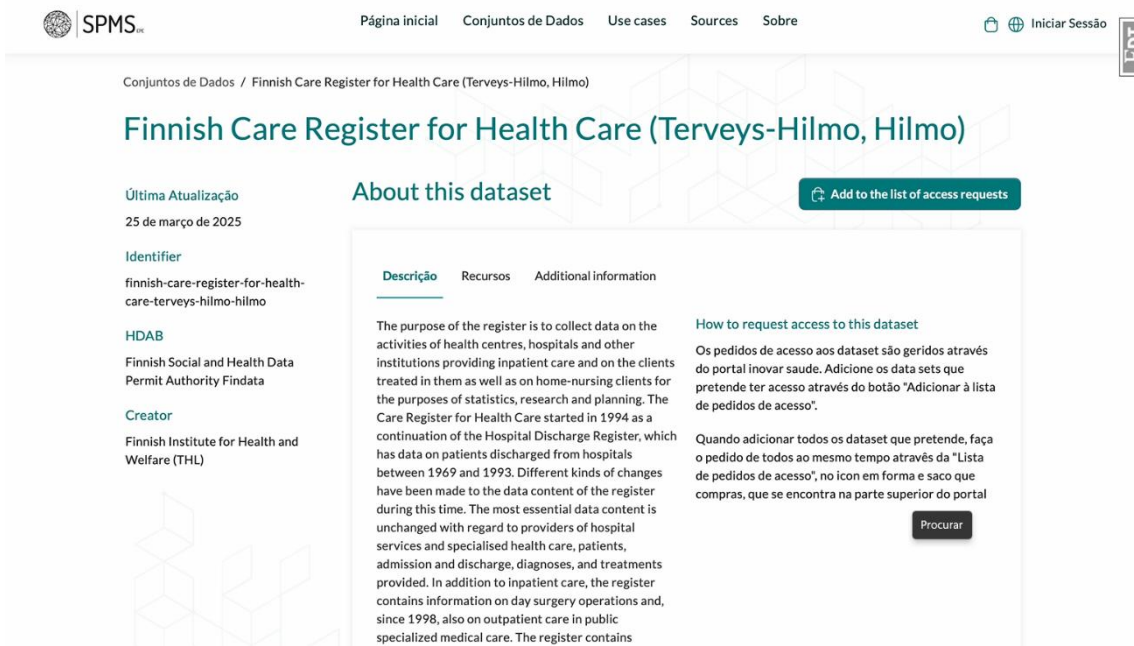


Figure 29. First development version of the Portuguese national datasets catalogue.

In Figure 30, an example of a dataset record can be seen: this page includes all key information related to the dataset, such as when it was last updated, the data holder, the HDAB, and a description of the dataset. Since the nHDsC and the DAAMs solutions will be connected, the page also instructs the data user on how to apply for data access, and features an interactable element to quickly add the dataset to the basket.



The screenshot shows a web interface for the 'Finnish Care Register for Health Care (Terveys-Hilmo, Hilmo)'. The header includes the SPMS logo and navigation links: 'Página inicial', 'Conjuntos de Dados', 'Use cases', 'Sources', and 'Sobre'. A user session link 'Iniciar Sessão' is also present. The main content area is titled 'Finnish Care Register for Health Care (Terveys-Hilmo, Hilmo)' and includes a sidebar with metadata: 'Última Atualização' (25 de março de 2025), 'Identifier' (finnish-care-register-for-health-care-terveys-hilmo-hilmo), 'HDAB' (Finnish Social and Health Data Permit Authority Findata), and 'Creator' (Finnish Institute for Health and Welfare (THL)). The main section, 'About this dataset', has tabs for 'Descrição', 'Recursos', and 'Additional information'. The 'Descrição' tab is active, showing a detailed description of the register's purpose and history. A 'How to request access to this dataset' section provides instructions on using the 'portal inovar saude' and adding datasets to a list of access requests. A 'Procurar' button is located at the bottom right of the description area.

Figure 30. Example of a dataset record in the first development version of the Portuguese national datasets catalogue.

5.4. National Implementation

The design and implementation of the national dataset catalogue will be built up in phases. These will be further detailed in HealthData@PT D6.2 and D6.3, which will report on the design and pilot implementation of the nHDS, respectively.

This D6.1 document comprises the first stage, laying down the requirements and specifications based on the obligations of the EHDS Regulation, on the technical specification from TEHDAS2 [3], and the overview of the dataset catalogues' landscape presented in section 3.3.

The following phases include continuing with the design of the technical solutions to implement the requirements and specifications laid in this document, and the implementation of the nHDS MVP. Both phases will leverage the national infrastructure and European advances to create a service completely compatible with the EC's HealthData@EU central services catalogue that answers to the national data discovery needs.

6. Concluding Remarks

The European landscape of dataset catalogues/metadata registries contains a few different solutions, all tailored to the current national infrastructures and to their data discovery needs. This creates a fragmented system where data users struggle to make the best use of international data, hindering cooperation and the free flow of information. A need to create a better knowledge and data-sharing environment for all European Union Member States led to the drafting of the EHDS Regulation, aiming to, among other things, improve healthcare at the point of care (primary use), to accelerate medical advancements and research and facilitate response to public health crises (secondary use).

To fully establish this cross-border cooperation, it is important to ensure that the multiple Member States solutions are interoperable, and for the national datasets' catalogues, this comes with the establishment of a metadata standard. HealthDCAT-AP as a standard that all metadata descriptions present in the European datasets catalogue need to follow is a major step in ensuring interoperability, doubly so when member-states adopt the standard for their own national catalogues.

This document (D6.1) constitutes one of the steps required to establish the Portuguese health datasets catalogue, in outlining the requirements and specifications for the implementation of the MVP of this infrastructure. For this, the state-of-the-art of the current European metadata registries and dataset catalogues was consulted, as well as the outputs of projects that aimed to tackle the construction of the European Health Data Space (such as TEHDAS, HealthyCloud, EHDS2 Pilot and TEHDAS2).

The ongoing outputs from the TEHDAS2 Joint Action, tasked with issuing a series of recommendations for the implementation of the necessary DBCs under the EHDS Regulation, were key for the construction of this report. Furthermore, as these recommendations mature and more documents are issued, the national efforts will take future guidelines into account for the implementation plan and design of the nHDsC. In that vein, current national efforts already count with the design and implementation of a Data Lake, and as such the nHDsC will leverage this infrastructure in its web services.

In closing, the Portuguese nHDsC will be the gateway for Portuguese health data, both at national and European level, and aims to facilitate the data discovery process that is key for effective (re)use of health data.

7. Bibliography

- [1] European Commission, European Health Data Space Regulation (EHDS) - European Commission, (2025). https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en (accessed April 11, 2025).
- [2] TEHDAS2 Consortium, M5.3 Draft technical specification on the national metadata catalogue, 2025. <https://tehdas.eu/wp-content/uploads/2025/01/2025-01-20-tehdas2-milestone-5.3.pdf>.
- [3] B. Barros, M. Englund, A. Gustafsson, T. Korsgaard, M. Peolsson, Ö. Petersson, C. Sebire, D. Welter, P. Wiklander, D5.3 Technical specification on the national metadata catalogue, 2025. <https://tehdas.eu/results>.
- [4] V. Kalokyri, H. Kondylakis, S. Sfakianakis, K. Dovrou, G. Manikis, K. Marias, M. Tsiknakis, M. Ghosh, M. Sambres, M. Vaterkowski, I. Blanquer, J. Mora, A. Vergara, L. Saint-Aubert, X. Rafael-Palou, P. Laras, G. Tsakou, T. Kussel, C. Calderón, A. Kosvira, D. González, C. Hernandez-Ferrer, L. Cerdá, M. Fernández, P. Martínez, D5.1. Early release of the Data Federation Framework, 2023. https://cancerimage.eu/wp-content/uploads/2023/10/D5.1_Early-release-of-the-Data-Federation-Framework_vf.pdf.
- [5] L.A. Tedd, Digital Libraries: Principles and Practice in a Global Environment, De Gruyter, Inc, Berlin/Boston, 2005.
- [6] J. Riley, Understanding metadata: what is metadata, and what is it for, National Information Standards Organization, Baltimore, MD, 2017.
- [7] B.E. Bargmeyer, D.W. Gillman, Metadata Standards and Metadata Registries, (2000). <https://www.bls.gov/osmr/research-papers/2000/pdf/st000010.pdf>.
- [8] Publications Office of the European Union, Models - EU Vocabularies - Publications Office of the EU, (n.d.). <https://op.europa.eu/en/web/eu-vocabularies/models> (accessed August 5, 2024).
- [9] I. Antunes, E. Bernal-Delgado, A. Bódi, F. Cascini, P. Derycke, S. Dinis, F. Estupiñán-Romero, Y. Fonseca, J. González-García, L. Grondin, I. Kesisoglou, T. Korsgaard, V. Lima, H. Lodenius, K. Lundgren, J. Lähteenmäki, V. Mendes, J. Pajula, C. Pinto, M. Pirttivaara, P. Schardax, K. Schneider, A. Skogholt, D. Spalding, C. Tellería-Orriols, D7.2 Options for the services and services' architecture and infrastructure for secondary use of data in the EHDS, 2023. <https://tehdas.eu/tehdas1/results/tehdas-proposals-for-the-implementation-of-ehds-technical-infrastructure>.
- [10] C. Alvarez-Romero, A. Martínez-García, D3.2 Guidelines to standardise metadata templates and assessment of FAIRness maturity levels, 2023. https://healthycloud.eu/wp-content/uploads/2023/04/D3.2_v1.2_to_submit.pdf.
- [11] C. Shona, K. Irene, D. Pascal, D3.3 Landscape analysis using a health-related data catalogue matrix, (2023) 46. <https://doi.org/10.5281/ZENODO.10226557>.
- [12] TEHDAS2 Consortium, M5.1 Draft guideline on data description, 2024. <https://tehdas.eu/wp-content/uploads/2025/01/2025-01-20-tehdas2-milestone-5.1.pdf>.
- [13] PwC EU Services, Requirements catalogue for scale-up-version (D02.03), 2024.
- [14] Guus Schreiber, Yves Raimond, W3C Working Group Note: RDF 1.1 Primer, (2014). <https://www.w3.org/TR/rdf11-primer/> (accessed August 5, 2025).
- [15] Richard Cyganiak, David Wood, Markus Lanthaler, W3C Recommendation: RDF 1.1 Concepts and Abstract Syntax, (2014). <https://www.w3.org/TR/rdf11-concepts/> (accessed August 6, 2024).

- [16] R. Albertoni, D. Browning, S.J.D. Cox, A.G. Beltran, A. Perego, P. Winstanley, Data Catalog Vocabulary (DCAT) - Version 3, (2024). <https://www.w3.org/TR/vocab-dcat-3/>.
- [17] Bert Van Nuffelen, DCAT-AP 3.0, (n.d.). <https://semiceu.github.io/DCAT-AP/releases/3.0.0/>.
- [18] DCAT-AP: How to extend DCAT-AP? | Interoperable Europe Portal, (n.d.). <https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-implementation-guidelines/release-1> (accessed August 12, 2025).
- [19] Pascal Derycke, HealthDCAT-AP, (2023). <https://healthdcat-ap.github.io/> (accessed June 20, 2024).
- [20] Mirna El Ghosh, Melanie Sambres, Catherine Duclo, Ferdinand Dhombres, Xavier Tannier, Valia Kalokyri, Stelios Sfakianakis, Manolis Tsiknanis, Christel Daniel, D5.2. The EUCAIM CDM and Hyper-Ontology for Data Interoperability: final version, (2024). <https://cancerimage.eu/achievements/>.
- [21] S. Hourri, L. Pery, E. Demnati, D5.1 nHDsC(9): Requirements/Specifications, (2024). https://ec.europa.eu/info/funding-tenders/opportunities/grants/docs/080166e514988156/Attachment_0.pdf.
- [22] L.O.B. Da Silva Santos, K. Burger, R. Kaliyaperumal, M.D. Wilkinson, FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication, Data Intell. 5 (2023) 163–183. https://doi.org/10.1162/dint_a_00160.
- [23] Luiz Olavo Bonino, Kees Burger, Rajaram Kaliyaperumal, FAIR Specifications: FAIR Data Point, (2023). <https://specs.fairdatapoint.org/fdp-specs-v1.2.html> (accessed May 5, 2025).
- [24] European Commission. Directorate General for Health and Food Safety., HealthData@EU central platform: open source release 4 : architecture model and technical specification., Publications Office, 2025. <https://data.europa.eu/doi/10.2875/2439009> (accessed August 8, 2025).
- [25] QUANTUM project, Deliverable 1.1 Specification of the data sets' quality and utility label, (2025). <https://zenodo.org/records/14937423> (accessed April 14, 2025).
- [26] QUANTUM project, Deliverable 1.2 Specification for the assessment of data holders' maturity, (2025). <https://zenodo.org/records/14944767> (accessed April 14, 2025).

Annexes

Annex 1. HealthyCloud WP3 recommendations

Table 6. HealthyCloud WP3 recommendations to uphold FAIR data principles in Health Data that are related with datasets' catalogues.

Principle	Recommendations
Findability	Expose machine-readable metadata and keywords to search engines and catalogues; Utilization of data referencing systems (e.g. Digital Object Identifier - DOI) and open data registries (e.g. ELIXIR); Construction of metadata catalogues utilizing descriptive metadata standards, such as DCAT-AP.
Accessibility	Robust and public data access systems, that should contain: Communication protocols; Clear data access policies and legal basis for data access; Data sharing agreements.
Interoperability	Utilization of internationally recognized standards, ontologies and vocabularies, such as: Fast Healthcare Interoperability Resources (FHIR) Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) Standardized data and metadata.
Reusability	Definition of clear and accessible data sharing, reuse and usage licences, such as Creative Commons; Publishing of metadata in publicly accessible and findable repositories like Zenodo.

Annex 2. Mandatory Properties for the HealthDCAT-AP Standard

Table 7. Mandatory properties for the main classes of the HealthDCAT-AP Standard (for non-public data). Classes without mandatory properties are instead hyperlinked to their definition [19].

Class	Mandatory Property	Description
Agent	Name	Name of an Agent
Catalogue	Applicable legislation	Legislation applicable to the resource
	Description	An account of the Catalogue
	Publisher	The Agent responsible for making the Catalogue available
	Title	Name of the Catalogue
Catalogue Record	Modification date	Latest date when the Catalogue was changed
	Primary topic	Link to the resource described in the record
Catalogue Resource	N/A	N/A
Concept	N/A	N/A
Dataset	Access rights	Information regarding whether the Dataset is publicly available, not public, or if access restrictions apply
	Applicable legislation	Legislation applicable to the creation or management of the Dataset
	Contact point	Contact information for communication regarding the resource
	Dataset distribution	An available Distribution for the Dataset
	Description	An account of the Dataset
	Geographical coverage	Spatial coverage of the Dataset
	Health category	Health category according to the electronic data for secondary use regulation, Art.33
	Health data access body	Information regarding the HDAB the Dataset originates from
	Health theme	Wikidata URIs containing the themes related to the Dataset
	Identifier	Main URI or other identifier for the Dataset
	Keyword	Tag describing the Dataset
	Provenance	Statement regarding the origin and custody of the Dataset
	Publisher	The Agent making the Dataset available
	Publisher note	Short description of the Publisher activities
	Publisher type	Type of publisher providing the Dataset
	Purpose	Finality of the processing of data or personal data.
	Sample	Sample distribution of the Dataset
	Theme	A Category of the Dataset (a Dataset can have multiple Themes, e.g. Vaccines, COVID-19 – All linked through wikidata)

Class	Mandatory Property	Description
Dataset	Title	Name of the Dataset
	Type	Nature of the Dataset (e.g. Personal Data)
Distribution	Access URL	URL that provides access to a Distribution of a Dataset
	Applicable legislation	Legislation applicable to the Distribution of the Dataset
Location	N/A	
Relationship	N/A	
Rights Statement	N/A	



Setting up a Health Data Access Body in Portugal and laying the national foundations for EHDS2



Copyright © 2025 HealthData@PT. All rights reserved.