M5.1 Draft guideline on data description

Guideline for health data holders on their duties regarding data description

TEHDAS2 – Second Joint Action Towards the European Health Data Space

17 January 2024

This project has been co-funded by the 4th EU Health Programme (2021–2027) under Grant Agreement no 101176773.

0 Document info

Disclaimer

Views and opinions expressed in this deliverable represent those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

0.1 Authors

Lead Author(s)	Lead organisation
Nienke Schutte	Sciensano, Belgium
Beatriz Barros	Sciensano, Belgium
Pascal Derycke	Sciensano, Belgium
Charles-Andrew Vande Catsyne	Sciensano, Belgium

0.2 Keywords

Keywords	TEHDAS2, Joint Action, Health Data, European Health Data Space, Health Data Holders, Metadata, Data Description, Dataset Catalogue, DCAT-AP

0.3 Document history

Date	Version	Editor	Change	Status
19/12/2024	0.1	Sciensano team	Finalised draft	Draft
17/10/2024	0.2	Sciensano team	Updated document based on TEHDAS2 and EC comments	Draft

Accepted in Project Steering Group on 14.01.2024

Contents

Contents 3
1 Executive summary 4
2 Introduction 5
2.1 Purpose 5
2.2 Problems being addressed 6
2.3 Target audience/intended users 7
3 Key terminology 9
4 Scope 12
5 Methodology: design of a common descriptive metadata model for health data 12
6 Proposed solution for personal electronic health data: HealthDCAT-AP 13
6.1 What is HealthDCAT-AP? 13
6.2 Key Features of HealthDCAT-AP 13
6.3 How was HealthDCAT-AP developed? 14
6.4 DCAT-AP foundational framework 14
6.5 Metadata records for different types of data access 16
7 Extending DCAT-AP for health: HealthDCAT-AP 20
7.1 DCAT-AP properties of the Dataset Class 21
7.2 DCAT-AP Adjusted Properties 26
7.3 HealthDCAT-AP new properties 30
8 Next steps toward a complete guideline 37
Annex I: Links to the EHDS Regulation 38
Annex II: HealthDCAT-AP UML Diagram 40
Annex III: Workshops on Article 51 – health categories 41
Annex IV: HealthDCAT-AP technical working groups 42

1 Executive summary

This draft guideline provides guidance for health data holders on a practical metadata model/standard for describing datasets, according to the EHDS framework, to facilitate secondary use of health data.

This document outlines the principles, rationale, and methodology behind the development of HealthDCAT-AP, a domain-specific extension of the DCAT-AP (Data Catalogue Application Profile). DCAT is a W3C standard providing a vocabulary to describe datasets in data catalogues, while DCAT-AP is its application profile tailored for Europe to ensure semantic interoperability. HealthDCAT-AP, developed in the HealthData@EU project (www.ehds2pilot.eu), adapts the DCAT-AP framework to meet the specific requirements of health data management under the European Health Data Space (EHDS) Regulation. It aims to enhance dataset discoverability and interoperability, especially in the context of secondary use of health data.

This first draft of the guideline is based on the initial version of HealthDCAT-AP, which incorporates health-specific metadata elements and properties. It provides examples and guidance to assist data holders in creating metadata records compliant with this extension. The document also explains the rationale for modifying and expanding the baseline DCAT-AP specification to address the unique characteristics of health data.

These guidelines are designed to help health data holders—including healthcare providers, institutions, and administrators—comply with their obligations under the EHDS Regulation (Articles 60 and 77). Specifically, this document:

- Explains how to use HealthDCAT-AP to describe datasets;
- Provides clear, practical steps to ensure metadata is accurate, interoperable, and compliant with legal requirements;
- Highlights the benefits of improved data discoverability and collaboration across Europe.

2 Introduction

As part of the European Health Union, the European Union (EU) is advancing the use of health data for secondary purposes, including research, innovation and policy making. Smooth and secure access to data will drive the development of new treatments and medicines and optimise resource utilisation—all with the overarching goal of improving the health of citizens across Europe.

TEHDAS2, the second joint action Towards the European Health Data Space, represents a significant step forward in this vision. The project will develop guidelines and technical specifications to facilitate cross-border use of health data, and support data holders, data users and the new Health Data Access Bodies (HDABs) in fulfilling their responsibilities and obligations outlined in the EHDS regulation.

TEHDAS2 focuses on several critical aspects of health data use, including:

- **Data discovery**: findability and availability of health data, ensuring it is accessible for secondary purposes;
- **Data access**: developing harmonised access procedures and establishing standardised approaches for granting data access across member states;
- **Secure processing environment**: defining technical specifications for environments where sensitive health data can be processed safely;
- **Citizen-centric obligations:** providing guidance on fulfilling obligations to citizens, such as communicating significant research findings that impact their health, informing them about research outcomes and ensuring transparency in how their data is used;
- **Collaboration models**: developing guidance on collaboration and guidelines on fees and penalties as well as third country and international access to data.

TEHDAS2 will contribute to harmonised implementation of the EHDS regulation through the concrete guidelines and technical specifications. Some of these documents and resources will also provide input to implementing acts of the regulation. Hence, the joint action will increase the preparedness for the EHDS implementation and lead to better coordination of member states' joint efforts towards the secondary use of health data, while also reducing fragmentation in policies and practices related to secondary use.

TEHDAS2 Work Package (WP) 5 aims to facilitate the discovery of datasets for secondary use of electronic health data under the EHDS framework. Task 5.1 is set to build a set of requirements towards data holders to fulfil their duties for data description. This document presents the first draft of Task 5.1, focusing on how health data holders can fulfil these duties utilising the HealthDCAT-AP.

2.1 Purpose

The EHDS Regulation aims to create a common unified framework for securely sharing and exchanging electronic health data across Europe. The regulation establishes clear rules and processes for data availability, usage conditions and supports repurposing data originally collected for other purposes, such as research, innovation or policymaking, through a shared European infrastructure, HealthData@EU.

The EHDS is part of the broader **EU Data Strategy**, which seeks to create interconnected **European data spaces** across strategic fields, including health. These data spaces are designed to promote the responsible and secure use of data while improving collaboration and accessibility for societal benefit. By harmonising practices and ensuring interoperability, the EHDS will simplify data-sharing processes, foster collaboration among Member States, and support the efficient use of resources. Standards for **interoperability** and **security**, such as DCAT-AP (Data Catalogue Application Profile), play a critical role in this effort, ensuring consistency across the EHDS and other European data spaces.

The use of **DCAT Application Profile (DCAT-AP) for data portals in Europe** as baseline specification for metadata records is a cornerstone for semantic interoperability in the EHDS and with other European data spaces. In this context, HealthData@EU pilot (GA 101079839, www.ehds2pilot.eu) started the work on the **common descriptive metadata model HealthDCAT-AP**, tailored to the health domain. HealthDCAT-AP refines the DCAT model to support the discovery and understanding of health datasets, improving their accessibility while ensuring privacy and security.

These guidelines are designed to help health data holders—including healthcare providers, institutions, and administrators—comply with their obligations under the EHDS Regulation (Articles 60 and 77). Specifically, this document:

- Explains how to use HealthDCAT-AP to describe datasets;
- Provides clear, practical steps to ensure metadata is accurate, interoperable, and compliant with legal requirements;
- Highlights the benefits of improved data discoverability and collaboration across Europe.

This guideline is designed to provide health data holders with a comprehensive description regarding HealthDCAT-AP as a solution for ensuring semantic interoperability and improving data discoverability across the EHDS and other European data spaces. By following these guidelines, health data holders can fulfil their legal obligations while contributing to the broader goals of a secure and interoperable European Health Data Space.

2.2 Problems being addressed

Recital 80 of the EHDS regulation emphasises linking European data spaces, enabling the secondary use of health data with data from sectors like environment, agriculture, and social fields to gain insights into health determinants, highlighting the aim of the European Commission to strive for **interoperability between the common European data spaces**. The harmonisation of dataset descriptions supports these interoperability principles through the common use of the DCAT-AP for data portals in Europe, already acknowledged in the first joint action TEHDAS[1].

Furthermore, a single data description standard is needed to **enhance data discovery** by ensuring consistency in how data is described, organised and shared across different platforms. A unified

[1] Recommendations to enhance interoperability within HealthData@EU- a framework for semantic, technical and organisational interoperability (2023). Joint Action Towards the European Health Data Space https://tehdas.eu/tehdas1

data description standard enables datasets to be easily catalogued and understood, regardless of their origin or system, facilitating seamless search and retrieval. This harmonisation reduces fragmentation, improves interoperability and allows health data users to efficiently locate relevant, high-quality data across national and EU-level infrastructures.

Finally, the FAIR[2] data principles stand for Findability, Accessibility, Interoperability, and Reusability, which are guidelines to ensure that data and metadata are FAIR. The EHDS Regulation endorses the FAIR data principles to ensure health data sharing and responsible reuse across the EU. A standardised data description is essential for implementing the FAIR principles by ensuring that data is findable, accessible, interoperable, and reusable. It enables data discovery through standardised descriptions, provides information on how to access datasets, ensures compatibility across systems, and documents the context and quality of data for future use. By facilitating these aspects, metadata supports better data management and sharing, making data more useful and accessible for various purposes.

2.3 Target audience/intended users

This document is primarily aimed at health data holders, who are responsible for describing and sharing their datasets in accordance with the EHDS Regulation. The definition of 'health data holder' is described in Article 2 of the EHDS Regulation, and can include a wide range of entities, such as:

- Healthcare providers (e.g., hospitals, clinics, and general practitioners);
- Public authorities or agencies involved in health or care services;
- Organisations managing reimbursement systems;
- Developers of health-related products and services, including wellness applications;
- Research institutions and mortality registries;
- EU institutions, bodies, and agencies that manage or process health data.

These entities may handle personal or non-personal health data and are responsible for ensuring that their datasets are properly described and shared with Health Data Access Bodies (HDABs).

Under the EHDS Regulation (Articles 60 and 77), health data holders are required to:

- **Describe their datasets:** Provide clear, accurate metadata about the health data they hold, to the health data access body, in a standardised manner;
- **Update this information regularly:** Verify and update dataset descriptions at least once a year to ensure accuracy.

This document will aid the health data holder in these requirements, through a clear description of the properties in HealthDCAT-AP.

The categories of electronic health data that should be made available for secondary use, and, therefore, should be described by its health data holder, this is described in Article 51 and can be

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

found in Table 1. TEHDAS2 Work Package 5 is working on providing full text definitions and examples for each of the categories (foreseen to be published in Deliverable 5.1, expected summer 2025).

Table 1: Minimum categories of electronic health data that should be made available for secondary use (Article 51 EHDS).

Nr	Category
a	Electronic health data from Electronic Health Records (EHRs)
b	Data on factors impacting on health, including socio-economic, 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, 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

Furthermore, **HDABs** could also benefit from clear descriptions of the properties in HealthDCAT-AP. Article 77 outlines that these bodies have the obligation to provide a description in the form of metadata of the available datasets. These guidelines support the HDABs in the establishment of standardised dataset catalogues (see technical specifications TEHDAS2 Milestone 5.3).

These guidelines will also serve the **health data user**, making it easier for them to understand the data. These guidelines help ensure that key aspects such as data quality, context, and accessibility are well-documented, reducing ambiguity and the need for users to seek additional clarification.

Finally, the **European Commission** is, through Article 79, tasked with establishing a publicly available EU dataset catalogue connecting the national dataset catalogues established by the health data access bodies in each Member State as well as the dataset catalogues of authorised

participants in HealthData@EU. In this process, a thorough understanding of the properties in HealthDCAT-AP is essential.

3 Key terminology

Community of Practice (CoP): The CoP comprises competent authorities and affiliated entities from EU Member States and European Economic Area (EEA) countries. Its members, involved in establishing the health data access bodies are responsible for the secondary use of health data within the EHDS in their respective Member States, including holders of the direct grants for setting up HDABs. It aims to foster collaboration and knowledge sharing among Competent Authorities and Affiliated Entities involved in establishing the HDABs and responsible for the secondary use of health data within the EHDS.

Content negotiation: Content negotiation refers to mechanisms defined as a part of the W3C HTTP protocol that make it possible to serve different versions of a document (or more generally, representations of a resource) at the same Uniform Resource Identifier (URI), so that user agents can specify which version fits their capabilities the best (Wikipedia). This mechanism can, for example, be used to serve a Resource Description Framework (RDF) representation of a DCAT metadata record for data exchange or an HTML format for browsers to display as a web page.

Controlled vocabulary: A controlled vocabulary is a predefined, standardised set of terms and phrases used to ensure consistency in naming and categorising concepts within a dataset. It restricts the use of alternative terms or synonyms to avoid ambiguity and maintain uniformity in data description. Controlled vocabularies are commonly used in metadata, taxonomies and classification systems to improve data discoverability, interoperability, and accuracy in search and retrieval across datasets or systems. Examples include thesauri, ontologies and code lists.

Dataset[3]: A collection of data featuring a data catalogue that describes datasets and offers services to facilitate their discovery and reuse. In the context of the EHDS Regulation, accessing these datasets must comply with the principles of data minimisation and purpose limitation. This ensures that only data relevant and necessary for the intended processing purpose are provided, in either anonymised or pseudonymised form, depending on what is feasible for meeting the processing objectives.

Dataset catalogue[4]: A collection of dataset descriptions, arranged in a systematic manner and including a user-oriented public part, in which information concerning individual dataset parameters is accessible by electronic means through an online portal.

Data dictionary: A data dictionary is a centralised repository of metadata that provides definitions, descriptions and details about the structure, fields and variables within a dataset. It typically includes information such as data types, allowed values, relationships between fields and the meaning of

[3] EHDS Regulation Article 2(2)w

[4] EHDS Regulation Article 2(2)y

each element. A data dictionary helps data users understand the content and structure of a dataset, facilitating proper use and interpretation of the data.

Data quality[5]: The degree to which the elements of electronic health data are suitable for their intended primary use and secondary use.

Data quality and utility label[6]: A graphic diagram, including a scale, describing the data quality and conditions of use of a dataset.

Health data access body (HDAB)[7]: A body designated by the Member State that is responsible for carrying out the tasks and obligations set out in Articles 57 (Tasks of health data access bodies), 58 (Obligations of health data access bodies towards natural persons) and 59 (Reporting by health data access bodies). Where a Member State designates several health data access bodies, it shall designate one HDAB to act as coordinator, with responsibility for coordinating tasks with the other health data access bodies both within the territory of that Member State and in other Member States.

Health data holder[8]: Natural or legal person, public authority, agency or other body in the healthcare or the care sectors, including reimbursement services where necessary, as well as any natural or legal person developing products or services intended for the health, healthcare or care sectors, developing or manufacturing wellness applications, performing research in relation to the healthcare or care sectors or acting as a mortality registry, as well as any Union institution, body, office or agency, that has either: (i) the right or obligation, in accordance with applicable Union or national law and in its capacity as a controller or joint controller, to process personal electronic health data for the provision of healthcare or care or for the purposes of public health, reimbursement, research, innovation, policy making, official statistics or patient safety or for regulatory purposes; or (ii) the ability to make available non-personal electronic health data through the control of the technical design of a product and related services, including by registering, providing, restricting access to or exchanging such data.

Health data user[9]: Natural or legal person, including Union institutions, bodies, offices or agencies, which has been granted lawful access to electronic health data for secondary use pursuant to a data permit, a health data request approval or an access approval by an authorised participant in HealthData@EU.

Interoperability[10]: the ability of organisations, as well as of software applications or devices from the same manufacturer or different manufacturers, to interact through the processes they support, involving the exchange of information and knowledge, without changing the content of the data, between those organisations, software applications or devices.

[5] EHDS Regulation Article 2(2)z
[6] EHDS Regulation Article 2(2)aa
[7] EHDS Regulation Article 55(1)
[8] EHDS Regulation Article 2(2)t
[9] EHDS Regulation Article 2(2)u
[10] EHDS Regulation Article 2(2)f

Knowledge graph: In knowledge representation and reasoning, a knowledge graph is a knowledge base that uses a graph-structured data model or topology to represent and operate on data. Knowledge graphs are often used to store interlinked descriptions of entities – objects, events, situations or abstract concepts – while also encoding the free-form semantics or relationships underlying these entities (Wikipedia).

Linked data: Linked data is a general term for a set of best practices for exposing data in machine-readable form using the content-negotiation feature of the standard HTTP web protocol. These best practices support the development of tools to link and make use of data from multiple web sources without the need to deal with many different proprietary and incompatible application programming interfaces (APIs), and use of HTTP to request data in structured form meant for machines instead of human-readable displays (doi.org).

Namespace: In the context of Linked Data, a namespace helps records have unique names. A namespace is a component of the URI. In a group of URIs produced as part of a dataset the shared part of the URI is often the namespace. For example, all concepts of the Language of Bindings thesaurus start with "https://w3id.org/lob/" which is the namespace for the thesaurus. In Linked Data the namespace may be declared with a shortcut using the keyword prefix. For example: @prefix lob: <https://w3id.org/lob/>. The prefix lob can then be used instead of the full namespace.

Personal electronic health data[11]: Data concerning health and genetic data, processed in an electronic form.

Semantic Web: The Semantic Web, sometimes known as Web 3.0 is an extension of the World Wide Web through standards set by the World Wide Web Consortium (W3C). The goal of the Semantic Web is to make internet data machine-readable. To enable the encoding of semantics with the data, technologies such as RDF and Web Ontology Language (OWL) are used. These technologies are used to formally represent metadata (Wikipedia).

Simple Knowledge Organisation System (SKOS): SKOS provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary. As an application of the, SKOS allows concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into other concept schemes. In basic SKOS, conceptual resources (concepts) are identified with URIs, labelled with strings in one or more natural languages, documented with various types of notes, semantically related to each other in informal hierarchies and association networks and aggregated into concept schemes.

Tabular data: Tabular data refers to data organised in a structured format of rows and columns, where each row represents a single record or entity, and each column represents a specific attribute or variable. This structure is commonly found in spreadsheets or relational databases, making it easy

[11] EHDS Regulation Article 2(2)a

to store, query, and analyse. Tabular data is often used for structured datasets where relationships between variables are well-defined[12].

Universal Resource Identifier (URI): URI is a unique address for a documentation record that we want others to refer to. The concept of 'paper', as defined in the Getty Arts & Architecture Thesaurus (AAT), can be referenced by the URI: 'http://vocab.getty.edu/aat/300014109'. One of the benefits of using URIs is machine disambiguation, i.e. it is clear to a machine where to point users when a record refers to 'paper' according to the Getty AAT definition. Also, a URI can be matched with other words for 'paper' in different languages, thus making records language independent.

4 Scope

This document describes the principles and the specification of the HealthDCAT-AP as delivered by the HealthData@EU Pilot in order to support health data holders in their data description duties. It describes the methodology and the changes made to the DCAT-AP to accommodate health data and their rationale. It includes a description of the HealthDCAT-AP properties and examples.

As mentioned in section 2.3, the final deliverable on guidelines for health data holders pertaining to their data description duties will be delivered in summer 2025 and will include a full text description of the minimum categories of health data as shown in Table 1. Moreover, the data quality and utility label that would need to be provided by health data holders according to Article 78 will be developed by the QUANTUM project (GA nr 101137057, https://quantumproject.eu/) and is out of scope of this draft guideline. Finally, guidelines for health data access bodies on limitations in relation to the purposes that are enlisted for secondary use according to EHDS and technical specifications on the national dataset catalogue will be published in separate reports (based on the work in TEHDAS2 WP5 task 5.2).

5 Methodology: design of a common descriptive metadata model for health data

The HealthDCAT-AP framework builds upon the DCAT-AP standard by adding health-specific metadata elements, properties and guidelines tailored to the unique requirements of health data management and sharing. This extension addresses the challenges of managing, sharing and discovering health-related datasets within the EHDS, ensuring compliance with EU data protection regulations, including the GDPR and the EHDS Regulation.

The HealthData@EU Pilot project initiated the development of the HealthDCAT-AP specification in January 2023 to enhance health data interoperability. Through a landscape analysis of existing metadata catalogues and DCAT profiles, and guided by a Technical Working Group (TWG), the project identified and refined metadata requirements. From September to December 2024, bi-weekly TWG sessions ensured alignment with FAIR principles and EU policies. A **draft version of HealthDCAT-AP**, hosted on GitHub (W3C ReSpec documentation style) for community feedback, extends DCAT-AP to meet health data needs. The final output of the HealthData@EU Pilot with

[12] https://w3c.github.io/csvw/metadata/

regards to this work encompasses **a deliverable with recommendations on further development and deployment for possible EU-wide uptake**[13]. As the HealthDCAT-AP model was only piloted by a few consortium partners of the HealthData@EU Pilot, **there is a need to collect feedback on the specification and its rationale**. TEHDAS2 provides the opportunity to finetune and validate the HealthDCAT-AP in the broader health community, including all EHDS relevant stakeholders.

6 Proposed solution for personal electronic health data: HealthDCAT-AP

6.1 What is HealthDCAT-AP?

HealthDCAT-AP is a metadata model designed to enhance the management and interoperability of health-related datasets within the EHDS. By standardising how health datasets are described and how dataset catalogues are conceived, it aids health data holders meet the requirements of the EHDS Regulation while supporting efficient data discovery, accessibility, and sharing across Europe. HealthDCAT-AP extends DCAT-AP, the European standard for dataset descriptions, to address the specific needs of describing personal and non-personal electronic health data. DCAT-AP is widely recognised by semantic experts as the most effective technical solution currently available for implementing EU data spaces. It ensures that key metadata—such as titles, descriptions, keywords, and access details—is consistently formatted and interoperable across national and EU-level catalogues. This standardisation simplifies data sharing and enables catalogues to "speak the same language". While **DCAT-AP provides a minimal common basis for sharing datasets** cross-border and across domains within Europe, it allows for extensions to meet the specific needs of different data landscapes.

The HealthDCAT-AP extension is one such adaptation, extending the core DCAT-AP model by introducing new properties, recommendations, and controlled vocabularies, specifically tailored to the unique requirements of electronic health data. Through this approach, **HealthDCAT-AP directly supports the implementation of the EHDS Regulation** by providing a structured framework for describing health datasets. This approach helps data holders comply with legal obligations (e.g., Articles 60 and 77 in the EHDS Regulation) while ensuring that datasets are discoverable and usable across the EHDS. By promoting a unified metadata approach, **HealthDCAT-AP enhances interoperability** and improves the efficiency of data sharing and integration.

6.2 Key Features of HealthDCAT-AP

- **Standardised metadata:** Defines consistent metadata elements to ensure health data descriptions are uniform and easy to understand.
- **Interoperability:** Facilitates seamless data exchange between national catalogues, HDABs, and the EU's central dataset catalogue.

[13] Recommendations on further development and deployment for possible EU-wide uptake (2024). HealthData@EU pilot study Deliverable 6.2

- **Machine readability:** Leverages the Resource Description Framework (RDF) to enable the implementation of a coherent and machine actionable semantic framework for the discoverability of datasets, supporting exchange, AI-driven tools and interfaces.

6.3 How was HealthDCAT-AP developed?

HealthDCAT-AP builds on best practices from DCAT-AP, ensuring full compatibility with DCAT-AP to support interoperability across domains. It introduces:

- New metadata properties specific to the health domain;
- New controlled vocabularies for consistent descriptions of concepts within the health domain;
- Recommendations to improve interoperability and compliance with EU standards;
- Refined usage guidelines for DCAT-AP properties to address the specific requirements of the health domain.

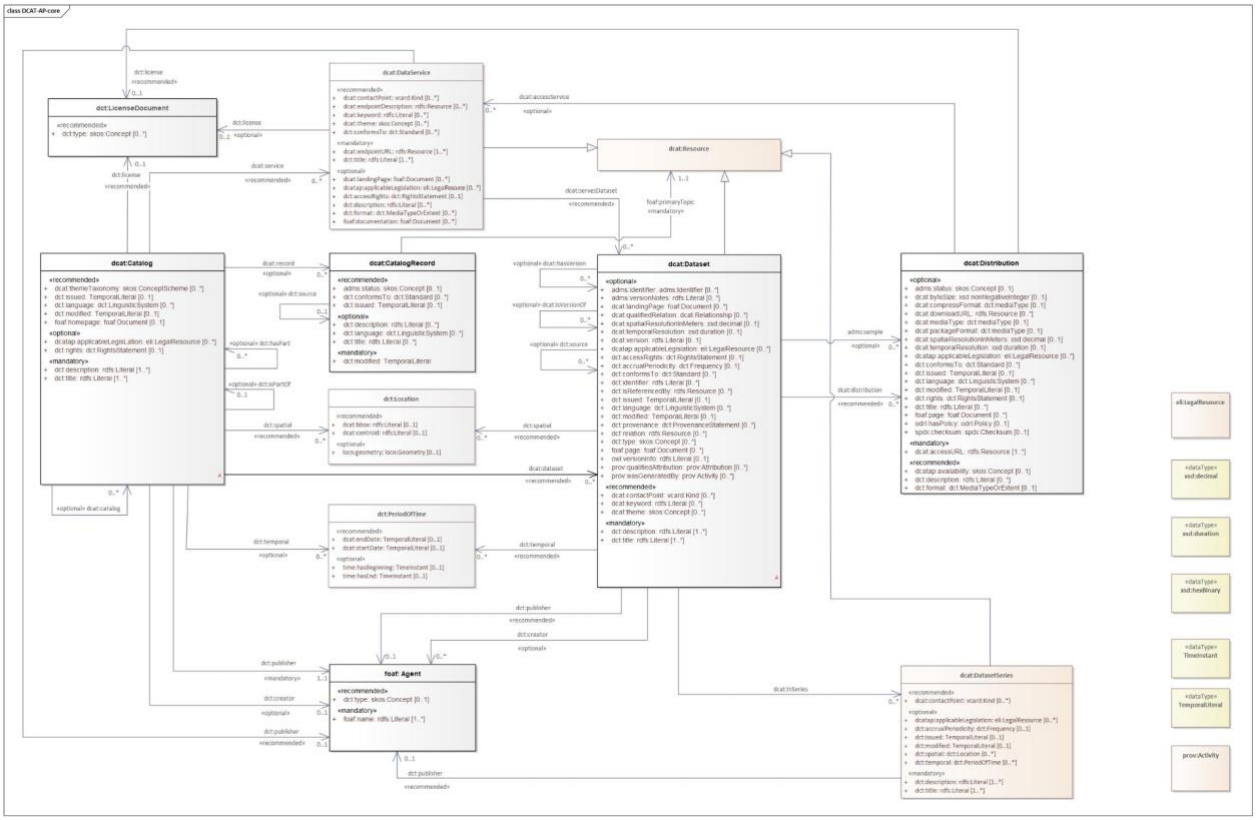
This approach ensures that HealthDCAT-AP remains adaptable to the evolving needs of electronic health data while maintaining compliance with EU regulations. For a complete and detailed explanation of HealthDCAT-AP, including technical examples in RDF, the draft specification of HealthDCAT-AP is available on GitHub. Users are encouraged to actively contribute to its refinement by submitting issues, commenting on content sections, or proposing improvements and changes. This collaborative approach ensures that HealthDCAT-AP evolves to meet the diverse needs of the health data landscape while adhering to EU regulations.

6.4 DCAT-AP foundational framework

HealthDCAT-AP extends DCAT-AP, which means that **HealthDCAT-AP retains all the elements of DCAT-AP**, maintaining the principal classes of DCAT-AP such as dcat:Catalog, dcat:CatalogRecord, dcat:Dataset, dcat:Distribution,dcat:DataService, among others while also adding new features tailored to health data specific needs. Therefore, **for applications to be compliant with HealthDCAT-AP, they must first conform to DCAT-AP**. This requirement is non-negotiable and serves to preserve the interoperability of dataset catalogues across the common European data spaces. This guideline focuses on HealthDCAT-AP, but due to its nature as an extension of DCAT-AP, it is crucial to first understand the foundational DCAT-AP standard. To assist data holders in visualising the structural organisation of DCAT-AP, Figure 1 below presents the UML[14] representation of the DCAT-AP standard version 3.0 (14 June 2024). This diagram illustrates the key classes and their relationships, providing a comprehensive overview of the base model upon which HealthDCAT-AP builds its health-specific extensions.

[14] Unified Modeling Language

Figure 1: DCAT-AP 3.0 UML Class Diagram.



In DCAT-AP, the importance and usage requirements of properties are defined by their cardinalities. Cardinalities specify the number of times an element can or must occur in relation to its parent element. This concept helps structure the metadata in a clear and consistent manner. The cardinalities in DCAT-AP are expressed as follows:

1. **Mandatory properties (1..1 or 1..*):** These MUST be included in the metadata. They are essential for describing datasets. The cardinality 1..1 means exactly one occurrence is required, while 1..* indicates at least one, but potentially multiple occurrences are needed.
2. **Recommended properties (0..1 or 0..*):** These SHOULD be included in the metadata. Their inclusion is strongly encouraged as they provide valuable additional information, but is not strictly required. The cardinality 0..1 allows for zero or one occurrence, while 0..* permits zero to many occurrences.
3. **Optional properties (0..1 or 0..*):** These MAY be included in the metadata. They offer supplementary details, and their usage is at the discretion of the metadata provider. The cardinalities are the same as for recommended properties, but with less emphasis on inclusion.

Additionally, cardinalities define how many elements can be included in a property. For instance, a property with cardinality 0..* can have multiple values, allowing for a more comprehensive description when needed. For instance, a property can be repeated to accommodate parallel language versions or multiple available resources. In HealthDCAT-AP, cardinalities have been carefully updated to meet the specific requirements of health data. These modifications ensure that the metadata

structure is tailored to the unique needs of the healthcare sector while maintaining compatibility with the original DCAT-AP standard. Further details about these cardinality updates are provided in the next section.

6.5 Metadata records for different types of data access

The EHDS Regulation mentions different types of health data access:

1. Non-personal electronic health data available as open data [open data], see also EHDS Regulation Article 60 (Annex I);
2. Non-personal electronic health data available as non-open data [protected or restricted data];
3. Personal electronic health data [sensitive data].

By categorising dataset accesses correctly and providing the appropriate metadata, data holders can ensure compliance with the EHDS Regulation while making their data discoverable and usable in a secure, regulated manner. To aid with this challenge, the different health data categories are defined in HealthDCAT-AP through the DCAT-AP **Access Rights property (dct:accessRights)** with the controlled vocabulary Access Right Authority maintained by the Publications Office. Therefore, the following rationale applies:

- If a dataset has non-personal electronic health data and is freely available to the public, the property access rights have to be defined as **"PUBLIC"**. For such datasets, the EHDS Regulation also mandates that at least one distribution is made accessible, aligning with the original purpose of DCAT-AP to describe publicly available data across European Open Data Portals.
- If a dataset has non-personal electronic health data but is access-controlled, it classifies as Protected/Restricted Data. Therefore, the property access rights have to be defined as **"RESTRICTED"**. Metadata must also include fields like publisher and distribution details specifying the format, size, and reuse conditions, to assure alignment with the Data Governance Act[15].
- If a dataset contains personal electronic health data, it classifies as Personal Data therefore the property access rights have to be defined as **"NON_PUBLIC"** and include details about the HDAB managing access. An in-depth description on how to describe personal electronic health datasets through HealthDCAT-AP is provided in the following section.

Besides the access rights property, HealthDCAT-AP establishes distinct sets of cardinalities for the metadata elements, tailored to the different level of sensitivity of health data, in accordance with the provisions of the EHDS Regulation. A resume of the mandatory, recommended and optional metadata elements for each sensitivity level of health data (e.g. public, restricted, non-public) is provided on Table 2.

[15]The DGA introduces the concept of NSIP (National Single Information Point) metadata, allowing non-open health datasets to be indexed in the European Data Portal. For compliance, HealthDCAT-AP accounts for NSIP metadata elements, to ensure that protected data, while not openly available, remains discoverable under regulated conditions.

Table 2: List of mandatory, recommended and optional HealthDCAT-AP metadata elements for open data, protected or restricted data and sensitive data. Properties reused without modification from DCAT-AP are marked in red, while new and adapted properties (through updated cardinalities) are highlighted in green.

NON-PERSONAL ELECTRONIC HEALTH DATA		PERSONAL ELECTRONIC HEALTH DATA
OPEN (PUBLIC)	PROTECTED (RESTRICTED)	SENSITIVE (NON PUBLIC)
Mandatory properties		
dct:description: rdfs:Literal [1..*] dct:title: rdfs:Literal [1..*] dct:identifier: rdfs:Literal: **xsd:anyURI [1..*]** dcatap:applicableLegislation rdfs:Resource **[1..*]** dcat:theme (dct:subject): skos:Concept **[1..*]** dct:accessRights: dct:RightsStatement **[1..1]** dcat:distribution: dcat:Distribution **[1..*]** **healthdcatap:hdab foaf:Agent [1..1]** **heathdcatap:healthCategory: (dct:subject) skos:Concept [1..*]**	dct:description: rdfs:Literal [1..*] dct:title: rdfs:Literal [1..*] dct:identifier: rdfs:Literal: **xsd:anyURI [1..*]** dcatap:applicableLegislation rdfs:Resource **[1..*]** dcat:theme (dct:subject): skos:Concept **[1..*]** dct:accessRights: dct:RightsStatement **[1..1]** dct:publisher: foaf:Agent **[1..1]** dcat:distribution: dcat:Distribution **[1..*]** **healthdcatap:hdab foaf:Agent [1..1]** **heathdcatap:healthCategory: (dct:subject) skos:Concept [1..*]**	adms:sample: dcat:Distribution **[1..*]** dcat:contactPoint: vcard:Kind **[1..*]** dcat:distribution: dcat:Distribution **[1..*]** dcat:keyword: rdfs:Literal **[1..*]** dcat:theme (dct:subject): skos:Concept **[1..*]** dcatap:applicableLegislation rdfs:Resource **[1..*]** dct:accessRights: dct:RightsStatement **[1..1]** dct:description: rdfs:Literal [1..*] dct:identifier: rdfs:Literal: **xsd:anyURI [1..*]** dct:provenance: dct:ProvenanceStatement **[1..*]** dct:publisher: foaf:Agent **[1..1]** dct:spatial: dct:Location **[1..*]** dct:title: rdfs:Literal [1..*] dct:type: skos:Concept **[1..1]** **dpv:hasPurpose dpv:Purpose [1..*]** **healthdcatap:hdab foaf:Agent [1..1]** **heathdcatap:healthCategory: (dct:subject) skos:Concept [1..*]** **healthdcatap:healthTheme: (dct:subject) skos:Concept [1..*]**
Recommended properties		
dcat:contactPoint: vcard:Kind **[0..*]** dcat:keyword: rdfs:Literal **[0..*]** **[healthdcatap:trustedDataHolder] xsd:boolean [0..1]**	dcat:contactPoint: vcard:Kind **[0..*]** dcat:keyword: rdfs:Literal **[0..*]** **[healthdcatap:trustedDataHolder] xsd:boolean [0..1]**	dcat:landingPage: foaf:Document [0..*] dcat:temporalResolution rdfs:Literal: xsd:duration [0..1] dct:accrualPeriodicity: dct:Frequency [0..1] dct:conformsTo: dct:Standard [0..*] dct:isReferencedBy: rdfs:Resource [0..*] dct:language: dct:LinguisticSystem [0..*] dct:relation: rdfs:Resource [0..*] dct:source: dcat:Dataset [0..*] dct:temporal: dct:PeriodOfTime [0..*] **dpv:hasLegalBasis dpv:LegalBasis [0..*]** **dpv:hasPersonalData dpv:PersonalData [0..*]** **dqv:hasQualityAnnotation dqv:QualityCertificate [0..*]** **healthdcatap:analytics: dcat:Distribution [0..*]** **healthdcatap:hasCodeValues: skos:Concept [0..*]** **healthdcatap:hasCodingSystem dct:Standard [0..*]** **healthdcatap:minTypicalAge rdfs:nonNegativInteger [0..1]** **healthdcatap:maxTypicalAge rdfs:nonNegativInteger [0..1]**

		healthdcatap:numberofRecords rdfs:nonNegativInteger 0..1] healthdcatap:numberofUniqueIndividuals rdfs:nonNegativInteger 0..1] healthdcatap:populationCoverage rdfs:Literal [0..*] [healthdcatap:trustedDataHolder] xsd:boolean [0..1]
Optional properties		

adms:identifier: adms:Identifier [0..*] adms:sample: dcat:Distribution **[0..*]** adms:versionNotes: rdfs:Literal [0..*] dcat:landingPage: foaf:Document [0..*] dcat:qualifiedRelation: dcat:Relationship [0..*] dcat:spatialResolutionInMeters: rdfs:Literal: xsd:decimal [0..1] dcat:temporalResolution rdfs:Literal: xsd:duration [0..1] dct:accrualPeriodicity: dct:Frequency [0..1] **dct:alternative: rdfs:Literal [0..1]** dct:conformsTo: dct:Standard [0..*] dct:creator: foaf:Agent [0..*] dct:hasVersion: dcat:Dataset [0..*] dct:inSeries: dcat:DataSeries [0..*] dct:isReferencedBy: rdfs:Resource [0..*] dct:issued: rdfs:Literal: xsd:date [0..1] dct:language: dct:LinguisticSystem [0..*] dct:modified: rdfs:Literal: xsd:date [0..1] dct:provenance: dct:ProvenanceStatement **[0..*]** dct:publisher: foaf:Agent **[0..1]** dct:relation: rdfs:Resource [0..*] dct:source: dcat:Dataset [0..*] dct:spatial: dct:Location **[0..*]** dct:temporal: dct:PeriodOfTime [0..*] dct:type: skos:Concept [0..1] **dpv:hasLegalBasis dpv:LegalBasis [0..*]** **dpv:hasPersonalData dpv:PersonalData [0..*]** **dpv:hasPurpose dpv:Purpose [0..*]** **dqv:hasQualityAnnotation dqv:QualityCertificate [0..*]** **healthdcatap:analytics: dcat:Distribution [0..*]** **healthdcatap:hasCodeValues: skos:Concept [0..*]** **healthdcatap:hasCodingSystem dct:Standard [0..*]** **healthdcatap:healthTheme: (dct:subject) skos:Concept [0..*]** **healthdcatap:maxTypicalAge rdfs:nonNegativInteger [0..1]** **healthdcatap:minTypicalAge rdfs:nonNegativInteger [0..1]** **healthdcatap:numberofRecords rdfs:nonNegativInteger [0..1]** **healthdcatap:numberofUniqueIndividuals rdfs:nonNegativInteger [0..1]** **healthdcatap:populationCoverage rdfs:Literal [0..*]** **healthdcatap:retentionPeriod dct:PeriodOfTime [0..1]** owl:versionInfo: rdfs:Literal [0..1] prov:qualifiedAttribution prov:attribution [0..*] prov:wasGeneratedBy: prov:Activity [0..*]	adms:identifier: adms:Identifier [0..*] adms:sample: dcat:Distribution **[0..*]** adms:versionNotes: rdfs:Literal [0..*] dcat:landingPage: foaf:Document [0..*] dcat:qualifiedRelation: dcat:Relationship [0..*] dcat:spatialResolutionInMeters: rdfs:Literal: xsd:decimal [0..1] dcat:temporalResolution rdfs:Literal: xsd:duration [0..1] dct:accrualPeriodicity: dct:Frequency [0..1] **dct:alternative: rdfs:Literal [0..1]** dct:conformsTo: dct:Standard [0..*] dct:creator: foaf:Agent [0..*] dct:hasVersion: dcat:Dataset [0..*] dct:inSeries: dcat:DataSeries [0..*] dct:isReferencedBy: rdfs:Resource [0..*] dct:issued: rdfs:Literal: xsd:date [0..1] dct:language: dct:LinguisticSystem [0..*] dct:modified: rdfs:Literal: xsd:date [0..1] dct:provenance: dct:ProvenanceStatement **[0..*]** dct:relation: rdfs:Resource [0..*] dct:source: dcat:Dataset [0..*] dct:spatial: dct:Location **[0..*]** dct:temporal: dct:PeriodOfTime [0..*] dct:type: skos:Concept [0..1] **dpv:hasLegalBasis dpv:LegalBasis [0..*]** **dpv:hasPersonalData dpv:PersonalData [0..*]** **dpv:hasPurpose dpv:Purpose [0..*]** **dqv:hasQualityAnnotation dqv:QualityCertificate [0..*]** **healthdcatap:analytics: dcat:Distribution [0..*]** **healthdcatap:hasCodeValues: skos:Concept [0..*]** **healthdcatap:hasCodingSystem dct:Standard [0..*]** **healthdcatap:healthTheme: (dct:subject) skos:Concept [0..*]** **healthdcatap:maxTypicalAge rdfs:nonNegativInteger [0..1]** **healthdcatap:minTypicalAge rdfs:nonNegativInteger [0..1]** **healthdcatap:numberofRecords rdfs:nonNegativInteger [0..1]** **healthdcatap:numberofUniqueIndividuals rdfs:nonNegativInteger [0..1]** **healthdcatap:populationCoverage rdfs:Literal [0..*]** **healthdcatap:retentionPeriod dct:PeriodOfTime [0..1]** owl:versionInfo: rdfs:Literal [0..1] prov:qualifiedAttribution prov:attribution [0..*] prov:wasGeneratedBy: prov:Activity [0..*]	adms:identifier: adms:Identifier [0..*] adms:versionNotes: rdfs:Literal [0..*] dcat:qualifiedRelation: dcat:Relationship [0..*] dcat:spatialResolutionInMeters: rdfs:Literal: xsd:decimal [0..1] **dct:alternative: rdfs:Literal [0..1]** dct:creator: foaf:Agent [0..*] dct:hasVersion: dcat:Dataset [0..*] dct:inSeries: dcat:DataSeries [0..*] dct:issued: rdfs:Literal: xsd:date [0..1] dct:modified: rdfs:Literal: xsd:date [0..1] **healthdcatap:retentionPeriod dct:PeriodOfTime [0..1]** owl:versionInfo: rdfs:Literal [0..1] prov:qualifiedAttribution prov:attribution [0..*] prov:wasGeneratedBy: prov:Activity [0..*]

7 Extending DCAT-AP for health: HealthDCAT-AP

This section presents a comprehensive overview of the enrichments made to DCAT-AP, categorised into two groups: 1) properties reused from DCAT-AP but refined with additional usage notes or customised cardinalities, and 2) properties newly introduced by HealthDCAT-AP to address requirements not previously covered by DCAT-AP. To illustrate these adaptations, Table 3 describes the HealthDCAT-AP new properties and adapted properties (through updated cardinalities) highlighted in green, while properties reused without modification from DCAT-AP are displayed In black. It also presents the metadata properties into mandatory, recommended and optional groups. This information is also presented in the HealthDCAT-AP UML diagram (see in Annex II).

Table 3: HealthDCAT-AP metadata elements organised by mandatory, recommended and optional properties to describe personal electronic health data (NON_PUBLIC). Properties reused without modification from DCAT-AP are marked in black, while new and adapted properties (through updated cardinalities) are highlighted in green.

HEALTH DCAT-AP METADATA ELEMENTS FOR PERSONAL ELECTRONIC HEALTH DATA (NON_PUBLIC)		
Mandatory metadata	**Recommended metadata**	**Optional metadata**
adms:sample: dcat:Distribution [1..*] dcat:contactPoint: vcard:Kind [1..*] dcat:distribution: dcat:Distribution [1..*] dcat:keyword: rdfs:Literal [1..*] dcat:theme (dct:subject): skos:Concept [1..*] dcatap:applicableLegislation rdfs:Resource [1..*] dct:accessRights: dct:RightsStatement [1..1] dct:description: rdfs:Literal [1..*] dct:identifier: rdfs:Literal: xsd:anyURI [1..*] dct:provenance: dct:ProvenanceStatement [1..*] dct:publisher: foaf:Agent [1..1] dct:spatial: dct:Location [1..*] dct:title: rdfs:Literal [1..*] dct:type: skos:Concept [1..1] **dpv:hasPurpose dpv:Purpose [1..*]** **healthdcatap:hdab foaf:Agent [1..1]** **heathdcatap:healthCategory: (dct:subject) skos:Concept [1..*]** **healthdcatap:healthTheme: (dct:subject) skos:Concept [1..*]**	dcat:landingPage: foaf:Document [0..*] dcat:temporalResolution rdfs:Literal: xsd:duration [0..1] dct:accrualPeriodicity: dct:Frequency [0..1] dct:conformsTo: dct:Standard [0..*] dct:isReferencedBy: rdfs:Resource [0..*] dct:language: dct:LinguisticSystem [0..*] dct:relation: rdfs:Resource [0..*] dct:source: dcat:Dataset [0..*] dct:temporal: dct:PeriodOfTime [0..*] **dpv:hasLegalBasis dpv:LegalBasis [0..*]** **dpv:hasPersonalData dpv:PersonalData [0..*]** **dqv:hasQualityAnnotation dqv:QualityCertificate [0..*]** **healthdcatap:analytics: dcat:Distribution [0..*]** **healthdcatap:hasCodeValues: skos:Concept [0..*]** **healthdcatap:hasCodingSystem dct:Standard [0..*]** **healthdcatap:minTypicalAge rdfs:nonNegativInteger [0..1]** **healthdcatap:maxTypicalAge rdfs:nonNegativInteger [0..1]** **healthdcatap:numberofRecords rdfs:nonNegativInteger 0..1]** **healthdcatap:numberofUniqueIndividuals rdfs:nonNegativInteger 0..1]** **healthdcatap:populationCoverage rdfs:Literal [0..*]** **[healthdcatap:trustedDataHolder] xsd:boolean [0..1]**	adms:identifier: adms:Identifier [0..*] adms:versionNotes: rdfs:Literal [0..*] dcat:qualifiedRelation: dcat:Relationship [0..*] dcat:spatialResolutionInMeters: rdfs:Literal: xsd:decimal [0..1] **dct:alternative: rdfs:Literal [0..1]** dct:creator: foaf:Agent [0..*] dct:hasVersion: dcat:Dataset [0..*] dct:inSeries: dcat:DataSeries [0..*] dct:issued: rdfs:Literal: xsd:date [0..1] dct:modified: rdfs:Literal: xsd:date [0..1] **healthdcatap:retentionPeriod** **dct:PeriodOfTime [0..1]** dcat:version: rdfs:Literal [0..1] prov:qualifiedAttribution prov:attribution [0..*] prov:wasGeneratedBy: prov:Activity [0..*]

The new properties introduced by HealthDCAT-AP primarily focus on the DCAT-AP class Dataset, which encompasses metadata about the dataset itself. For instance, the Dataset class in DCAT-AP includes information such as the dataset's title, a description in any language, the author, the category to which the dataset belongs, and much more.

The following sections begin by reviewing the existing DCAT-AP properties of the Dataset class. This is followed by an examination of the "DCAT-AP Adjusted Properties" and the newly introduced "HealthDCAT-AP Properties" with a focus on the personal electronic health data (sensitive - NON_PUBLIC data)

7.1 DCAT-AP properties of the Dataset Class

Title
DCAT definition: A name given to the dataset.
Note: The title should describe the main topic of the dataset in a clear and concise way, ideally in one short sentence. This property can be repeated for parallel language versions of the name.

Example - Title	Technical reference: **[dct:title]**
Linking of registers for COVID-19 vaccine surveillance	

Description
DCAT definition: Free-text property to describe the dataset.
Note: Provide a comprehensive overview of the dataset's content, purpose, and key characteristics, using simple language to ensure clarity for data applicants.

Example - Description	Technical reference: **[dct:description]**
The LINK-VACC project, organised by Sciensano, involves the communication of pseudonymized data related to COVID-19 vaccines. The data includes demographic information (age, gender, postal code, date of death) of individuals who have received at least one dose of a COVID-19 vaccine in Belgium, details about the vaccinator, vaccination location, administered vaccine, and observed side effects. Additionally, the project encompasses data from the HealthData COVID-19 database and Clinic database, which provide demographic and clinical details of individuals tested for or hospitalized with COVID-19, as well as information from the Common Base Registry for HealthCare Actor (CoBRHA), STATBEL, and the Agency for Health Insurance Fund (AIM) database .	

Keywords
DCAT-AP definition: Keywords or tags that describe the dataset.
Note: Choose specific and relevant terms that accurately represent the dataset's content. This helps improve discoverability for users searching for related information.

Example - Keywords	Technical reference: **[dcat:keywords]**
"COVID-19" , "SARS-CoV-2", "corona virus", "vaccine", "vaccine effectiveness", "surveillance"	

Contact Point
DCAT-AP definition: Contact information associated with the dataset.
Note: Include up-to-date contact information for the person or organisation responsible for the dataset. This ensures users can easily reach out for questions or access requests.

Example - Contact Point	Technical reference: **[dcat:contactPoint]**
Organization: Sciensano Address - Country name: Belgium, “BEL” - Locality: Elsene – Ixelles - Postal-Code:B-1050 - Street address:Rue Juliette Wytsmanstraat 14	

URL: https://sciensano.be/
Email: covacsurv@sciensano.be

Landing Page

DCAT-AP definition: A web page that provides access to the dataset, its distributions and/or additional information.
Note: Provide a direct link to a webpage that offers comprehensive information about the dataset and access to its distributions.

Example - Landing Page	Technical reference: **[dcat:landingPage]**
https://sciensano.service-now.com/sp	

Temporal Resolution

DCAT-AP definition: The minimum time period resolvable in the dataset.
Note: Specify the finest time interval that can be distinguished in the dataset, such as hourly, daily, or monthly. The **temporal duration is expressed** in the example in the **ISO 8601 duration format**.

Example - Temporal Resolution	Technical reference: **[dcat:temporalResolution]**
"P1D"^^ <http://www.w3.org/2001/XMLSchema#duration>;	

Accrual Periodicity

DCAT Definition: The frequency at which the dataset is updated.
Note: Indicate how often the dataset is updated, using standard terms like daily, weekly, monthly, or annually from the EU Vocabularies Frequency Named Authority List.

Example - Accrual Periodicity	Technical reference: **[dct:accrualPeriodicity]**
<http://publications.europa.eu/resource/authority/frequency/DAILY>;	

Conforms to

DCAT Definition: An implementing rule or other specification.
Note: List any standards, schemas, or regulations that the dataset adheres to, ensuring data quality and interoperability.

Example - Conforms to	Technical reference: **[dct:conformsTo]**
https://www.wikidata.org/entity/Q19597236;	

Referenced by

DCAT Definition: A related resource, such as a publication, that references, cites or otherwise points to the dataset.
Note: Include citations or links to publications, reports, or other resources that have used or mentioned this dataset.

Example - Referenced By	Technical reference: **[dct:isReferencedBy]**

https://doi.org/10.1186/s13690-021-00709-x;

Language

DCAT Definition: Language of the dataset, to be indicated from the Language Controlled Vocabulary[16].
Note: Clearly state the language(s) used in the dataset to help users assess its relevance and usability.

Example - Language	Technical reference: **[dct:language]**
English - http://publications.europa.eu/resource/authority/language/ENG	

Relation

DCAT Definition: Indicate resources related to the dataset with an unspecified relationship.
Note: Mention any related datasets or resources that might provide context or complementary information.

Example - Relation	Technical reference: **[dct:relation]**
dct:relation <...>;	

Source

DCAT Definition: A related dataset from which the described dataset is derived.
Note: If applicable, identify the original dataset from which this one was derived or processed.

Example - Source	Technical reference: **[dct:source]**
<https://fair.healthdata.be/dataset/e703dd1e-0bca-438d-a7cb-d3429dc6d118>, <https://fair.healthdata.be/dataset/6777a689-7fd7-49a0-adc7-d28fef161843>;	

Temporal

DCAT Definition: A temporal period that the dataset covers.
Note: Specify the time period covered by the dataset, using precise start and end dates.

Example - Temporal	Technical reference: **[dct:temporal]**
End Date "2024-12-31" <http://www.w3.org/2001/XMLSchema#date>; Start Date "2021-01-01" <http://www.w3.org/2001/XMLSchema#date>	

Other Identifier

A secondary identifier of the dataset, such as MAST/ADS, DataCite, DOI, EZID or W3ID.
Note: Include any additional unique identifiers associated with the dataset to improve discoverability.

Example - Other Identifier	Technical reference: **[adms:identifier]**
"HDBP0250"	

[16] Language - EU Vocabularies - Publications Office of the EU

Version

DCAT-AP definition: The version indicator (name or identifier) of the dataset.
Note: Provide a clear version number or identifier to help users track changes and cite the correct dataset version.

Example - Version	Technical reference: **[dcat:version]**
Version "Project HDBP0250"	

Has Version

DCAT-AP definition: A related dataset that is a version, edition or adaptation of the described dataset.
Note: If applicable, link to other versions of this dataset, such as updates or adaptations.

Example - Has Version	Technical reference: **[dcat:hasVersion]**
<https://.../dataset/91ca0886-0090-497a-8fc0-bd3055d72191>;	

Version Notes

A description of the differences between this version and a previous version of the dataset.
Note: Briefly describe what has changed in this version compared to previous ones, highlighting key updates or improvements.

Example - Version Notes	Technical reference: **[adms:versionNotes]**
Dataset continuously updated;	

Issued

DCAT definition: The date of formal issuance (e.g.: publication) of the dataset (Not to be confused by date of creation of the metadata records itself).
Note: State the official release date of the dataset, which may differ from its creation or last modification date.

Example - Issued	Technical reference: **[dct:issued]**
Issued on 2023-08-19	

Modified

DCAT definition: The most recent date on which the dataset was changed or modified.
Note: Indicate the most recent date when any changes were made to the dataset, helping users identify updates.

Example - Modified	Technical reference: **[dct:modified]**
Modified on 2024-12-31	

Creator

DCAT definition: An entity responsible for producing the dataset.
Note: Name the person, group, or organisation primarily responsible for producing the dataset.

Example - Creator	Technical reference: **[dct:creator]**

Example when URI is available:
<https://org.belgif.be/id/CbeRegisteredEntity/0693876830>;

Example when URI is not available:
Creator
Agent:
- Name: Sciensano
- Homepage: http://sciensano.be
- Email: info@sciensano.be

Generated By

An activity that generated, or provides the business context for, the creation of the dataset. **Note:** Describe the process or project that led to the creation of this dataset, providing context for its existence.

Example - Generated By	Technical reference: **[prov:wasGeneratedBy]**

Activity type: http://dbpedia.org/resource/Record_linkage ;
Page: https://www.sciensano.be/sites/default/files/information_letter_linkvacc_20221010_en.pdf
Label: "Pseudonymisation and Anonymisation"
Started at time "2021-01-01"

Associated with:
Agent
- Name First name, last name
- Mail: Email address
- Page: https://www.sciensano.be/fr/people/person
- Affiliation: Organization, name "Sciensano"

Qualified Attribution

Definition: An Agent having some form of responsibility for the dataset.
Note: List any individuals or organisations that contributed significantly to the dataset's creation or maintenance.

Example - Qualified Attribution	Technical reference: **[prov:qualifiedAttribution]**

Agent
- Type: Organization
- Name "healthdata.be (Sciensano)"
- Address: Rue Juliette Wytsmanstraat 14, 1050, Elsene - Ixelles, Brussels, Belgium
- Homepage: https://healthdata.be
- Email: healthdata@sciensano.be
- Phone <tel:+3227930142>

Spatial Resolution in Meters

DCAT definition: The minimum spatial separation resolvable in a dataset, measured in meters. **Note:** For geospatial datasets, specify the smallest spatial detail that can be reliably distinguished, measured in meters.

Example - Spatial Resolution in Meters	Technical reference: **[dcat:spatialResolutionInMeters]**
"10"	

7.2 DCAT-AP Adjusted Properties

Access Rights

DCAT definition: Information that indicates whether the Dataset is publicly accessible, has access restrictions or is not public.
Note: Clearly specify the dataset's accessibility status, indicating whether it's public, restricted, or non-public. Use predefined values from the Access Rights Authority List for datasets under EHDS Regulation: For personal electronic health data, use 'NON_PUBLIC' from the Access Rights Vocabulary. It's mandatory property in HealthDCAT-AP.

Example - Access Rights	Technical reference: **[dct:accessRights]**
Non-public - http://publications.europa.eu/resource/authority/access-right/NON_PUBLIC>;	

Applicable Legislation

The Applicable Legislation property has been set to mandatory in HealthDCAT-AP to acknowledge datasets that are in the scope of EHDS Regulation, as each dataset identified by an HDAB as being within the scope of the EHDS Regulation **must include the European Legislation Identifier (ELI) of the EHDS Regulation in the Applicable Legislation property**. This property is important to help data users understand the legal frameworks and reuse conditions applicable to health datasets.

DCAT-AP definition: The legislation that mandates the creation or management of the Dataset.
Note: Specify all relevant legal frameworks that govern the dataset's creation, management, and use. It may include multiple legislations (e.g., EHDS Regulation, INSPIRE Directive). For datasets within EHDS scope, always include the ELI of the EHDS Regulation.

Example - Applicable Legislation	Technical reference: **[dcatap:applicableLegislation]**
Applicable Legislation <http://data.europa.eu/eli/reg/2022/868/oj>	

Identifier

In HealthDCAT-AP, it is mandatory to use Persistent Uniform Resource Identifiers (PURIs) as identifiers for metadata records and datasets. These persistent identifiers serve as **stable and unique references, pointing directly to the metadata or dataset description**. In practice, this means that each metadata record in HealthDCAT-AP is assigned a PURI that links back to the primary datasets catalogue, ensuring that any changes are accurately reflected across all platforms, maintaining the integrity of the metadata. This approach aligns with DCAT-AP best practices and facilitates precise linking, retrieval, and interoperability of health data.

DCAT definition: The main identifier for the Dataset, e.g. the URI or other unique identifier in the context of the Catalogue.

Note: Assign a Persistent Uniform Resource Identifier (PURI) as the main identifier for the dataset. This PURI should directly link to the metadata or dataset description in the primary catalogue.

Example - Identifier	Technical reference: **[dct:identifier]**
Identifier: "https://fair.healthdata.be/dataset/d43a158e-7d13-4660-bbc3 9d3f8d5501e5"	

Distribution

In HealthDCAT-AP, it is defined as mandatory that **every metadata record includes at least one dataset distribution**. For personal electronic health data (sensitive data), this must take the form of an HDAB landing page that provides access to the Data Service and/or additional information for accessing the Dataset. The Applicable Legislation property of the Dataset Distribution must reference the EHDS Regulation's ELI. For protected and sensitive health datasets, HealthDCAT-AP also requires the inclusion of specific properties in the Distribution, such as an access URL, byte size, format and rights, to comply with DGA guidelines.

DCAT-AP definition: An available Distribution for the Dataset.
Note: Provide at least one distribution for every dataset, tailored to its access rights. For personal electronic health data, include a mandatory HDAB landing page distribution.

Example - Distribution	Technical reference: **[dcat:distribution]**
Distribution: - Applicable Legislation <http://data.europa.eu/eli/reg/2022/868/oj>; - AccessURL <https://hda.be> ; - Format <http://publications.europa.eu/resource/authority/file-type/XML>; - Byte Size "800000"; - Rights Statement "Access to data is conditional on the issuance of a permit by the HDAB after submission of a data request application" - Media Type <http://www.iana.org/assignments/media-types/text/tab-separated-values>	

Geographical Coverage

The Geographical Coverage property defines the geographical area represented by the dataset. This property serves as a critical filter, particularly when analysing datasets that pertain to specific populations. While accurately specifying geographical coverage can be technically challenging due to the range of possible representations, the use of, at least, a controlled vocabulary is recommended to ensure consistency and interoperability across datasets' descriptions. Multiple representations are recommended (i.e., GEONAMES or geographic coordinates).

DCAT-AP definition: A geographic region that is covered by the Dataset.
Note: Specify the geographic area represented by the dataset using standardised terms, identifiers from controlled vocabularies and geographic coordinates.

Example - Geographical Coverage	Technical reference: **[dct:spatial]**
Belgium - *http://publications.europa.eu/resource/authority/country/BEL*	

Provenance

The Provenance property has been designated as mandatory in HealthDCAT-AP and is implemented in a free-text format to provide transparency on the origin and history of a dataset. It allows users to understand how a dataset was created, modified and by whom—key information for evaluating its quality, relevance and overall usability.

DCAT definition: A statement about the lineage of a Dataset.
Note: Provide a clear, concise statement describing the dataset's origin, creation process, and any significant modifications. Include information about the entities involved in its production and maintenance as well as key milestones in the dataset's history that may impact its interpretation or application.

Example - Provenance	Technical reference: **[dct:provenance]**
Provenance Statement: The data for the LINK-VACC project is sourced from several existing databases, including Vaccinnet+, HealthData COVID-19 database(Contact tracing and Clinic database), CoBRHA, STATBEL, and the AIM database. These databases collectively provide comprehensive demographic, clinical, and socio-economic data relevant to the project's objectives	

Publisher

The Publisher property is mandatory in HealthDCAT-AP, also due to the requirements of DGA. Serving as the data holder, the publisher is responsible for ensuring the dataset is made accessible and available to users. This mandatory requirement reinforces accountability and transparency, clearly identifying the entity that manages and provides access to the dataset. The publisher Agent class has also received two new properties specific to healthDCAT-AP: healthdcatap:publisherType and healthdcatap:publisherNote.

DCAT definition: An entity (organisation) responsible for making the Dataset available.
Note: Specify the organisation responsible for making the dataset available. This entity acts as the data holder and ensures dataset accessibility.

Publisher Type

The Publisher Type property identifies the type of organisation responsible for making the dataset available. A controlled vocabulary is being developed within the scope of TEHDAS2 WP5 to denote commonly recognised types of health data publishers, ensuring clarity and consistency in identifying the type of organisation behind the dataset. This categorisation helps data users understand the context and credibility of the dataset provider.

HealthDCAT-AP definition: A type of organisation that makes the Dataset available.
Note: Specify the category of organisation responsible for making the dataset available, using terms from the controlled vocabulary being developed in TEHDAS2 WP5.

Publisher Note

The Publisher Note is a free-text property added to HealthDCAT-AP, allowing the publisher to provide a description of their activities and scope. This property offers valuable context regarding

the publisher's role, helping users better understand the dataset. By providing insight into the publisher's operations, the note enhances transparency and supports users in evaluating the dataset's reliability and relevance.

HealthDCAT-AP definition: A description of the publisher activities.
Note: Provide a concise description of the publisher's activities, scope, and role in relation to the dataset.

Example – Publisher	Technical reference: **[dct:publisher]**

Agent:
- Name "Sciensano"
- Email *info@sciensano.be*
- Homepage *https://sciensano.be*
- Publisher Type <.../authority-table/publisher-type/NationalPublicHealthInstitut> **[healthdcatap:publisherType]**
- Publisher Note "Sciensano is a research institute and the national public health institute of Belgium. It is a so-called federal scientific institution that operates under the authority of the federal minister of Public Health and the federal minister of Agriculture of Belgium." **[healthdcatap:publisherNote]**

Sample

For sensitive data (NON_PUBLIC), the "Sample" property has been defined as mandatory. For such datasets, a Sample Distribution must be provided, which should consist of a data dictionary providing a representation of the dataset structure. An additional Distribution could consist of an anonymised or synthetic data to offer a representation of the original dataset that retains the essential characteristics of the dataset while ensuring the protection of sensitive information. This way, the sample property provides users with a safe-to-share representation of the dataset, enabling them to understand its structure and characteristics without exposing sensitive information. The data dictionary must describe clearly and comprehensively the structure and content of the dataset including data structures with data formats, vocabularies, classification schemes, taxonomies and code lists). If the dataset is subject to intellectual property (IP) rights, a redacted version of the data dictionary may be provided. The data dictionary must be published in a machine-readable and actionable format, specifically CSVW, an extension of the widely used CSV format.

Definition: A sample distribution of the dataset.
Note: Provide a representative subset of the dataset that retains essential characteristics while protecting sensitive information.

Examples - Sample	Technical reference: **[adms:sample]**

Sample Distribution - Example 1 (Synthetic data)

- Description: "Proxy data generating for the EHDS2 Pilot project Sciensano Use Case"
- Download URL: *https://github.com/CAVDgit/EHDS2_UC_Sciensano/blob/main/use_case_1_synthetic_data_10K_individuals.csv;*
- Media Type: <http://www.iana.org/assignments/media-types/text/tab-separated-values>

Sample Distribution - Example with RDF data dictionary

- Distribution description: "Structural metadata expressed using the csvw RDF vocabulory";
- Download URL: <https://www.healthinformationportal.eu/health-information-sources/link-vacc>;
- Media Type: "http://www.iana.org/assignments/media-types/text/turtle"

Theme

The Theme property in DCAT-AP ensures that datasets are consistently categorised across the common European data spaces. For all datasets falling within the scope of the EHDS Regulation, it is mandatory that the Theme property is assigned the specific value http://publications.europa.eu/resource/authority/data-theme/HEAL.

DCAT-AP definition: A category of the Dataset.
Note: Assign the appropriate category to the dataset using the Dataset Theme Controlled Vocabulary maintained by the Publications Office. For health datasets within the EHDS Regulation scope, use the mandatory value http://publications.europa.eu/resource/authority/data-theme/HEAL.

Example - Theme	Technical reference: **[dcat:theme]**
Theme: <http://publications.europa.eu/resource/authority/data-theme/HEAL>	

Type

The Type property in HealthDCAT-AP plays a pivotal role in distinguishing between different classifications of electronic health data, particularly in compliance with Article 51 of the EHDS Regulation. To achieve this, HealthDCAT-AP uses the [dct:type] property in conjunction with the Dataset-Type Controlled Vocabulary[17] maintained by the Publications Office. This controlled vocabulary includes categories such as geospatial data, statistical data, and ontologies, and it should be expanded to include a specific entry for "PERSONAL_DATA". Health data holders are mandated to use this value to clearly identify datasets containing personal electronic health data, ensuring accurate classification and a clear understanding of the dataset's accessibility and sensitivity.

DCAT definition: A type of the Dataset (A recommended controlled vocabulary data-type is foreseen. For health datasets containing personal level information, the type of the dataset MUST take the value "personal data").
Note: Specify the dataset's classification using the Dataset-Type Controlled Vocabulary maintained by the Publications Office. For instance, for datasets containing personal electronic health data, use the mandatory value "PERSONAL_DATA".

Example - Type	Technical reference: **[dct:type]**
Personal data - <http://publications.europa.eu/resource/authority/dataset-type/PERSONAL_DATA>;	

7.3 HealthDCAT-AP new properties

Purpose

[17] Dataset type - EU Vocabularies - Publications Office of the EU

The Purpose property was introduced in HealthDCAT-AP to clearly articulate the intention or objective behind the collection and use of the sensitive data.

Definition: A free text statement of the purpose of the processing of data or personal data.
Note: Clearly state the intention or objective behind the dataset's collection and use in a free-text format.

Example - Purpose	Technical reference: **[dpv:hasPurpose]**
Purpose description: The primary objective of Sciensano's LINK-VACC project is to monitor COVID-19 vaccines post-authorization and evaluate the public health value of prioritizing vaccination for people with comorbidities. This involves assessing the vaccines' effectiveness and safety in the broader population context, beyond the limited scope of clinical trials, and determining future vaccination policies in public health emergencies such as epidemics or pandemics.	

Legal Basis

The Legal Basis property in HealthDCAT-AP has been introduced to specify the legal or regulatory basis under which data is collected and processed. In the context of personal data, the GDPR mandates that any data collection must have a valid legal basis. Including this information in metadata helps data users understand the foundational premise of the dataset's collection and its potential for secondary use, ensuring compliance and transparency. Please note that this legal basis is related to the initial processing of data and does not affect the processing of data under the EHDS Regulation.

Definition: The legal basis used to justify processing of personal data.
Note: Specify the legal or regulatory foundation for data collection and processing. This property helps users understand the dataset's compliance and potential for secondary use.

Example - Legal Basis	Technical reference: **[dpv:hasLegalBasis]**
Legal Basis description: "CSI Deliberation no. 21/028 of february 18, 2021, last amended on june 18, 2021, relating to the communication of data to pseudonymized personal character relating to the health of vaccinnet+, healthdata covid-19 database i and ii, healthdata covid-19 clinical database, cobrha, statbel and the agency intermutualist in sciensano, as part of the link-vacc project and the subsequent processing of personal data pseudonymised by the federal drug agency in view monitoring the safety of covid-19 vaccines"; Source: https://www.ehealth.fgov.be/ehealthplatform/file/view/AXkNfdPml9vUUfvGGfJr?filename=21-028-f212-AFMPS-vaccinnet-modifi%C3%A9e%20le%2018%20juin%202021.pdf>, https://www.ehealth.fgov.be/ehealthplatform/file/view/AX_-9sZSuwVJMAnC0ENo?filename=21-028-f166-LINK-VACC-modifi%C3%A9e%20le%205%20avril%202022.pdf>	

Personal Data

The Personal Data property in HealthDCAT-AP provides a detailed description of the key elements within a dataset that contribute to its classification as sensitive data, particularly those elements containing personal information. This approach enhances filtering capabilities, enabling users to perform more precise and targeted searches for relevant datasets.

Definition: Key elements that represent an individual in the dataset.
Note: Specify the types of personal data present in the dataset.

Example - Personal Data	Technical reference: **[dpv:hasPersonalData]**

Here are some examples of personal data elements taken out of the code list:
Gender, Age, Location, Nationality, Education, Health Record Data, Contact information

Population Coverage

The Population Coverage property is important for understanding the scope and context of the dataset, specifically when it pertains to health data involving a defined population. It ensures that users can gain clarity on the demographic or population group represented, making it easier to assess the dataset's relevance and applicability. This property is only applicable to datasets that include a population.

Definition: A definition of the population within the dataset.
Note: Specify the key elements that represent an individual in the dataset, using the Personal Data Categories from the DPV ontology for precise classification.

Example - Population Coverage	Technical reference: **[healthdcatap:populationCoverage]**
Population coverage: "The population targeted by the LINK-VACC project comprises all individuals in Belgium who have received a COVID-19 vaccine, undergone testing for COVID-19, or have been hospitalized with a confirmed diagnosis of COVID-19. The project also considers healthcare professionals and the general Belgian population for understanding vaccination coverage and effectiveness, especially among those with comorbidities and varying socio-economic backgrounds"	

HealthDCAT-AP has introduced new properties that use Wikidata as a large authority vocabulary. By aligning these properties with Wikidata's HTTP URIs, HealthDCAT-AP enhances consistency and improves semantic annotation, consequently improving interoperability across health data systems. Properties based on Wikidata vocabulary include:

Health Theme

The Health Theme property is a categorisation or tag for a dataset, similar to keywords but designed to be machine-readable and actionable. By relying on Wikidata as a semantic knowledge base, this property enhances the discoverability and usability of datasets, ensuring that datasets are categorised in a way that can be understood and used by automated systems.

Health DCAT-AP definition: A category of the Dataset or tag describing the Dataset.
Note: Use Wikidata entities to categorise or tag the dataset with relevant health-related themes, enhancing machine-readability and discoverability. It works as keywords but for providing concepts for machines.

Example - Health Theme	Technical reference: **[healthdcatap:healthTheme]**
Health Theme: https://www.wikidata.org/entity/Q58624061 , https://www.wikidata.org/entity/Q7907952	

Coding System

This property indicates the coding system used within a dataset, providing crucial information for its reuse and discoverability. For example, if a dataset uses ICD-10 for disease classification, this

property allows data users to search for datasets with the same coding system. As a machine-actionable property, it also facilitates automated processes, making dataset discovery more efficient.

HealthDCAT-AP definition: Coding systems in use (ex: ICD-10-CM, DGRs, SNOMED=CT, ...). **Note:** Specify the standardised coding systems used within the dataset, such as ICD-10-CM, DRGs, or SNOMED-CT, using Wikidata concept URIs.

Example - Coding System	Technical reference: **[healthdcatap:hasCodingSystem]**
Coding system: https://www.wikidata.org/entity/P1690, https://www.wikidata.org/entity/P4229	

Code Values

The Code Values property complements the coding system property by detailing the specific codes used within a dataset. It improves the discoverability of datasets by enabling searches for specific diseases or conditions using coding systems like the ICD-10 example provided. Like the Coding System property, this is machine-actionable, allowing for automated searches and processes to enhance dataset accessibility.

HealthDCAT-AP definition: Health classifications and their codes associated with the dataset. **Note:** Specify the specific health classification codes used within the dataset, enhancing discoverability and enabling precise searches for particular diseases or conditions.

Example - Code Values	Technical reference: **[healthdcatap:hasCodeValues]**
The coding system is ICD-10 and the identifiers are Y59.0 and U07.1: Identifier: https://www.wikidata.org/entity/P494 Label: International Classification of Diseases, 10th Revision (ICD-10) Definition: ICD-10 is a medical classification list by the World Health Organization. Notation: ICD-10 Version Info: Version 2019 Identifier: https://icd.who.int/browse10/2019/en#/Y59.0 Notation: "Y59.0"; Label: "Viral vaccines" Identifier : https://icd.who.int/browse10/2019/en#/U07.1 Notation: "U07.1"; Label: "COVID-19, virus identified"	

Health Category

HealthDCAT-AP introduces the Health Category property to indicate the specific category to which a dataset belongs, as defined in Article 51 of the EHDS Regulation proposal. This property relies on a controlled vocabulary, which must be used to ensure standardisation and consistency across datasets. As the categorisation process can be challenging for health data holders, TEHDAS2 is supporting the creation of short labels for the categories and a full text definition for each category defined in the EHDS Regulation (see also Chapter 8).

HealthDCAT-AP definition: The health category to which this dataset belongs as described in the Commission Regulation on the European Health Data Space laying down a list of categories of electronic data for secondary use.
Note: Specify the dataset's category using the controlled vocabulary defined in Article 51 of the EHDS Regulation proposal.

Example - Health Category	Technical reference: **[healthdcatap:healthCategory]**
Health category: http://healthdata.ec.europa.eu/.../(o)	

Health Data Access Body

The EHDS Regulation introduces health data access bodies, which serve as a gateway for data access, overseeing requests and ensuring compliance with data protection regulations. Including this information ensures transparency and facilitates compliance with EHDS requirements for cataloguing health datasets and managing data access.

HealthDCAT-AP definition: Health data access body supporting access to data in the Member State.
Note: Specify the health data access body (HDAB) responsible for managing access to the dataset.

Example - Health Data Access Body	Technical reference: **[healthdcatap:hdab]**
Organization: - Name: "Belgian Health Data Agency" - Address: Galileelaan 5, Bus 2, Saint-Josse-ten-Noode, 1210, BEL - Email: info@hda.fgov.be - Homepage: *https://www.hda.belgium.be* **When the HDAB registry is published:** HDAB name (will all the associated information)	

Trusted Data Holder [healthdcatap:trustedDataHolder]

Next to HDABs, the EHDS Regulation also introduces the new role of Trusted Health Data Holder (see also EHDS Regulation Article 71). The HealthDCAT-AP property questions whether or not health data holder is a trusted health data holder.

Quality Annotation

The Quality Annotation property is introduced in HealthDCAT-AP to include information on any applicable quality and utility labels for a dataset, specified in Article 78 of the EHDS Regulation.

Definition: A statement related to quality of the Dataset, including rating, quality certificate, feedback that can be associated to the dataset.
Note: Include a data quality and utility label for the dataset, as defined in EHDS Regulation. This label should provide a graphic representation of the dataset's quality and conditions of use, reflecting criteria such as findability, accessibility, interoperability, and reusability based on FAIR principles

Example - Quality Annotation	Technical reference: **[dqv:hasQualityAnnotation]**
Quality Certificate: - Target: https://fair.healthdata.be/dataset/d43a158e-7d13-4660-bbc3-9d3f8d5501e5	

- Body: https://acertificateserver.eu/mycertificate
- Purpose: quality assessment

Number of Unique Individuals

The Number of Unique Individuals property is a new facet introduced in Health DCAT-AP to provide quantitative information about the dataset. This property not only informs about the size of the dataset but also supports users in estimating its potential value since by understanding the scale of the dataset in terms of the number of individuals represented, users can better assess its utility and relevance for their analysis or research.

Definition: Number of records for unique individuals.
Note: Specify the average total count of unique individuals represented in the dataset, helping data applicants to assess the dataset's size and potential value for analysis or research.

Example - Number of Unique Individuals	Technical reference: **[healthdcatap:numberOfUniqueIndividuals]**
Number of unique individuals: "8914700"	

Age Ranges

HealthDCAT-AP has introduced two new properties related to age ranges: Maximum Typical Age and Minimum Typical Age. These facets are particularly relevant for population-based datasets and offer useful filters for exploring health datasets in the catalogues, allowing users to understand the typical age coverage of individuals in a dataset.

Definition: Minimum typical age of the population within the dataset.
Note: Specify the minimum typical age of the population represented in the dataset.

Example - Min typical age	Technical reference: **[healthdcatap:minTypicalAge]**
Minimal typical age: "18"	

Definition: Maximum typical age of the population within the dataset.
Note: Specify the maximum typical age of the population represented in the dataset.

Example - Max typical age	Technical reference: **[healthdcatap:maxTypicalAge]**
Maximal typical age: "64"	

Analytics

This property enables health data holders to link datasets to associated resources such as visualisations, analytics dashboards, technical reports, quality and usability indicators and querying services like APIs. These linked resources provide insights and metrics about the dataset, allowing users to explore and comprehend its content without directly accessing the underlying data.

The property can link to external resources, such as:

- Dashboards: advanced interfaces offering interactive analytics;
- Technical reports: statistical summaries and dataset metrics;

- APIs: tools for querying high-level statistics, such as the Beacon API.

Definition: An analytics distribution of the dataset.
Note: Provide URLs pointing to data discovery services or document repositories where users can access or request associated resources such as dashboards, technical reports of the dataset, quality measurements, usability indicators or analytics services.

Example - Analytics	Technical reference: **[healthdcatap:analytics]**

Example of the Antimicrobial resistance surveillance data in the EU/EEA - ECDC (EARS-Net) mapping tool:

Title: Surveillance Atlas of Infectious Diseases (en)
Description: The Surveillance Atlas of Infectious Diseases is a tool that interacts with the latest available data about a number of infectious diseases. The interface allows users to interact and manipulate the data to produce a variety of tables and maps. The information contained in the dataset provided through ATLAS is made available by ECDC collating data from the Member States collected through The European Surveillance System (TESSy).
URL to access the tool: Surveillance Atlas of Infectious Diseases

Number of Records

The Number of Records property is strongly recommended to provide additional quantitative information about a dataset. By including the Number of Records, data users can better assess the dataset's potential utility and relevance.

Definition: Size of the dataset in terms of the number of records.
Note: Specify the total count of records in the dataset, or alternatively, provide an average when exact numbers are unavailable. This property is strongly recommended as it offers valuable quantitative insights, helping users assess the dataset's potential utility and relevance.

Example - Number of Records	Technical reference: **[healthdcatap:numberOfrecords]**

Number of records: "124866488"

Retention Period

The Retention Period property defines the temporal period during which a dataset is available for secondary use. This property becomes mandatory if the dataset are not more available for reuse after a certain period of time. While the dataset itself may be removed, its metadata record must remain accessible, ensuring that any references or citations to the dataset in publications remain valid.

Definition: A temporal period during which the dataset is available for secondary use.
Note: Specify the time frame during which the dataset can be used for secondary purposes.

Example - Retention Period	Technical reference: **[healthdcatap:retentionPeriod]**

Retention period:

- End Date "2034-12-31"
- Start Date "2020-03-01"
- Comment: "Provide complementary information"

8 Next steps towards a complete guideline

TEHDAS2 WP 5 aims to facilitate the discovery of datasets for secondary use purposes. Task 5.1 is set to build a set of requirements towards health data holders to fulfil their duties for data description and addresses the question: **Which data should be described by the health data holder and how?** Key elements of these requirements are: 1) the scope of the EHDS; data holders need to be able to identify datasets that are covered by the categories of health data described in the Regulation, and 2) HealthDCAT-AP to describe these datasets.

To this end, task 5.1 organises two workshop series:

1) A workshop series on **Article 51**, designed to provide clarification on the categories of health data in scope of the EHDS. The goal of these meetings is to work on full text explanations of the categories, include examples and set short label titles for the categories. The full workshop series can be found in Annex III.

2) A series of technical working groups dedicated to validating and finetuning specific elements of the HealthDCAT-AP. The full workshop series can be found in Annex IV.

Furthermore, a *big data* analysis on existing health datasets will be conducted to shed light on the categories of Article 51. The https://data.europa.eu/en is already hosting thousands of metadata records pertaining to health. In addition, there are already many data sources findable online. This analysis will scrape information from https://data.europa.eu/en and the internet and will use advanced methods to arrive at good examples and classify the themes of health data out there. This will aid in clarifying the minimum categories in Article 51. This also allows finding indications for data type and publisher type.

Finally, the European Health Information Portal, hosted by Sciensano, is glad to offer to all health data holders the possibility to create a HealthDCAT-AP compliant record of their dataset, in RDF formats, through its HealthDCAT-AP editor. The HealthDCAT-AP editor aims to help users to create DCAT RDF metadata to describe their datasets.

TEHDAS2 task 5.1 will analyse the outputs from the working group series, the public consultation, the feedback on Github, discussions in the Community of Practice and the submitted metadata records through the HealthDCAT-AP editor in order to arrive at a final guideline on data description duties of health data holders (TEHDAS2 Deliverable 5.1), due summer 2025.

Annex I: Links to the EHDS Regulation

Relevant articles for this guideline:

Article 60: Duties of health data holders

3. The health data holder shall communicate to the health data access body a description of the dataset it holds in accordance with Article 77. The health data holder shall, at a minimum on an annual basis, check that its dataset description in the national dataset catalogue is accurate and up to date.

5. Health data holders of non-personal electronic health data shall provide access to data through trusted open databases to ensure unrestricted access for all users and data storage and preservation. Trusted open public databases shall have in place robust, transparent and sustainable governance and a transparent model of user access.

Article 77: Dataset description and dataset catalogue

1. Health data access bodies shall, through a publicly available and standardised machine-readable dataset catalogue, provide a description in the form of metadata of the available datasets and their characteristics. The description of each dataset shall include information concerning the source, scope, main characteristics, and nature of the electronic health data in the dataset and the conditions for making those data available.

4. By ... [two years from the date of entry into force of this Regulation], the Commission shall, by means of implementing acts, set out the minimum elements health data holders are to provide for datasets and the characteristics of those elements.

Article 78 Data quality and utility label

1. Datasets made available through health data access bodies may have a Union data quality and utility label applied by the health data holders.

2. Datasets with electronic health data collected and processed with the support of Union or national public funding shall have a data quality and utility label covering the elements set out in paragraph 3.

3. The data quality and utility label shall cover the following elements, where applicable:

(a) for data documentation: metadata, support documentation, the data dictionary, the format and standards used, the source of the data and, where applicable, the data model;

(b) for assessment of technical quality: completeness, uniqueness, accuracy, validity, timeliness and consistency of the data;

(c) for data quality management processes: the level of maturity of the data quality management processes, including review and audit processes, and bias examination;

(d) for assessment of coverage: the period, population coverage and, where applicable, representativity of the population sampled, and the average timeframe in which a natural person appears in a dataset;

(e) for information on access and provision: the time between the collection of the electronic health data and their addition to the dataset and the time needed to provide electronic health data following the issuing of a data permit or a health data request approval;

(f) for information on data modifications: merging and adding data to an existing dataset, including links with other datasets.

6. By ... [two years from the date of entry into force of this Regulation], the Commission shall, by means of implementing acts, set out the visual characteristics and technical specifications of the data quality and utility label, based on the elements referred to in paragraph 3 of this Article.

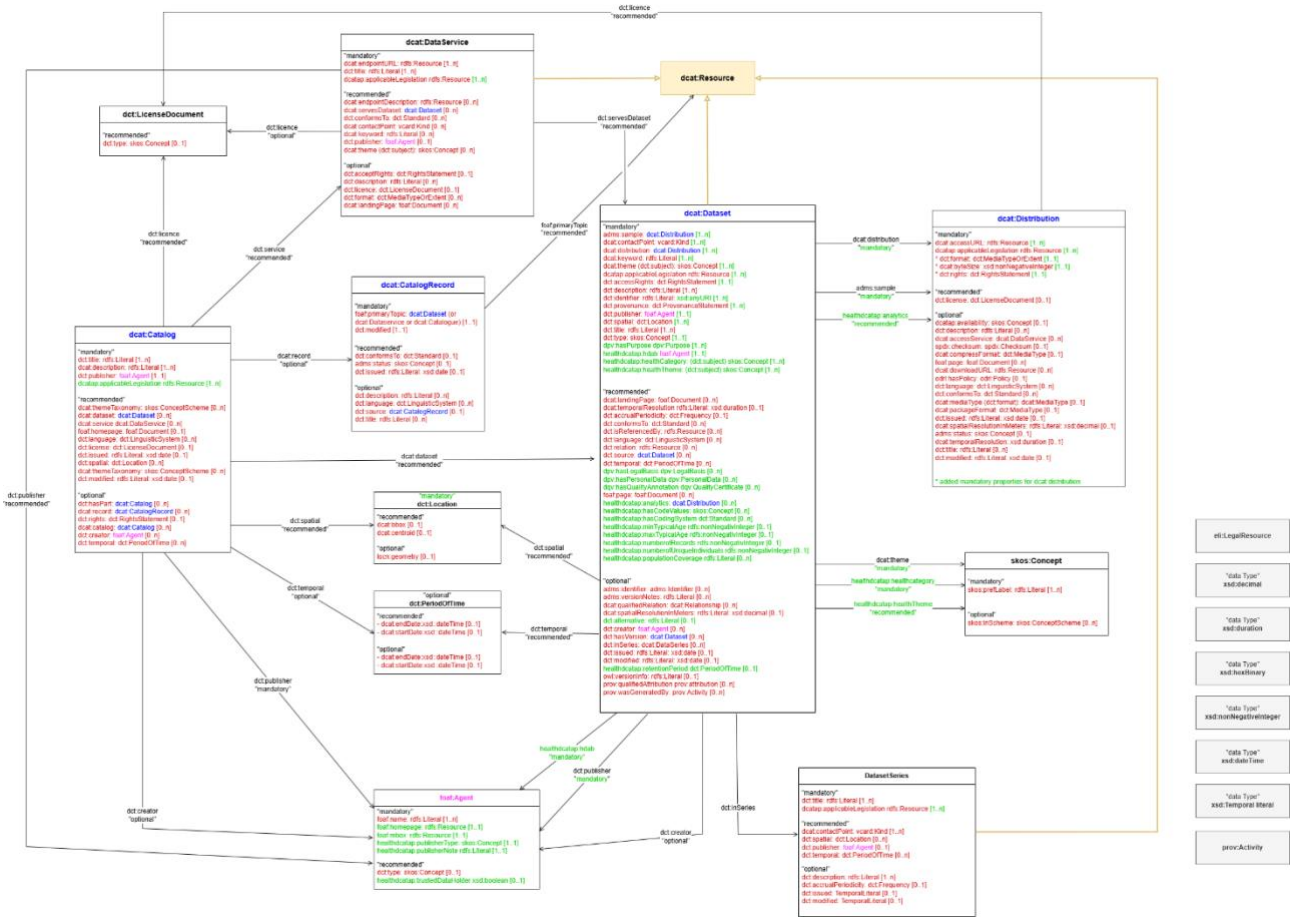
Article 79: EU dataset catalogue

1. The Commission shall establish an EU dataset catalogue connecting the national dataset catalogues established by the health data access bodies in each Member State as well as the dataset catalogues of authorised participants in HealthData@EU.

2. The EU dataset catalogue, the national dataset catalogues and the dataset catalogues of authorised participants in HealthData@EU shall be made publicly available.

Annex II: HealthDCAT-AP UML Diagram

Figure 2: HealthDCAT-AP UML Class Diagram (extending DCAT-AP 3.0). New properties and adapted properties (through updated cardinalities) are highlighted in green, while properties reused without modification from DCAT-AP are marked in red. Source: HealthDCAT-AP



Annex III: Workshops on Article 51 – health categories

Date	Topic	Article 51 Category
19 November 2024 14:00-16:00	Administrative Health Data	**(c)** aggregated data on healthcare needs, resources allocated to healthcare, the provision of and access to healthcare, healthcare expenditure and financing **(e)** healthcare-related administrative data, including dispensation, claims, and reimbursement data **(j)** data on professional status, and on the specialisation and institution of health professionals involved in the treatment of a natural person
26 November 2024 13:00-15:00	Registries & Research Data	**(k)** data from population-based health data registries such as public health registries **(l)** data from medical registries and mortality registries **(p)** data from research cohorts, questionnaires and surveys related to health, after the first publication of the related results
29 November 2024 10:00-12:00	Electronic Health Records & Clinical Data	**(a)** electronic health data from EHRs **(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, Regulation (EU) 2017/745 and Regulation (EU) 2017/746
7 January 2025 10:00-12:00	Socio-Economic, Environment and Behavioural Determinants	**(b)** data on factors impacting on health, including socio-economic, environmental and behavioural determinants of health
10 January 2025 10:00-12:00	Medicinal Products & Devices Registries	**(h)** personal electronic health data automatically generated through medical devices **(i)** data from wellness applications **(n)** other health data from medical devices **(o)** data from registries for medicinal products and medical devices
14 January 2025 10:00-12:00	Biobank Data	**(q)** health data from biobanks and associated databases.
17 January 2025 10:00-12:00	Genomic, Molecular and Pathogen Data	**(d)** data on pathogens that impact human health **(f)** human genetic, epigenomic and genomic data **(g)** other human molecular data such as proteomic, transcriptomic, metabolomic, lipidomic and other omic data

Note. All times are CET

Annex IV: HealthDCAT-AP technical working groups

Date	Topic	Explanation	Goal
31 October 2024 13:00-14:30	**Data lineage** dct:provenance prov:qualifiedAttribution prov:wasGeneratedBy	[dct:provenance] provides transparency about the origins and history of a dataset, ensuring data reliability and trustworthiness. It helps users understand how the dataset was created, modified, and by whom, which is crucial for assessing its quality and relevance. Furthermore, for [prov:qualifiedAttribution] and [prov:wasGeneratedBy] it could be that dedicated controlled vocabulary is needed.	**Improve the usage note and, if needed, controlled vocabulary.**
21 November 2024 13:00-14:30	**Health themes** healthdcatap:healthTheme	Similar to keywords, but designed for machines, the health theme property relies on Wikidata — a large-scale, human-readable, machine-readable, multilingual, multidisciplinary, centralised, editable, structured, and linked knowledge base. This significantly enhances the quality and usability of dataset descriptions for machines.	**Evaluate the use of Wikidata and confirm its suitability**
5 December 2024 13:00-14:30	**Publisher type** healthdcatap:publishertype	Controlled vocabulary needs to be created to allow health data holders to list the publisher type (example NPHI, Statoffice etc).	**Create controlled vocabulary for publisher types (and avoid 'other' as option)**
23 January 2025 13:00-14:30	**Cardinalities**	In the HealthData@EU Pilot study, a first decision was made on the cardinalities.	**Validation and consolidation of the decisions on the cardinalities.**
20 February 2025 13:00-14:30	**Distribution**	[healthdcatap:analytics] Publishers are encouraged to provide URLs pointing to API endpoints or document repositories where users can access or request associated resources such as technical reports of the dataset, quality measurements, usability indicators,... or analytics services [adms:sample]: When a health Dataset is categorised as personal electronic health data, implementers MUST provide descriptions for one sample Distribution of the dataset. These samples could be anonymised or synthetic subsets that retain the original dataset's essential characteristics without revealing any personal information, or it might solely exhibit the dataset's structure.	**Validation of these distributions, improve usage notes and define cardinalities**
6 March 2025 13:00-14:30	Other topics TBD	e.g. Dataset dimensions, Has personal data	
20 March 2025 13:00-14:30	Other topics TBD	e.g. Utilisation of CSVW for Data Dictionaries/Codebooks	
April 2025	Other topics TBD	If needed	

Note. All times are CET