
# D5.1 Guideline for data holders on data description

## Guideline for health data holders on their duties regarding data description

TEHDAS2 – Second Joint Action Towards the European Health Data Space

1 July 2025

**Co-funded by the European Union**

# 0 Document info

**Disclaimer**
Funded by the European Union. Views and opinions expressed are however 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


| Author(s) | Organisation |
|---|---|
| Nienke Schutte | Sciensano, Belgium |
| Beatriz Jacinto Barros | Sciensano, Belgium |
| Pascal Derycke | Sciensano, Belgium |
| Charles-Andrew Vande Catsyne | Sciensano, Belgium |
| Annelise Nyvold Lundbye | The Danish Healthcare Quality Institute, Denmark |
| Danielle Welter | Luxembourg National Data Service, Luxembourg |
| David Rosendahl | The Danish Healthcare Quality Institute, Denmark |
| Doupi Persephone | Finnish Institute for Health and Welfare, Finland |
| Jerome Daumas | Health Data Hub, France |
| Mari Mäkinen | Finnish Institute for Health and Welfare, Finland |
| Marianne Benderra | Health Data Hub, France |
| Marja-Riitta Rautiainen | Finnish Institute for Health and Welfare, Finland |
| Michael Peolsson | Swedish eHealth Agency, Sweden |
| Minna Liikala | Finnish Institute for Health and Welfare, Finland |
| Pieta Näsänen-Gilmore | Finnish Institute for Health and Welfare, Finland |
| Régis Lassalle | Health Data Hub, France |
| Wei Gu | Luxembourg National Data Service, Luxembourg |


## 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 |
|---|---|---|---|---|
| 10/03/2025 | 0.1 | Sciensano team | Initial draft developed based on feedback from the public consultation | Draft |
| 26/06/2025 | 0.2 | Sciensano team + T5.1 contributors | Review and refinement of content | Draft |
| 03/06/2025 | 0.3 | Sciensano team | Final formatting and quality check | Draft |

Accepted in Project Steering Group on 24 June 2025.

# Contents

# 1 Executive summary

This guideline provides practical and legal support for health data holders subject to the obligation to describe and make datasets available for secondary use under the European Health Data Space (EHDS) Regulation. It addresses two fundamental questions:

(1) are the electronic health data I hold within the scope of the secondary use framework established by the EHDS Regulation; and

(2) if so, how can I comply with the dataset description obligation laid down in Article 77, using the common metadata model (HealthDCAT-AP) that will be formalised through an implementing act?

To address these questions, the document provides:

- **Clarification of the categories of electronic health data** (as defined in Article 51 of the EHDS Regulation) that must be made available for secondary use. These are accompanied by full-text definitions, legislative context, and real-world examples to help data holders assess whether the data they hold fall within the scope of the secondary use framework.

- **Guidance on the dataset description obligation** (Article 77), providing an explanation of the legal requirement for data holders to describe datasets made available for secondary use, and of the role of metadata in ensuring transparency, discoverability, and interoperability.

- **Introduction to the HealthDCAT Application Profile (HealthDCAT-AP):** a comprehensive presentation of the common metadata model to be used to comply with Article 77, including its structure, rationale, and alignment with the broader DCAT-AP framework used across the EU.

- **Detailed, practical instructions for each property of the HealthDCAT-AP** , with tips and examples based on questions and challenges raised during the TEHDAS2 activities. These instructions are designed to help health data holders accurately complete their metadata records.

- A summary of the **analytical and conceptual work carried out in the TEHDAS2 project** to assess the applicability of HealthDCAT-AP and to support its refinement and validation, informed by expert feedback.

- A dedicated chapter on **cross-cutting implementation considerations**, based on TEHDAS2 workshops and expert input from across EU Member States, including key challenges and recommended mitigation strategies.

By following the guidance provided in this document, health data holders can better understand and prepare to meet their dataset description obligations under the EHDS framework. This guideline may also inform the work of the the European Commission and other stakeholders seeking to understand the technical and organisational considerations involved in the future implementation HealthDCAT-AP across the EU.

# 2 Introduction

**Advancing health data use in the European Health Union**

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 policymaking. 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 smooth cross-border use of health data, and support data holders, data users and the new health data access bodies in fulfilling their responsibilities and obligations outlined in the European Health Data Space (EHDS) Regulation[1].

TEHDAS2 focuses on several critical aspects of health data use.

- **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.

## 2.1 Purpose

The EHDS Regulation aims to create a common 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 the reuse of 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**

[1] Regulation - EU - 2025/327 - EN - EUR-Lex

**Strategy**[2], 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 harmonizing practices and ensuring interoperability, the EHDS will simplify data-sharing processes, foster collaboration among Member States, and support the efficient use of resources. Specifications such as DCAT-AP (Data Catalogue Application Profile)[3], play a critical role for interoperability, ensuring consistency across the EHDS and other European data spaces.

The use of **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 started the work on the **common descriptive metadata model HealthDCAT-AP**, tailored to the health domain. HealthDCAT-AP extends the DCAT model to support the discovery and understanding of health datasets, improving their accessibility while ensuring privacy and security.

This guideline supports health data holders in understanding and preparing to meet their obligations under the EHDS Regulation (Articles 60 and 77). Specifically, this document:

- Provides clarification on the categories of electronic health data that must be described and made available for secondary use;
- 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 understand and prepare to meet their legal obligations while contributing to the broader goals of a secure and interoperable European Health Data Space.

## 2.2 Problems being addressed

This guideline addresses to two main questions:

- ❑ **Which categories of electronic health data fall within the scope of the secondary use obligations under the EHDS Regulation (Article 51)?**
- ❑ **How should these datasets be described to comply with the dataset description obligation laid down in Article 77, using the HealthDCAT-AP application profile?**

[2] European data strategy - European Commission
[3] DCAT-AP 3.0

Recital 80 of the EHDS regulation emphasizes 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[4].

Furthermore, a common metadata model is needed to **enhance data discovery** by ensuring consistency in how datasets are described, organised and shared in the national dataset catalogues provided by HDABs and in the HealthData@EU Central Catalogue. A unified description model 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[5] 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. An harmonised data description is essential for implementing the FAIR principles by ensuring that data is findable, accessible, interoperable, and reusable. It enables data discovery through harmonised descriptions, provides information on how to access datasets, ensures compatibility across systems, and documents the context and quality of data for future use. The EHDS Regulation endorses the FAIR data principles as a foundation for responsible data reuse, and a harmonised metadata model is key to putting these principles into practice.

## 2.3 Target audience

This document is designed to serve **three main purposes** within the framework of the EHDS, that identifies the main target audience:

1. **Support health data holders** in understanding their legal obligations under the EHDS Regulation, learning the fundamentals of the HealthDCAT-AP metadata model, and applying it in practice to create metadata records for secondary use of health data.

2. **Inform the European Commission** on the development, refinement, and validation of the HealthDCAT-AP common metadata model for secondary use under the EHDS. This supports the future Implementing Act, which will define the minimum metadata elements required and formally establish HealthDCAT-AP as the EU-wide metadata model for secondary use under the EHDS.

[4] 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

[5] 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

3. **Provide guidance to HDABs** on the HealthDCAT-AP metadata model, helping them understand how metadata will be displayed in national dataset catalogues and how specific properties can support tailored search functionalities. More details on this are available in Deliverable D5.3 *Technical specification on the national metadata catalogue*.

**How to consult this document depending on your role:**

- **If you want to understand whether you are a health data holder under the EHDS**: Refer to **Chapter 5**, which explains the legal obligations and provides **full-text definitions and examples** for each of the data categories listed in Article 51. It also reflects on implementation challenges identified during dedicated TEHDAS2 workshops.

- **If you are a health data holder and want to learn how to describe your data using HealthDCAT-AP**: Consult **Chapter 6**, which explains why HealthDCAT-AP extension was developed as the common metadata model, provides foundational knowledge on RDF, DCAT-AP, and clarifies what HealthDCAT-AP is and is not. Then refer to **Chapter 7**, which offers practical, step-by-step instructions for completing each property of HealthDCAT-AP, tailored to common questions raised by data holders during implementation.

- **If you represent the European Commission and seek insights on the TEHDAS2 work supporting HealthDCAT-AP validation and refinement**: Explore **Chapter 8**, which summarises the key outcomes of the TEHDAS2 validation work, including proposed technical solutions, identified challenges, and areas for future development.

- **If you are interested in cross-cutting implementation considerations** (regardless of your role): See **Chapter 9**, which presents a high-level overview of the main challenges and proposed mitigation strategies discussed within the TEHDAS2 expert community.

# 3 Key terminology

| Term | Definition |
|---|---|
| Controlled vocabularies[6] | Controlled vocabulary refers to a curated and structured list of words, terms, and phrases which provide a uniform method of describing content in a given subject area, resulting in a harmonized terminology. |
| Data minimisation | A principle mandating organisations to only collect, store and process the minimum necessary amount of personal data for a specific purpose. This principle is fundamental under GDPR and |

[6] Controlled vocabularies - EU Vocabularies - Publications Office of the EU

| | |
|---|---|
| | relevant to all health data uses outlined in EHDS. (GDPR Article 5(1)(c)) |
| Dataset | A structured collection of electronic health data. (EHDS Article 2(2)(w)) |
| Dataset Catalogue | 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. (EHDS Article 2(2)(y)) |
| Health data access body (HDAB) | A Member State-designated authority responsible for facilitating the secondary use of electronic health data. HDABs evaluate data access applications, authorise and issue data permits, obtain datasets from data holders, and make them available in Secure Processing Environments. They also track all requests and permits issued, and are responsible for ensuring transparency by publishing information on granted data permits (EHDS Regulation, Articles 55 and 58, and Recital 52) |
| Health data holder | An entity that processes electronic health data as a data controller for the purposes of provisioning care or healthcare, developing healthcare products, services, wellness applications and undertaking healthcare research, or processing healthcare data for innovation, policy development, official statistics or patient safety or for regulatory purposes. (EHDS Article 2(2)(t)) |
| Health data user | A natural or legal person, including European Union institutions, bodies or agencies, which has been granted lawful access to electronic health data for secondary use pursuant to a data permit, health data request approval or an access approval by an authorised participant in HealthData@EU etc. (EHDS Article 2(2)(u)) |
| Interoperability | 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 (EHDS Article 2(2)(f)) |
| Non-personal electronic health data | Electronic health data other than personal electronic health data, including both data that have been anonymised so that they no longer relate to an identified or identifiable natural person (the ‘data |

| | |
|---|---|
| | subject') and data that have never related to a data subject (EHDS Article 2(2)(b)) |
| Personal electronic health data | Data concerning health and genetic data, relating to an identified or identifiable natural person, processed in an electronic form (EHDS Article 2(2)(a)) |
| Secondary use | Processing of electronic health data for the purposes set out in Chapter IV of EHDS Regulation, other than the initial purposes for which they were collected or produced (EHDS Article 2(2)(e)) |
| Trusted health data holder | Member State designated health data holder for whom a "simplified" procedure can be followed for the issuance of data permits. as defined in Article 72 of the EHDS regulation. Trusted health data holders leverage their expertise on the data they hold to assist the HDAB by providing assessments of data requests or access applications. Once data permits are authorised, these trusted data holders provide the data within an SPE that they manage. (EHDS, Article 72) |

# 4 Scope

This document provides guidance on the description of electronic health data for secondary use under the EHDS Regulation. It focuses on the implementation of the HealthDCAT-AP and supports health data holders in understanding and preparing to comply with their obligations for metadata provision, as outlined in Articles 60 and 77 of the Regulation. More specifically, this document includes:

- A clarification of the categories of electronic health data that must be made available for secondary use under Article 51, along with explanations, legislative context, and practical examples.
- An overview of the HealthDCAT-AP extension, developed to accommodate the specificities of health data;
- A detailed explanation of HealthDCAT-AP properties, with practical instructions and tips to support the creation of metadata records by health data holders;
- A summary of the conceptual work, design decisions, and validation work carried out during the TEHDAS2 project.

The following topics are out of the scope of this guideline:

1. Dataset collection, structuring, and management practices, including how health data is generated or organised prior to metadata creation;

2. Integration of HealthDCAT-AP into national dataset catalogues: for more information on this, refer to the upcoming Deliverable 5.3 – Technical specifications on the national metadata catalogue, available at TEHDAS2 website (Link)
3. The data quality and utility label to be provided by data holders according to Article 78: this is being developed by the QUANTUM project and is not addressed in this document. For more details, see https://quantumproject.eu/
4. Guidelines for Health Data Access Bodies on assessing requests for secondary use based on purpose limitations under the EHDS Regulation and technical processes for data access provision, including secure processing environments, data linkage, and anonymisation, among others. These topics will be covered in separate reports published as part of TEHDAS2. Follow updates here: https://tehdas.eu/public-consultations/

# 5 Data for secondary use under the EHDS

## 5.1 Legal basis: Understanding the obligations for secondary use

Before diving into the specific categories of data that must be made available, it is important to remember that the EHDS framework for secondary use focuses on the **reuse of existing data**. The EHDS Regulation does not impose an obligation to collect new data or to digitise health data originally collected in paper form. Instead, health data holders are required to describe and make available, for secondary use, only those datasets they already hold in electronic format, when these fall within the categories listed in Article 51 and are requested through the procedures laid down in Chapter IV of the EHDS Regulation.

> Article 2(2) defines secondary use as:
>
> *"The processing of electronic health data for the purposes set out in Chapter IV of this Regulation, other than the initial purposes for which they were collected or produced."*

The emphasis is therefore on enabling access to existing electronic health data initially collected for public interest purposes such as research, innovation, policy making, public health, and regulatory activities (see Article 53), which may still be made available for secondary use under the EHDS, provided access is requested via the mechanisms established in Chapter IV of the EHDS Regulation.

### 5.1.1 Who shall make data available?

Under the EHDS Regulation, the obligation to make electronic health data available for secondary use applies only to entities that meet both of the following conditions:

1. They qualify as a "**health data holder**", as defined in Article 2(2)(t) of the Regulation
2. They hold or process data that fall within one or more of the categories listed in Article 51, which defines the scope of data eligible for secondary use.

Understanding this double condition is essential for determining whether your organisation is subject to the obligations laid down in Chapter IV of the EHDS Regulation.

The definition of a health data holder in Article 2(2)(t) includes:

- Any natural or legal person, public authority, agency or other body in the healthcare or the care sectors (including reimbursement services where necessary)

- 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
  - Acting as a mortality registry

- Any Union institution, body, office or agency, with:
  - 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;
  - 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.

### 5.1.2 Who is exempt?

According to Article 50, the following are exempt from these obligations:

- Natural persons, including individual researchers
- Legal persons qualifying as microenterprises under EU law

Additionally, Member States may further regulate these exemptions at national level.

### 5.1.3 What Data must be made available?

Entities that qualify as health data holders under Article 2 of the EHDS Regulation, such as natural or legal persons, or Union institutions, bodies, offices, or agencies, are expected to make available the data they hold or process as controllers or joint controllers, provided that the data falls within the categories of electronic health data eligible for secondary use, as defined in Article 51:

- ❑ **(a)** Data from electronic health records (EHRs)
- ❑ **(b)** Data on socio-economic, environmental, and behavioural determinants of health
- ❑ **(c)** Aggregated data on healthcare needs, resources, access, expenditure, and financing
- ❑ **(d)** Data on pathogens affecting human health
- ❑ **(e)** Healthcare-related administrative data (e.g. dispensations, reimbursements)
- ❑ **(f)** Human genetic, epigenomic, and genomic data
- ❑ **(g)** Other omics data (e.g. proteomic, transcriptomic, metabolomic)

- ❑ **(h)** Personal health data automatically generated by medical devices
- ❑ **(i)** Data from wellness applications
- ❑ **(j)** Data on professional status, specialisation, and affiliation of healthcare providers
- ❑ **(k)** Data from population-based health registries
- ❑ **(l)** Data from medical and mortality registries
- ❑ **(m)** Data from clinical trials, studies, and investigations under relevant EU legislation
- ❑ **(n)** Other health data from medical devices
- ❑ **(o)** Data from medicinal product and medical device registries
- ❑ **(p)** Data from research cohorts, surveys, and questionnaires (after publication of results
- ❑ **(q)** Data from biobanks and associated databases

It is important to note that simply holding data falling under the categories listed in Article 51 does not, by itself, create an obligation to make data available. Only entities that qualify as *health data holders* under Article 2 of the EHDS Regulation are subject to the obligations set out in Article 50. Entities outside this definition are not required to comply.

Article 51 also allows Member States to extend the scope of health data categories made available for secondary use. They may also define additional safeguards or rules for handling specific data types, such as genetic or omic data, data from wellness applications, and biobanks. Consider your national implementation for additional requirements or restrictions.

**Note:** According to Article 105, Article 51(1), points (b), (f), (g), (m), and (p), shall apply six years after the date of entry into force of the EHDS Regulation.

## 5.2 Clarifying EHDS Data Categories: Method and Inputs

### 5.2.1 Methodology for clarification on the EHDS Categories for secondary use

It can be a challenge for data holders to derive clear criteria from the legislative text to determine whether their datasets fall under Article 51 of the EHDS Regulation. While the EHDS Regulation defines the data categories in Article 51, further operational clarification was needed to support their practical interpretation and application by health data holders.". In the TEHDAS2 Joint Action, each category of data listed in Article 51 was examined through a series of dedicated workshops. The TEHDAS community, along with other stakeholders and domain experts, was invited to discuss the scope of the minimum categories of electronic health data for secondary use that health data holders shall make available under Article 51 and collect examples of datasets for labelling or tagging.

The workshops aimed to develop:

- **Clear, full-text definitions** of each category
- **Examples** of relevant data that fall under each category
- **Short label titles** to support easier categorization

This work was further supported by surveys and bilateral consultation with field experts.

### 5.2.2 Key Challenges Identified

Through the total of 7 workshops, several challenges were identified:

**Understanding and structuring the health data landscape**

Mapping data to the EHDS categories in Article 51 is not always straightforward. Data holders must first make an inventory of what data they have collect and then identify what data they are obliged to share for secondary use under the EHDS. This becomes especially complex given the vast amount of unstructured data in the health domain, which cannot easily be classified or reused without further processing. The discussions during TEHDAS2 activities highlighted that the EHDS Regulation promotes the FAIR data principles, which encourage making datasets findable, accessible, interoperable and reusable. While not directly prescriptive, these principles support the aims of data discoverability and interoperability. For many health data holders, achieving the objectives of the EHDS may require adapting internal practices for organising, structuring, and maintaining health data. The challenge of unstructured data is widely recognised. Nonetheless, the responsibility will ultimately lie with each data holder to define a structured approach suitable to their local context. EU-wide efforts in the form of support mechanisms, training, and shared tools will be essential to help data holders build the capacity required to meet the obligation of making both structured as well as unstructured data available for secondary use.

**Navigating diversity in health data and its purposes**

Even with clarified definitions, many data holders can face challenges to categorise their datasets due to the high variability in data types, collection methods, and local practices. Health data is not uniform - there are countless modalities and formats, and how data is collected and used varies significantly between Member States, regions, and even institutions. In practice, many datasets combine multiple categories of data from several sources, making categorisation a challenge. While Article 51 defines data categories based on the type of data rather than the original purpose of collection, stakeholder discussions suggested that the purpose for which data was initially collected may, in practice, help inform classification decisions, especially for complex or hybrid datasets.

**Linking with other data spaces**

Another key insight from the discussions was that a growing amount of non-health data is relevant for health-related questions. Although data from other domains, such as environmental monitoring, social media, or behavioural tracking, do not fall under the EHDS by default, their increasing importance underscores the need for interoperability across European data spaces, as emphasised in the EHDS Regulation (Recital 80). Ensuring that the EHDS is designed with interfaces and standards to combine and contextualise data across sectors is an important enabler of richer insights for public health, research, and policy.

**Non-alignment of definitions and concepts**

A key challenge raised across multiple discussions was the lack of alignment in how health-related indicators and concepts are defined across different contexts. Some existing initiatives, such as the Genomic Data Infrastructure (GDI), use definitions that differ from or

partially overlap with those in the EHDS. Mapping and alignment efforts will be needed to ensure coherence across domains. Similarly, organisations such as the OECD and Eurostat highlighted that established definitions for administrative data already exist at international level, which may differ from or overlap with those now introduced in the EHDS framework. The discussions also revealed different interpretations and definitions of key concepts that vary across national policies, institutional practices, or legal traditions (for instance, what constitutes a clinical trial). Addressing these conceptual divergences will require joint clarification efforts, alignment across data standards, and potentially the development of guidance or mappings that link existing definitions to the EHDS categories.

**The need for harmonisation and EU-level coordination**

Perhaps the most consistent message from the discussions was the deep fragmentation in how data is collected, organised, and managed across Europe. This fragmentation is rooted in long-standing national practices, local legislation, and institutional traditions, making harmonisation a significant challenge. Within the same country and even within the same organisation, data structures and procedures can differ, complicating efforts to align with the EHDS requirements. Implementing and sustaining the EHDS for secondary use will therefore require strong coordination mechanisms at both the national and EU levels. This will be an evolving process. As implementation progresses, the mechanisms and solutions underpinning the EHDS will need to adapt to the practical realities of how data is collected, linked, and used. Long-term success will depend on continuous dialogue, support for data holders, and flexible, collaborative approaches to standardisation.

## 5.3 Clarification on the EHDS categories for secondary use

Considering the challenges highlighted above, the definitions presented in this section were developed with several objectives in mind:

- The definitions must be broad and flexible, to accommodate the vast range and heterogeneity of health data within each category.
- Despite this flexibility, they must also provide clear boundaries to help health data holders assess whether their datasets fall under EHDS obligations and, if so, to describe them accordingly.
- The definitions should align with regulatory and technical standards already in use in the relevant fields.
- They should also reflect real-world data practices, examples, and use cases to support practical implementation.

This balance is what the core working group and expert contributors have aimed to achieve in the definitions presented below (Table 1-17).

For each category, a description of the legislative intent as provided by the European Commission is included.

**Important note:** Only data collected by entities considered health data holders under Article 2 of the EHDS Regulation, and not exempt from the obligations outlined in Article 50, are in scope for this category. Data collected by entities outside this definition are not covered. Additionally, the categories listed in Article 51 are not mutually exclusive, a single dataset

may fall under multiple categories simultaneously. For all categories, natural persons retain the right to opt out at any time from the processing of their personal electronic health data for secondary use.

### 5.3.1 Data from electronic health records (EHRs)

**Table 1. Clarification of Article 51(1)(a):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by RKKP (Denmark) and discussed during the TEHDAS2 workshops.

| Article 51 category | (a) Data from electronic health records (EHRs) |
|---|---|
| **Legislative intent** | **Description:** This category covers all information recorded in electronic health records (EHRs) in the context of healthcare provision. It includes data generated or documented during clinical encounters, diagnostic processes, and treatment, with the goal of supporting continuity of care and medical decision-making. The intent is to enable secondary use of comprehensive patient health information recorded during routine care.<br><br>**Data in scope:** Structured and unstructured EHR data, clinical notes, prescriptions, lab results, medical images, paramedical and nursing notes, appointment records, referrals, diagnostic reports, and pathology/histology data.<br><br>**Data out of scope:** Data not stored in the EHR system |
| **Definition** | A collection of electronic health data related to a natural person, processed for the purpose of the provision of healthcare and stored in an Electronic Health Records (EHRs) system.<br><br>Note: Electronic health data from Electronic Health Records, EHRs, refers to the comprehensive digital collection of health information about individuals, such as medical history, diagnosis codes, clinical notes, test results etc., processed for the purpose of providing healthcare. |
| **Supporting legislative definitions** | **"Electronic health record"** or **"EHR"** means a collection of electronic health data related to a natural person and collected in the health system, processed for *the purpose of the provision of* healthcare (Article 2, points 2.j)<br>**"Electronic health data"** means personal or non-personal electronic health data (Article 2, points 2.c) [NOTE: since the definition of "EHR" is restricted to "data related to a natural person" only the first part of the definition of "Electronic health data" is relevant here.]<br>**"Personal electronic health data"** means data concerning health and genetic data processed in an electronic form (Article 2, points 2.a)<br>**"Data concerning health"** means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status (Regulation (EU) 2016/679, Article 4, points 15)<br>**"Genetic data"** means personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question (Regulation (EU) 2016/679, Article 4, points 13) |
| **Examples** | The electronic health data in the EHRs are exemplified in the introductory remark to the regulation as follows : "In health systems, personal electronic health data are usually gathered in electronic health records, which typically contain a natural person's medical history, diagnoses and treatment, medications, allergies and vaccinations, as well as radiology images, laboratory results and other medical data, spread between different actors in the health system, such as general practitioners, hospitals, pharmacies or care services." (Recital 7).<br><br>The following is a non-exhaustive list of examples of different types of data that are typically part of the category: |

| | |
|---|---|
| | • **Medical History**: A historical collection of medical data, i.e. an account of past illnesses, surgeries, medical complications, family medical history and hospital admissions.<br>• **Patient Diagnostics:** Information regarding title of diagnosis and diagnosis code.<br>• **Treatment**: medical, surgical or therapeutic interventions aimed at treating a specific illness, injury, or condition.<br>• **Care Plan**: Personalized statement of planned healthcare activities relating to one or more specified health issues.<br>• **Medication:** A drug or drugs (medication can be 1 or several medicinal products) prescribed or given as medical treatment to treat or prevent disease.<br>• **Allergy Information**: Documentation of known allergies to medications, foods, or environmental factors.<br>• **Immunization Records**: Documentation of vaccinations received by the patient, including dates and types of vaccines.<br>• **Images**: Information from imaging tests and studies, e.g. X-rays, MRIs and CT scan. Either in the form of summaries and descriptions or including the images<br>• **Laboratory Results**: Results from laboratory tests, e.g., blood tests, urine tests.<br>• **Clinical Notes**: Observation and progress notes recorded by healthcare providers, e.g. descriptions or symptoms that are not registered as part of the diagnosis.<br>• **Vital Signs**: Regularly recorded measurements, such as blood pressure, heart rate, respiratory rate, and temperature.<br>• **Referral Information**: Details of referrals to specialists or other healthcare providers, including the reason for referral and any follow-up instructions.<br>• **Patient Reported Outcome (PRO)** data reported directly by a patient on his or her own health condition, without interpretation by a doctor or anyone else.<br>• **Test**: Information regarding exercises, physical exams and other test.<br>• **Risk factors:** Information on medical dispositions, smoking, drugs and alcohol, workplace risk factors.<br>• **Basic patient information**: Information such as name, age, gender, height, weight, ethnicity, address, and contact details.<br><br>Examples of data sets or metadata descriptions of data sets:<br>• A research database with data from general practices in the Netherlands: Link<br>• A research database with data from general practices in the UK: Link<br><br>**Examples of excluded datasets:**<br>• Personal health records (PHRs) managed by patients themselves.<br>• Wellness app data.<br>• General population health surveys and registers.<br>• Patient-collected health data outside clinical settings. (Applications that store data in EHRs are not excluded.)<br>• Data for other systems that are not imported to or synchronized with EHRs, e.g. Radiology Information Systems (RIS), Cardiology Information Systems (CIS) or Picture Archiving and Communication System (PACS)<br>• Data from systems that do not register health data, e.g. door systems, archiving systems for health data, backup systems, clinic presentation systems, statistical processing of workflow |
| **Label** | EHR data |

### 5.3.2 Data on factors impacting on health, including socio-economic, environmental and behavioural determinants of health

**Table 2. Clarification of Article 51 (1)(b):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by LNDS (Luxembourg) and discussed during the TEHDAS2 workshops.

| Article 51 category | (b) Data on factors impacting on health, including socio-economic, environmental and behavioural determinants of health |
|---|---|
| **Legislative intent** | **Description:** This category pertains to individual-level data held by health data holders that reflect factors impacting health outcomes.<br><br>**Data in scope:** Socio-Economic Determinants includes data on economic status, occupation, and employment status. Environmental Determinants includes data on exposure to environmental pollution (e.g., air quality, chemical exposure, as mentioned explicitly in Recital 56). Behavioural Determinants includes data on smoking, alcohol consumption, and physical activity levels.<br><br>**Data out of scope:** Socio-economic data not systematically held by health data holders. For instance, tax offices have records about income, employment status and family composition, but are not health data holders under EHDS. Solely environmental data unrelated to personal health (e.g., general climate data in High Value Datasets under the Open Data Directive), though such data can be referenced by data users. |
| **Definition** | Data collected by Health Data Holders that do not constitute health data in their own right but that can provide context for health status, diagnoses and prognoses, population stratification and outcomes for individuals or population groups. They include but may not be limited to:<br>• **Socio-economic determinants**, i.e. factors relating to social and economic background, such as income, employment status, job quality, education, marital status, migration background, household size, living area (rural/urban) and social support systems.<br>• **Environmental determinants**, i.e. factors to which individuals are exposed to in their daily life, such as air and water quality, mobility, housing conditions or workplace conditions.<br>• **Behavioural determinants**, i.e. actions taken intentionally or unintentionally by individuals that may impact their health, such as diet, exercise, tobacco, alcohol and drug use, hand washing or screen time. |
| **Examples** | Examples in scope:<br>Example 1 - Norwegian Institute of Public Health<br>Example 2 - Norwegian Environmental Biobank<br>Example 3 - Sweden Regional Comparisons Public Health 2019<br><br>**Examples out of scope:**<br>• Open data that represents population-based information about socio-economic, environmental and behavioral determinants, especially data that has never been personal.<br>• Data falling under the definition of “behavioural determinants” should not overlap with category “(i) data from wellness applications”. The former might for example include whether a person does any exercise while the latter includes activity data from a smartwatch or fitness app. |
| **Label** | Socio-economic, environmental and behavioral determinants of health |

### 5.3.3 Aggregated data on healthcare needs, resources allocated to healthcare, the provision of and access to healthcare, healthcare expenditure and financing

**Table 3. Clarification of Article 51 (1)(c):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by HDH (France) and discussed during the TEHDAS2 workshops.

| Article 51 category | (c) Aggregated data on healthcare needs, resources, access, expenditure, and financing |
|---|---|
| **Legislative intent** | **Description:** This category covers aggregate-level data related to the functioning of health systems, including healthcare needs, resource allocation (financial and human), access to services, service provision, and health expenditure. It does not include individual-level data, and may include data derived from other categories (e.g. medical or administrative data), as long as it has been aggregated.<br><br>**Data in scope:** National or regional statistics on healthcare access and coverage, health workforce and infrastructure capacity data, aggregated data on healthcare spending and financing, resource allocation metrics<br><br>**Data out of scope:** Individual-level data on patients or professionals, raw data prior to aggregation, financial or administrative data not linked to health system analysis. |
| **Definition** | Combination of related healthcare data categories collected and maintained by Member States' administrations and public institutions (such as health insurance bodies) collected by any healthcare provider and other relevant entities at the national, regional, and local (e.g. hospital…) levels. These data encompass key aspects of healthcare systems, including needs, access, coverage, provision, resources, and financial flows. They can fall into three main categories:<br><br>• Aggregated data registered for ensuring and maintaining **health care financing schemes** (Eurostat definition: types of financing arrangements through which people obtain health services, including both direct payments by households for services and goods and third-party financing arrangements).<br>• Aggregated data registered for ensuring and maintaining **health care functions expenditures** (see Eurostat definition: expenditure for curative cares, rehabilitative cares, outpatient care, day care, long-term care, home-based care, ancillary services, pharmaceuticals and other medical non-durable goods, therapeutic appliances and other medical goods, preventive care, governance and health systems and financing administration).<br>• Aggregated data on healthcare needs and access registered **to assess healthcare demand, coverage, service utilization**, and barriers to access. |
| **Supporting sources** | A System of Health Accounts 2011: revised edition<br>Health Care functions expenditures: webpage link<br>Aggregate definition (Eurostat): webpage link |
| **Examples** | Examples in scope:<br>• SAE Annual statistics on health institutions (France, holder: DREES) Example 1<br>• Data.DREES - Open data on health and social protection (France, holder: DREES) *Example 2*<br>• Open Data CNAM - Datasets related to healthcare and social security (France, holder: CNAM) *Example 3*<br>• ScanSanté - National hospital data for public health purposes (France, holder:Agence technique de l'information sur l'hospitalisation, ATIH) *Example 4*<br>• Odissé - National health and epidemiological data (France, holder: Santé Public France, SPF) *Example 5* |

| | |
|---|---|
| | • OECD Health data - Statistics on health and health systems across OECD countries (EU, holder: OECD) Example 6<br>• Dataset of the Belgian TCT - covers a linkage between hospital discharge data and reimbursement data of the health insurance companies with reimbursed costs aggregated at several levels (eg hospital level, by medical diagnosis) Example 7<br>• FPS Social Security - Belgian social security system - Data on healthcare expenditures calculated yearly and aggregated at several levels Example 8 |
| **Label** | Aggregated healthcare data |

### 5.3.4 Data on pathogens that impact human health

**Table 4. Clarification of Article 51 (1)(d):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by LNDS (Luxembourg) and discussed during the TEHDAS2 workshops

| **Article 51 category** | **(d) Data on pathogens affecting human health** |
|---|---|
| **Legislative intent** | **Description:** Genetic or molecular data describing pathogens that affect humans.<br><br>**Data in scope:** SARS-CoV-2 genome sequence data, bacterial strain identification through DNA sequencing, molecular data on viral mutations.<br><br>**Data out of scope:** Immune response data (e.g., antibody levels or cytokine profiles), pathogen data affecting animals with no human health link. |
| **Definition** | Data collected by health data holders related to pathogens (bacteria, viruses, fungi, protozoa and other parasites) affecting human health, including<br>• **Pathogen sample data:** data about physical samples of pathogens collected from infected individuals or environments, and that are used for clinical diagnostics, surveillance and research purposes;<br>• **Pathogen genomic data:** raw read sequences, assembled sequences, variant calling and analyses of pathogen genomes for identifying genetic variation in pathogens as well as understanding and tracking the evolution and spread of pathogens;<br>• **Pathogen epidemiological & environmental data:** information on the presence, incidence, distribution, and control of pathogens in various environments, such as water, soil and air, but also in animal species with an increased risk of animal to human transmission of pathogens, used to identify potential sources of infection or zoonotic transmission, assess risks, track outbreaks and implement public health measures;<br>• **Other pathogen-related laboratory data:** findings from laboratory experiments (excluding laboratory tests in a clinical setting) and field studies. |
| **Examples** | Examples in scope:<br>• Clinical microbiological tests results<br>• Phenotypic data<br>• Base-called raw sequence data (eg. fastq)<br>• Species typing data<br><br>Antimicrobial resistance surveillance data in the EU/EEA - ECDC (EARS-Net): Example 1 |
| **Label** | Pathogen data |

### 5.3.5 Healthcare-related administrative data (e.g. dispensations, reimbursements)

**Table 5. Clarification of Article 51 (1)(e):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by HDH (France) and discussed during the TEHDAS2 workshops.

| Article 51 category | (e) Healthcare-related administrative data, including on dispensations, reimbursement claims and reimbursements |
|---|---|
| **Legislative intent** | **Description:** This category covers person-level and aggregated data generated for administrative purposes within healthcare settings. It is often related to - but distinct from - electronic health records. Importantly, this category refers specifically to healthcare administrative data and should not be confused with the broader concept of "administrative data" as used by Eurostat (ESTAT), which encompasses data held by public administrations for general statistical purposes. The current scope refers to healthcare-specific administrative data, rather than broader public administration datasets. Examples include data produced to support dispensation, claims processing, reimbursements, billing, and other operational functions of healthcare systems<br><br>**Data in scope:** Reimbursement and claims data, dispensation records collected for administrative use, diagnosis or treatment codes used for billing or reporting<br><br>**Data out of scope:** Data collected solely for clinical care purposes (falls under EHR data), aggregated health statistics not linked to administrative processes, clinical notes without administrative relevance. |
| **Definition** | Healthcare administrative data consist of information recorded at both individual and aggregated levels during every interaction with the healthcare system for administrative purposes, such as claims, reimbursement, and dispensation. These data are generated from various healthcare encounters, including physician visits, diagnostic procedures, hospital admissions, pharmacy dispensations, but also occupational health services, rehabilitation, residential care, and home-based healthcare. Additionally, they encompass claims data from healthcare and travel insurance companies, as well as healthcare expenses covered by local governments. |
| **Supporting sources** | Definition of Health Care functions and Health Care providers: Link |
| **Examples** | Examples in scope:<br>• Base principale du SNDS - National healthcare data system (France, holder: Caisse nationale d'Assurance maladie, CNAM) Example 1<br>• RAC - Reste A Charge, database on healthcare expenditure and out-of-pocket expenses (France, holder: Direction de la recherche, des études, de l'évaluation et des statistiques, DREES) *Example 2*<br>• GePaRD - German Pharmacoepidemiological Research Database (Germany, holder: Leibniz institute BIPS) *Example 3*<br>• ARS Toscana database - Tuscany healthcare and public health database (Italy, holder: Agenzia Regionale di Sanita, ARS) *Example 4*<br>• Healthcare reimbursement data of the health insurance companies in Belgium, gathered by the Intermutualistic Agency (IMA-AIM) *Example 5*<br>• Vektis healthcare insurers claims data, WMO claims data from local government *Example 6* |
| **Label** | Healthcare administrative data |

### 5.3.6 Human genetic, epigenomic, and genomic data

**Table 6. Clarification of Article 51 (1)(f):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by LNDS (Luxembourg) and discussed during the TEHDAS2 workshops.

| Article 51 category | (f) Human genetic, epigenomic, and genomic data |
|---|---|
| **Legislative intent** | **Description:** Data derived from genetic material (e.g., DNA, RNA) that reveals information about an individual's genome.<br><br>**Data in scope:** Data derived from a sequencing process on DNA or RNA, somatic or germinal from a human sample than can be related with a specific phenotype/genotype. For instance, BRCA1 mutation detected through gene sequencing and whole-genome sequencing data.<br><br>**Data out of scope:** Generic information of genes. For instance, mention of "BRCA1 linked to breast cancer" in clinical notes, results from routine biomarker tests that do not directly involve genetic analysis. |
| **Definition** | Data collected by health data holders regarding the genetic information of humans, including its structure, function, evolution and mapping, such as:<br>• **Whole-genome sequences**, which encompass all or most of the genetic makeup of individuals or groups of individuals;<br>• **Targeted sequences**, where only specific parts of the genetic material are sequenced, for example as part of a genetic diagnostic process;<br>• **Epigenomic data**, namely information regarding any modifications to genetic material and its expression in response to environmental factors and that may affect the expression of genes;<br>• **Derived data** obtained by performing analyses on the raw or assembled data for all the above such as variant calling and aggregated data such as profile and association information about variant-disease relations;<br>• **Genetic data** derived from non-sequencing techniques such as polymerase chain reaction (PCR) assays, microarrays, or cytogenetics techniques including karyotyping or fluorescent in-situ hybridisation (FISH). |
| **Examples** | Examples in scope:<br>• Next Generation Sequencing *Example 1*<br>• European Genome-Phenome Archive (EGA) entry *Example 2*<br>• European Variation Archive (EVA) *Example 3*<br>• Aggregated data *Example 4* |
| **Label** | Genetics, epigenetics and genomics |

### 5.3.7 Other human molecular data such as proteomic, transcriptomic, metabolomic, lipidomic and other omic data

**Table 7. Clarification of Article 51 (1)(g):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by LNDS (Luxembourg) and discussed during the TEHDAS2 workshops.

| Article 51 category | (g) Other human molecular data such as proteomic, transcriptomic, metabolomic, lipidomic and other omic data |
|---|---|
| **Legislative intent** | **Description:** Data describing molecular processes or biological pathways unrelated to direct genetic sequencing.<br><br>**Data in scope:** Proteomic analysis identifying protein expression levels, lipidomic data showing metabolic profiles, transcriptomic data from RNA sequencing.<br><br>**Data out of scope:** Raw molecular data generated for industrial or non-health purposes, generic lab results like cholesterol or creatinine levels. |
| **Definition** | Any human molecular data other than genetic, genomic or epigenomic data, in either raw, cleaned or analysed form, including but not limited to:<br>• Data about the structure, function, interaction, identification and quantification of proteins produced by humans in a range of environments (proteomics), generated from techniques such as antibody-based immunoassays or affinity proteomics, mass spectrometry, nuclear magnetic resonance (NMR) spectroscopy or protein chips;<br>• Data about the expression of genes in the form of RNA transcripts produced from human genomes (transcriptomics), generated from techniques such as microarrays or RNA-seq;<br>• Data about small molecules (metabolites) in a sample that reflect the metabolic state of the sample (metabolomics), generated from techniques such as various forms of mass spectrometry or NMR spectroscopy;<br>• Data about the lipid profile of a cell, person or group of people (lipidomics), generated from techniques such as various forms of mass spectrometry or NMR spectroscopy;<br>• Data generated through the use of multiple omics types and technologies, including the ones from the above categories, in order to study complex biological systems and processes and their impact on human health (multi-omics);<br>• Any other human-related molecular "-omic" data, such as for example microbiomics or exposomics.<br><br>It should be noted that the omics types and associated technologies listed above are not intended to provide a comprehensive overview of all possible data types that might fall under this category but rather serve to illustrate some common examples. |
| **Examples** | **Examples in scope:**<br>Raw data directly generated from experimental techniques<br>Cleaned or post-processed versions of the raw data<br>Analyses derived from the raw or cleaned data<br><br>**Examples out of scope:**<br>Any data that falls under the definition for category "(f) Human genetic, epigenomic, and genomic data"<br>Generic lab results performed in a primary healthcare setting |
| **Label** | Human molecular data |

### 5.3.8 Personal health data automatically generated by medical devices

**Table 8. Clarification of Article 51 (1)(h):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by RKKP (Denmark) and discussed during the TEHDAS2 workshops.

| Article 51 category | (h) Personal health data automatically generated by medical devices |
|---|---|
| **Legislative intent** | **Description:** This category covers health data generated by medical devices used by individuals. The focus is on data created during routine or specific device use that aids in monitoring or managing health conditions.<br><br>**Data in scope:** National Continuous glucose monitoring readings, data from wearable devices (e.g., heart rate, steps, oxygen saturation, sleep pattern), data from implantable devices (e.g. pacemakers).<br><br>**Data out of scope:** Data stored locally on devices without sharing or syncing to health data holders. |
| **Definition** | Digital information regarding a person's health that is produced automatically by a medical device.<br><br>Note: Personal electronic health data automatically generated through medical devices refers to digital information related to the physical or mental health of an individual, that is produced or collected by various technologies such as machines, instruments or implants. The medical device collects or gathers the information by itself (usually after an initial start or setup by the person themselves or by a professional), but at some point, shares the data with a health data holder. The generated health data may be used to provide real-time measurements or for tracking the patient over time, or for monitoring, diagnosis, or treatment. Health and fitness trackers are not included in the category. |
| **Supporting legislative definitions and resources** | "**Personal electronic health data**" means data concerning health and genetic data, processed in an electronic form (Article 2, point 2.a)<br>"**Data concerning health**" means personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status (Regulation (EU) 2016/679, Article 4, point 15)<br>"**Genetic data**" means personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question (Regulation (EU) 2016/679, Article 4, point 13)<br><br>"For the purposes of this Regulation, the following definitions apply: … (e) the definitions of 'medical device' … laid down in Article 2, points (1) … of Regulation (EU) 2017/745." (Article 2, point 1.e)<br>"**Medical device**" means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:<br>• diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease,<br>• diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability,<br>• investigation, replacement or modification of the anatomy or of a physiological or pathological process or state,<br>• providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations,<br><br>and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means. |

| | |
|---|---|
| | The following products shall also be deemed to be medical devices:<br>• devices for the control or support of conception;<br>• products specifically intended for the cleaning, disinfection or sterilization of devices as referred to in Article 1(4) and of those referred to in the first paragraph of this point. (Regulation (EU) 2017/745, Article 2, point 1)<br><br>For clarification on which software constitutes a medical device, consult<br>European Commission *"Is your software a medical device?"*<br>Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR |
| **Examples** | The following is a non-exhaustive list intended to provide some examples of medical devices and the types of data they automatically generate:<br><br>• **Smart Blood Pressure Monitors** that produce blood pressure readings at given intervals, including systolic and diastolic values.<br>• **Oxygen Saturation Monitors (Pulse Oximeters)** that continuously monitors blood oxygen saturation levels and pulse rates.<br>• **Smart Scales** that provide weight measurements along with body composition data (e.g., body fat percentage, muscle mass)<br>• **Remote Patient Monitoring Devices** that monitor various health parameters and transmit data to healthcare providers, e.g. respiratory rate or pacemaker measurements.<br>• **Continuous Glucose Monitors (CGMs)** that collects data on real-time glucose levels, trends over time, and alerts for high or low blood sugar.<br>• **Smart Insulin Pumps** that automatically delivers insulin and collects data on bolus doses administered.<br>• **Sleep Apnea Monitors** that track breathing patterns, oxygen levels, and sleep quality during the night.<br>• **Cardiac Monitors/Heart Rate Monitors** that monitors heart rhythms continuously and detects irregularities like arrhythmias.<br>• **Electrocardiogram (ECG/EKG) Devices** that record the electrical activity of the heart.<br>• **Smart Thermometers** that read body temperature.<br>• **Smart Inhalers** that track medication usage for individuals with asthma or other respiratory conditions**.**<br><br>Examples of excluded datasets:<br>• Data from non-medical wellness apps<br>• Manually entered health data<br>• Self-reported health information<br>• Data that is only stored locally on devices and not shared or synchronized to a healthcare network or a health data holder<br>• Digital video and sound from medical teleconsultations and operating rooms<br>• Data exchanged during robotic operation for performance control purposes<br>• Virtual reality models created without relation to a person, e.g. based on synthetic data |
| **Label** | Medical Devices data |

### 5.3.9 Data from wellness applications

**Table 9. Clarification of Article 51 (1)(i):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by RKKP (Denmark) and discussed during the TEHDAS2 workshops.

| Article 51 category | (i) Data from wellness applications |
|---|---|
| **Legislative intent** | **Description:** This category covers health-related data generated by wellness applications designed to promote physical and mental well-being. Emphasis is on voluntary user input or device-synced information.<br><br>**Data in scope:** User-entered data (e.g. diet logs, mood tracking, sleep tracker), sensor data (e.g. step count, heart rate from fitness apps) accessible to the holder.<br><br>**Data out of scope:** Data not linked to health purposes (e.g. pure entertainment apps) / Locally stored data without access possibility (as in: not held by the provider). |
| **Definition** | Digital information about the health, behaviour, wellbeing or lifestyle of an individual collected by a wellness application.<br><br>Note: Data from wellness applications can come from any software, or any combination of hardware and software, such as a computer program, an app or a technology like an appliance, instrument or a wearable, and are at some point, shared with a health data holder. They provide users with insights into their health, behaviour and wellness status, often incorporating feedback and personalized recommendations. Wellness applications are used for processing electronic health data for other purposes than healthcare, such as well-being and pursuing healthy lifestyles. |
| **Supporting legislative definitions** | "**Wellness application**" means any software, or any combination of hardware and software, intended by the manufacturer to be used by a natural person, for the processing of electronic health data, specifically for providing information on the health of natural persons, or the delivery of care for purposes other than the provision of healthcare. (Article 2, point 2.ab) |
| **Examples** | The following is a non-exhaustive list of examples of different types of data that are part of the category:<br>• **Physical activity data**, e.g. step count, distance travelled, time spent on activities or workouts and type of activities, calories burned, heart rate.<br>• **Sleep related data**, e.g. bedtime, wake timer and sleep duration, sleep stages, sleep quality and interruptions.<br>• **Nutrition data**, e.g. calorie intake, water intake and meal logging.<br>• **Mental health data**, e.g. mood tracking, meditation, breathing exercises or mindfulness sessions and stress levels.<br>• **Biometric data**, e.g. weight, body mass index (BMI), body fat percentage, blood pressure.<br>• **Goals and lifestyle data**, e.g. goals tracking, progress reports, achievements, daily routine or habit tracking and work-life balance.<br>• **Symptom and health data**, e.g. medical reminders, scheduling and appointments, symptom logs.<br><br>**Note:** The same type of data, such as step count or heart rate, may fall under category (h) or (i) depending on whether it is generated by a certified medical device or a non-medical wellness application.<br><br>**Examples of excluded datasets:**<br>• Data from regulated medical devices<br>• Clinical health records<br>• Clinical diagnostic data<br>• Data not collected for health purposes, e.g. collected from apps for entertainment or educational purposes<br>• Data that is only locally stored and thereby not held by the app provider |

| | |
|---|---|
| | • Data from apps that enable easier communication in emergency cases<br><br>**Note:** Wellness applications, as defined in Article 2(2)(ab) of the EHDS Regulation, must be clearly distinguished from regulated medical devices (Recital 57). For example, applications used for clinical diagnosis or monitoring are classified under category (h), not (i). |
| **Label** | Wellness applications |

### 5.3.10 Data on professional status, and on the specialisation and institution of health professionals involved in the treatment of a natural person

**Table 10. Clarification of Article 51 (1)(j):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by HDH (France) and discussed during the TEHDAS2 workshops.

| **Article 51 category** | **(j) Data on professional status, and on the specialisation and institution of health professionals involved in the treatment of a natural person** |
|---|---|
| **Legislative intent** | **Description:** This category covers individual-level data on health professionals directly involved in the treatment of a natural person, including their professional status, qualifications, specialisation, and their affiliated institutions. The intent is to enable analysis of care pathways and treatment outcomes, such as referrals between professionals, which may not be fully captured in clinical records alone.<br><br>**Data in scope:** Information on treating professionals' status (e.g. GP, specialist), specialisation and qualifications of health professionals involved in care, affiliated institutions (e.g. hospital or clinic where care was delivered).<br><br>**Data out of scope:** Data on professionals not involved in treatment, any other information about those health professionals (names, financial, contact details...), generic workforce or licensing data unrelated to patient care delivery. |
| **Definition** | Individual level data on health professionals. These data encompass professional status, specialization,organizational affiliations, work location and affiliations status to health care providers for physicians (EU Regulation 2022/2294), nurses (EU Regulation 2022/2294) or any professional directly involved in treatment. It also covers registry-based data on the professional status, educational level, and specialization of healthcare staff at a broader population level. |
| **Supporting sources** | A System of Health Accounts 2011: revised edition<br>Health Care financing scheme (Link)<br>Health Care functions expenditures: (Link)<br>Aggregate definition (Eurostat): (Link)<br>Definition of *Nurse*: EU Regulation 2022/2294<br>Definition of *Physicians*: EU Regulation 2022/2294<br>Commonly agreed definitions of other healthcare professionals are available in the Eurostat Manual (Link)<br>Definition of *Healthcare providers*: (Link) |
| **Examples** | Examples in scope:<br>• RPPS - Shared directory of health system professionals (France, holder: Agence du Numérique en Santé, ANS) *Example 1*<br>• AGB Register (Data Management Register for the Healthcare Industry, Netherlands) - contains the codified data of healthcare providers in the Netherlands. *Example 2* |
| **Label** | Aggregated healthcare data |

### 5.3.11 Data from population-based health data registries such as public health registries

**Table 11. Clarification of Article 51 (1)(k):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by THL (Finland) and discussed during the TEHDAS2 workshops.

| Article 51 category | (k) Data from population-based health data registries such as public health registries |
|---|---|
| **Legislative intent** | **Description:** This category covers systematically collected data in registries established for public health purposes, typically maintained by public authorities and often mandated by law. These population-based registries aim to support epidemiological surveillance, disease monitoring, and health system planning, rather than individual care.<br><br>**Data in scope:** Cancer registries and infectious disease registries maintained by public health agencies, legally mandated national or regional registries tracking disease incidence or health indicators, population-wide registries used for surveillance and public health reporting.<br><br>**Data out of scope:** Clinical data collected solely for individual treatment, registries not intended for population-level monitoring. |
| **Definition** | Data from registries containing population-based health data. Registries are organised systems, such as databases that are used to collect and store specific types of information that are usually related in some way.<br><br>Population-based health data registries are systematic collections of health-related data from defined populations. The focus on the data collection is the health status of the defined population, often categorized by geographic areas or specific demographic groups. The aim of these registries is to track health outcomes and trends over time, and for these purposes, all new cases are continuously recorded, and data reflecting the health status in defined groups is gathered. These registries are typically used for epidemiology, public health surveillance, prevention, and to support public health decision-making, policy development, and healthcare planning.<br><br>Population-based health data registries encompass extensive data on health-related events, healthcare utilization, interventions, diseases, and conditions across diverse demographics. They are often maintained by public health authorities, government agencies, health organizations, or research institutions.<br><br>Not included in this category are hospital-based register data on individuals who seek care at specific hospitals, and clinical and medical registries that do not aim to cover the entire defined population. |
| **Examples** | Examples in scope:<br>• Finnish Care Register for Health Care (Terveys-Hilmo)<br>• Finnish Register of Primary Health Care Visits (AvoHilmo)<br>• Swedish National Patient Register<br>• Finnish Cancer Registry<br>• Netherlands Cancer Registry<br>• Cyprus Cancer Registry<br>• Finnish National Vaccination Register<br>• Medical Birth Registry of Norway<br>• Finnish National Infectious Diseases Register |
| **Label** | Population-based health register data |

### 5.3.12 Data from medical registries and mortality registries

**Table 12. Clarification of Article 51 (1)(l):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by THL (Finland) and discussed during the TEHDAS2 workshops.

| Article 51 category | (l) Data from medical and mortality registries |
|---|---|
| **Legislative intent** | **Description:** To cover (1) more targeted registries (i..e. not population-based) and (2) mortality registries (which aren't *health* registries as per point (k) in the first place) |
| **Definition** | Data from medical and mortality registries. Registries are organised systems, such as databases that are used to collect and store specific types of information that are usually related in some way.<br>• **Medical registries** are systematic collections of data on patients with a specific disease, condition, or characteristic. Medical registries collect detailed clinical information on patients diagnosed with specific diseases or undergoing specific treatments within the healthcare system. These registries are typically focused on patient outcomes, treatment effectiveness, and healthcare quality improvement, and their aim is to help guide clinical practice and improve patient care. Data may be sourced from multiple origins, such ashealthcare providers and patient surveys.<br>• **Mortality registries** are systematic collections of data on deaths, often including information on the cause, circumstances, and other factors related to mortality, as well as demographics. Cause of death registries, which contain data on the underlying cause of death, including information on diseases, injuries, or conditions, fall into this category.<br><br>The main difference between population-based health data registries (k) and medical registries (l) is that medical registries are usually more specialized, focusing on specific diseases or events and tracking their clinical outcomes and mortality data. Medical registries may also have less broad coverage than population-based health data registries, as they concentrate on patient groups rather than entire populations. |
| **Examples** | Examples in scope:<br>• Hospital District of Helsinki and Uusimaa (HUS) Quality Registers<br>• Finnish Spine Register (national quality register)<br>• Swedish National Quality Registry for Heart Failure<br>• Dutch Quality Register of Rheumatoid Arthritis<br>• Danish Lung Cancer Register (DLCR)<br>• Norwegian Pancreatic Cancer Registry<br>• Dutch Pediatric and Adult Registry of Diabetes<br>• Swedish National Cause of Death Register |
| **Label** | Medical and mortality register data |

### 5.3.13 Data from clinical trials, studies, and investigations under relevant EU legislation

**Table 13. Clarification of Article 51 (1)(m):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by RKKP (Denmark) and discussed during the TEHDAS2 workshops.

| Article 51 category | (m) Data from clinical trials, studies, and investigations under relevant EU legislation |
|---|---|
| **Legislative intent** | **Description:** It specifically refers to datasets collected in the context of clinical research activities governed by EU legislation (e.g. the Clinical Trials Regulation), distinguishing them from routine clinical care data, even when the content may overlap. The intent is to enable the secondary use of clinical research data, including data from non-conclusive or discontinued trials, for purposes beyond the original research objectives.<br><br>**Data in scope:** Completed clinical trials, studies, and investigations, data from non-conclusive or discontinued trials, post-trial datasets available after protection periods expire.<br><br>**Data out of scope:** Ongoing trials, trials still under data protection or exclusivity periods. |
| **Definition** | Data collected in the context of clinical studies, regulated under EU guidelines. Clinical studies include clinical observational studies (case studies, real-world-data studies, and cohort studies), clinical trials (randomized and non-randomized controlled trials), and clinical investigations.<br><br>Note: A **clinical trial** is any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects (treatment efficacy, effectiveness, and safety) on health outcomes. Clinical trials may also be referred to as interventional trials. Interventions include but are not restricted to drugs, cells and other biological products, surgical procedures, radiologic procedures, devices, behavioural treatments, process-of-care changes, and preventive care. A **clinical investigation** is a clinical study conducted to evaluate the clinical utility and effectiveness of interventions, which may include randomized pilot studies, feasibility studies, and post-marketing surveillance. |
| **Supporting legislative definitions** | The data collected are vital for regulatory approval processes, advancing medical knowledge, evaluating medical intervention, and improving patient care by providing evidence-based insights into treatment effectiveness and safety. Article 51 currently limits the category to clinical trials, clinical studies, and clinical investigations subject to Regulation (EU) 536/2014, Regulation [SOHO], Regulation (EU) 2017/745 and Regulation (EU) 2017/746, respectively. In particular, the category includes clinical studies and clinical trials as defined in the Regulation (EU) No 536/2014:<br><br>(1) 'Clinical study' means any investigation in relation to humans intended:<br>(a) to discover or verify the clinical, pharmacological or other pharmacodynamic effects of one or more medicinal products;<br>(b) to identify any adverse reactions to one or more medicinal products; or<br>(c) to study the absorption, distribution, metabolism and excretion of one or more medicinal products;<br>with the objective of ascertaining the safety and/or efficacy of those medicinal products;<br><br>(2) 'Clinical trial' means a clinical study which fulfils any of the following conditions:<br>(a) the assignment of the subject to a particular therapeutic strategy is decided in advance and does not fall within normal clinical practice of the Member State concerned;<br>(b) the decision to prescribe the investigational medicinal products is taken together with the decision to include the subject in the clinical study; or<br>(c) diagnostic or monitoring procedures in addition to normal clinical practice are applied to the subjects. |
| **Examples** | The data described and elaborated in the examples below are examples of the category clinical studies, clinical trials, and clinical investigations and include different types of data that |

| | |
|---|---|
| | could belong to this category. However, it should be emphasized that these are only examples and not an exhaustive list of the possible data within the category.<br><br>• **Basic Participant Information:** Age, gender, racial and ethnic background, information on participants' health status before the study starts.<br>• **Clinical Trial Data:** A clinical trial studies the changes in the outcome or course of a condition or disease to prevent harm or improve health through the use of treatments, medicinal products, medical devices or procedures/surgery. Data about the clinical trial is available in a treatment protocol and includes the objectives of the intervention, design (selection, randomization), number of participants needed, measurements, endpoints, methodology, anticipated bias, safety matters, and privacy of the data.<br>• **Outcome Measures:** Primary outcomes (main results the study aims to measure, e.g., reduction in disease symptoms, survival rates), secondary outcomes (additional effects that are measured, e.g., quality of life assessments, side effects), and adverse events (any unfavorable medical occurrence in a patient, or clinical trial participant receiving a medicine, and which does not necessarily have a causal relationship with this treatment).<br>• **Clinical Assessments:** Laboratory Test Results (blood tests, imaging results, and other diagnostic evaluations conducted during the study), data from wearables or apps - digital therapeutics/digital medical devices, telemonitoring devices, and clinical ratings (scores from scales used to assess symptoms or health conditions, e.g., depression scales, pain scales).<br>• **Follow-Up Data:** Longitudinal Data (information collected at various points during and after the study to assess long-term effects and outcomes) and survival data (data on the time until an event occurs, e.g., disease progression, death).<br>• **Patient Reported Outcome (Measure) (PRO/PROM):** Patient Reported Outcomes (PROs) are data reported directly by a patient on his or her own health condition, without interpretation by a doctor or anyone else. Patient-reported outcome measures (PROMs) are the tools used to measure and collect data on PROs. Outcome (PRO) instruments include surveys and questionnaires (information provided by participants regarding their health status, symptoms, quality of life, and patient satisfaction) and daily logs (participant-reported data on medication adherence, side effects, and daily functioning).<br>• **Compliance and Adherence Data:** Medication adherence (data on whether participants took their medication as prescribed) and protocol deviations: (records of any deviations from the study protocol, such as missed doses or withdrawal from the study).<br>• **Biological Samples Data:** Biospecimens - information on samples collected for analysis (e.g., blood, tissue, urine) and their results.<br>• **Genomic/Genetic/Molecular Data**: Samples and test results from genomic/genetic/molecular tests stored in biobanks and genomic data repositories.<br>• **Quality Control Data:** Quality Control (QC) is part of the system of ensuring high standards during research, trials and production for medicines. The data includes monitoring reports (data collected to ensure the study is conducted according to regulatory standards and protocols) and audit findings (records of inspections and findings from quality assurance processes).<br>• **Economic Data:** Cost-effectiveness data (information on the costs associated with interventions compared to outcomes achieved) and healthcare utilization data (data on healthcare services used by participants during the study).<br>• **Administrative Data** related to the clinical research conducted.<br><br>Examples of datasets or metadata descriptions of datasets:<br>An example of a metadata description is the Danish Medicines Agency: *Example 1*<br><br>Examples of excluded datasets:<br>• Routine EHR data<br>• Registry data outside a clinical study context.<br>• Synthetic health data (included in studies)<br>• Non-human data (no human participants in a study) |
| **Label** | Clinical study data |

### 5.3.14 Other health data from medical devices

**Table 14. Clarification of Article 51 (1)(n):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by RKKP (Denmark) and discussed during the TEHDAS2 workshops.

| Article 51 category | (n) Other health data from medical devices |
|---|---|
| **Legislative intent** | **Description:** This category covers a broader scope than category (h), including additional health data derived from non-wearable or diagnostic medical devices.<br><br>**Data in scope:** Imaging data (e.g., MRI or X-ray scans) stored in healthcare systems, data captured during diagnostic tests by medical equipment (e.g. lab results, ECG).<br><br>**Data out of scope:** Non-digital or paper-based test results / Device-specific logs unrelated to health metrics (e.g., error or maintenance reports) |
| **Definition** | Health data generated by medical devices, including manually entered or processed outputs, such as data from imaging or diagnostic equipment, particularly when human interpretation or intervention is involved.<br><br>Note: The health data in this category can be interpreted as patient-related and other information derived from medical devices, e.g., imaging data (MRI or X-ray scans) and other data resulting from further layers of data processing, or as data captured during diagnostic tests by medical equipment, e.g., laboratory and other test results, involving manual intervention. |
| **Supporting legislative definitions** | "**Medical device**" means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:<br>• diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease,<br>• diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability,<br>• investigation, replacement or modification of the anatomy or of a physiological or pathological process or state,<br>• providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations,<br><br>and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means.<br><br>The following products shall also be deemed to be medical devices:<br>• devices for the control or support of conception;<br>• products specifically intended for the cleaning, disinfection or sterilization of devices as referred to in Article 1(4) and of those referred to in the first paragraph of this point.<br>(Regulation (EU) 2017/745 on medical devices, Article 2, point 1) |
| **Examples** | The data described and elaborated in the examples below are examples of the category 'Other health data from medical devices' and include different types of data that could belong to this category. The examples demonstrate how medical devices collect health data, often requiring manual steps or advanced processing (e.g., software algorithms) to interpret or enhance the data for clinical use. However, it should be emphasized that these are only examples and not an exhaustive list of the possible data within the category.<br><br>• **Blood Pressure Readings**: A healthcare provider manually inflates a cuff to measure blood pressure and records the data. |

| | |
|---|---|
| | • **Temperature Measurements**: A thermometer that is manually used to record body temperature, which is subsequently entered into the patient's health record.<br>• **Manual Pulse Oximetry**: A nurse or patient manually uses a pulse oximeter to measure and read oxygen saturation levels and records the result.<br>• **Processed output data from ECG**: An electrocardiogram (ECG) device records the electrical activity of the heart and processes the raw data to generate a visual output, which is then interpreted by a healthcare provider.<br>• **Glucose Monitoring**: A continuous glucose monitor (CGM) collects glucose level data and uses built-in algorithms to provide alerts or predictions, which are then reviewed by healthcare providers.<br>• **Heart Rate Monitors**: A wearable device measures heart rate and processes the data to detect abnormal rhythms or fluctuations, alerting users or providers.<br>• **MRI Imaging**: An MRI machine captures images of internal body structures and uses data processing to generate detailed images that are then analyzed by radiologists for diagnosis.<br>• **Smart Wearables**: Devices like smartwatches that track steps, heart rate, and sleep patterns, then process this data to provide insights on health trends or potential issues, such as irregular heart rhythms.<br>• **Pulse Oximeters with Data Analysis**: Devices that not only measure oxygen levels but also analyze trends over time, providing actionable insights for healthcare providers.<br>• **Device Usage Logs and Service Data**: Information on the use of the device and the service provided.<br>• **Post-processing of Health Data:** Data generated by AI algorithms from multiple sources of health data generated by medical devices.<br><br>Examples of excluded datasets:<br>• Data from non-medical wellness devices, such as fitness apps<br>• Health records and general health monitoring unrelated to medical devices<br>• Non-digital or paper-based test results<br>• Device-specific logs unrelated to health metrics (e.g., error reports)<br>• Administrative data such as billing and insurance claims<br>• Artificially generated data (synthetic data)<br>• Data not within health services, e.g. for educational and training purposes |
| **Label** | Other health data medical devices |

### 5.3.15 Data from registries for medicinal products and medical devices

**Table 15. Clarification of Article 51 (1)(o):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by RKKP (Denmark) and discussed during the TEHDAS2 workshops.

| Article 51 category | (o) Data from registries for medicinal products and medical devices |
|---|---|
| **Legislative intent** | **Description:** This category covers data collected systematically in registries aimed at monitoring the use, safety and effectiveness of medicinal products and medical devices<br><br>**Data in scope:** Safety data (e.g. side effects, adverse events) / Efficacy studies (e.g. effectiveness of specific drugs or devices) / Usage patterns of medicinal products.<br><br>**Data out of scope:** General aggregate statistics not tied to health purposes. |
| **Definition** | Data from registries for medicinal products and medical devices refer to systematically collected data on the use, safety, effectiveness, and performance of medicinal products and medical devices.<br><br>Note: For medicinal products, registries track outcomes, adverse reactions, and long-term effects of drugs. For medical devices, they monitor safety, functionality, and post-market performance. These registries help ensure public safety, support regulatory decisions, and facilitate ongoing research and improvement of these products. |
| **Supporting legislative definitions** | **"Medicinal product"** means any substance or combination of substances presented for treating or preventing disease in human beings. Any substance or combination of substances which may be administered to human beings with a view to making a medical diagnosis or to restoring, correcting or modifying physiological functions in human beings is likewise considered a medicinal product. (Directive 2001/83/EC, Article 1, point 2)<br><br>"**Medical device**" means any instrument, apparatus, appliance, software, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes:<br>• diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease,<br>• diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability,<br>• investigation, replacement or modification of the anatomy or of a physiological or pathological process or state,<br>• providing information by means of in vitro examination of specimens derived from the human body, including organ, blood and tissue donations,<br><br>and which does not achieve its principal intended action by pharmacological, immunological or metabolic means, in or on the human body, but which may be assisted in its function by such means.<br><br>The following products shall also be deemed to be medical devices:<br>• devices for the control or support of conception;<br>• products specifically intended for the cleaning, disinfection or sterilization of devices as referred to in Article 1(4) and of those referred to in the first paragraph of this point. (Regulation (EU) 2017/745 on medical devices, Article 2, point 1) |
| **Examples** | The data described and elaborated in the examples below are examples of the category 'data from registries for medicinal products and medical devices' and include different types of data that could belong to this category. However, it should be emphasized that these are only examples and not an exhaustive list of the possible data within the category.<br><br>For medicinal products: |

| | |
|---|---|
| | • Adverse Event Reports: Data on negative side effects or reactions experienced by patients using the medicinal product.<br>• Treatment Outcomes: Information on how well the medicinal product is working in real-world settings, including improvements or failures in treating a disease or condition.<br>• Usage Patterns: Data on the frequency, dosage, and duration of use of the medicinal product by different populations.<br>• Efficacy Data: Information on the effectiveness of the medicinal product in achieving its intended therapeutic goals.<br>• Patient Demographics: Data on the age, gender, health status, and other demographic factors of patients using the medicinal product.<br>• Prescription Data: Information about how frequently and where the medicinal product is being prescribed.<br><br>For Medical Devices:<br>1) **Device Performance Data**: Information on how well a device functions during real-world use, including any malfunctions or failures.<br>2) **Adverse Event Reports**: Data on complications, injuries, or other adverse outcomes associated with the use of medical devices.<br>3) **Post-Market Surveillance Data**: Data collected after the device has been marketed, tracking its ongoing safety and effectiveness.<br>4) **Patient Outcomes**: Information on how the use of the medical device impacts patient health, including recovery rates or quality of life improvements.<br>5) **Device Longevity and Wear**: Data on the durability and lifespan of medical devices, such as implants or prosthetics.<br>6) **Recall Information**: Data related to any recalls of medical devices, including reasons for the recall and actions taken.<br><br>Examples of datasets or metadata descriptions of datasets:<br>Netherlands Pharmacovigilance Centre Lareb (side-effects of medication) *Example 1*<br>Norwegian Adverse Drug Reaction Registry *Example 2*<br>Norwegian Pacemaker and ICD Registry *Example 3*<br><br>Examples of excluded datasets:<br>Datasets including patient health records<br>Datasets that do not directly relate to the safety, performance, or efficacy of approved medicinal products or medical devices<br>Marketing-oriented dataset |
| **Label** | Registries medicinal products & medical devices |

### 5.3.16 Data from research cohorts, questionnaires and surveys related to health, after the first publication of the related results

**Table 16. Clarification of Article 51 (1)(p):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by THL (Finland) and discussed during the TEHDAS2 workshops.

| Article 51 category | (p) Data from research cohorts, questionnaires and surveys related to health, after the first publication of the related results; |
|---|---|
| **Legislative intent** | **Description:** Information collected from groups of individuals or populations to understand health-related phenomena, behaviours, or outcomes. These data are often used to identify risk factors, track trends, or evaluate the effectiveness of public health interventions. The goal is to 'open up' such data, which might otherwise not get reused.<br><br>**Data in scope:** Only data from research projects that have yielded at least partially publicly available results.<br><br>**Data out of scope:** Such data, before the first publication of results (then it's out of scope of obligation to declare to the catalogue). |
| **Definition** | Research data refers to data from health research cohorts or studies that is collected, recorded, studied, processed, or created to produce original research results and answer research questions. This data typically originates from sources such as surveys, experiments, observations, and existing databases. Health-related research data is often collected from groups of individuals or populations to understand health-related phenomena, behaviours, or outcomes. This data is typically used to identify risk factors, track trends, or evaluate the effectiveness of public health.<br>• Such data are included in this category once the research results have been published for the first time. Research publication involves making research findings available to the general public through scientific journals, conference proceedings, or other academic platforms. This process typically includes peer review to ensure the quality and reliability of the research.<br><br>A research cohort is a group of individuals who share a common characteristic or experience, which may be within a defined period and may be followed over time, to study specific outcomes, typically related to particular exposures or interventions. The definition of 'research cohort' includes both prospective and retrospective cohorts. Prospective cohorts are followed forward in time from the point of recruitment, while retrospective cohorts are identified from existing records and followed backward in time. A research cohort can also be population-based, drawn from the general population. These cohorts are often maintained by public sector entities or other organisations, either within the health or academic sector.<br>It should be noted that research datasets that merely duplicate existing health data registries, without any transformation, aggregation, or derivation, remain covered only under their original Article 51 category (e.g. (a), (e), (l)) and do not fall under category (p). In contrast, once a dataset involves processing, aggregation, or the creation of derived variables and is published for secondary use, it qualifies as a derived research dataset and falls under category (p). The original source datasets shall remain available individually. |
| **Examples** | Examples in scope:<br>• Healthy Finland research data<br>• School Health Promotion study 2019<br>• National FinHealth 2017 Study data<br>• Manifold more interviews 2021<br>• Northern Finland Birth Cohorts \| University of Oulu<br>• Norwegian Counties Public Health Surveys<br>• The Tromsø Study<br>• The German National Cohort (NAKO) |
| **Label** | Research data |

### 5.3.17 Data from biobanks and associated databases

**Table 17. Clarification of Article 51 (1)(q):** Includes the full-text definition, interpretation of the legislative intent, and practical examples. The definition and examples were developed by THL (Finland) and discussed during the TEHDAS2 workshops.

| Article 51 category | (q) Data from biobanks and associated databases |
|---|---|
| **Legislative intent** | **Description:** This category covers data derived from biological samples stored in biobanks, such as genomic, proteomic or metabolic data, as well as related health data (e.g., medical records, lifestyle information) that are collected and stored for research purposes. The regulation strictly covers existing derived data stored within biobanks or associated databases that are accessible for secondary use under the EHDS framework. Any request for additional analysis or generation for new data from samples falls outside the regulatory scope.<br><br>**Data in scope:** Data already derived from biological samples (e.g. genomic or proteomic data), health data associated with these samples (e.g. medical records, environmental exposures), data from population-based, disease-specific or tissue biobanks, as long as it is existing and stored for secondary use.<br><br>**Data out of scope:** Physical biological samples (e.g. DNA, tissue, blood), data that has not yet been generated (e.g. requiring new analysis on samples), administrative metadata unrelated to research purposes |
| **Definition** | According to Article 1 of the European Commission Implementing Decision 2013/701/EU, the term "Biobanks" refers to collections, repositories, and distribution centers of various types of human biological samples, including but not limited to blood, tissues, cells, or DNA, as well as related data such as associated clinical and research information. Additionally, it encompasses biomolecular resources, including model organisms and microorganisms, which may contribute to the understanding of human physiology and diseases.<br><br>Biobanking involves the process of acquiring and storing these biological materials, along with activities related to their collection, preparation, preservation, testing, analysis, and distribution. This comprehensive approach ensures that defined biological materials, as well as related information and data, are managed effectively to support research and clinical applications.<br><br>Health data from biobanks and associated databases can include, but may not be limited to:<br><br>• Data about samples: General information regarding the samples stored in a biobank and intended for research purposes.<br>• Data from samples: Information obtained through the analysis of samples stored in a biobank, or other data associated with the samples and acquired through various means.<br>• Data about individuals: Information necessary for biobank research pertaining to the individual from whom the sample was collected.<br><br>Additionally, biobanks may store:<br><br>• Data on the biobanks themselves: Information about the biobank's infrastructure, management, and operational procedures. This data is not considered health data unless it impacts the research conducted on the samples.<br><br>In more detail, the health data stored in biobanks can include:<br>• Data on the preanalytical processing of samples, encompassing detailed records of the conditions under which samples were collected, processed, and stored, including all associated quality parameters. |

| | |
|---|---|
| | • Genomic data and epigenomic data, such as DNA/RNA and exome sequences, genotypes, gene expression levels, epigenotypes, data from methylation scan, and histone modification.<br>• Proteomics and metabolomics data, such as protein expression levels in cells, and metabolites and their concentration in cells.<br>• Sociodemographic data, such as age, gender, place of residence, education, and migration background.<br>• Clinical data and longitudinal health data from surveys and follow-ups, including but not limited to diagnoses, treatments, and laboratory results.<br>• Lifestyle data, such as diet, smoking, alcohol consumption, physical activity, nutrition, drug use etc.<br><br>**Excluded from this category are:**<br>Physical human biological samples (e.g., DNA, tissue, blood)<br>Data that has not yet been generated (e.g., requiring new analysis on samples)<br>Administrative metadata unrelated to research purposes.<br>Data of biobanks themselves, unless it impacts the research conducted on the samples |
| **Examples** | Examples in scope:<br>• Norwegian Mother, Father and Child Cohort Study (MOBA)<br>• Collections available in Sample Locator of German Biobank Node<br>• Leiden Longevity Study<br>• LifeLines *Link*<br>• PALGA: *Link*<br>• GEOLynch *Link* |
| **Label** | Biobank data |

# 6 HealthDCAT-AP: Why is this the proposed common metadata model for EHDS for secondary use

In Chapter 5, the question “Which data should be described under the EHDS for secondary use?” was addressed. Building on that, this chapter responds to *how* the data should be described by providing an overview of the legal context and the framework for establishing a common metadata model for secondary use under the EHDS Regulation.

## 6.1 Legal context

The EHDS Regulation establishes clear legal requirement for the **description and cataloguing of datasets made available for secondary use**. These obligations are set out in several key provisions of the Regulation:

- **Article 60** provides that health data holders must communicate a description of the datasets they hold to the competent HDAB and must verify at least once per year that this description is accurate and up to date in the national dataset catalogue.

- **Article 77** requires HDABs to maintain and make available a harmonised, publicly accessible, machine-readable dataset catalogue. Each dataset must be described *in the form of metadata* and include essential information on the source, scope, main characteristics, type of the data, and the conditions for access.

- **Article 2** provides the relevant definitions underpinning these provisions:
  - A **dataset** is defined as a structured collection of electronic health data (article 2(2)(s))
  - A **dataset catalogue** is defined as a systematically organised collection of dataset descriptions, including a user-oriented public part accessible via an online portal. (Article 2(2)(u))

Taken together, these provisions require health data holders to provide dataset descriptions in a **structured and common format**, and HDABs to ensure the **cataloguing, accessibility, and discoverability** of these descriptions at national level. Moreover, **Article 77(4)** empowers the European Commission to adopt an implementing act that specifies **the minimum set of metadata elements** that must be included in these dataset descriptions and their characteristics. This mechanism enables the development of a **common metadata model**, applicable across Member States, to support **semantic and technical interoperability**, enhance transparency, and facilitate cross-border data reuse..

In this context, it is necessary to adopt and implement a robust and harmonised metadata model capable of accommodating the structured description of the wide variety of health datasets covered under article 51 the EHDS. This objective underpins the development of **HealthDCAT-AP** and the ongoing efforts to refine and validate the model, ensuring it is fit for purpose as the common metadata model for secondary use within the EHDS framework.

## 6.2 Understanding Metadata in the EHDS context

### 6.2.1 What is metadata and why do we need a metadata model in the EHDS for secondary use?

In the context of the EHDS for secondary use, **metadata** is structured information that describes a dataset, including its origin, content, structure, update frequency, and intended use. Metadata helps data users understand what the dataset contains, how it was generated, and whether it is suitable for their purpose - all without needing direct access to the data itself.

To better understand metadata, it's essential to distinguish it from data itself:

- **Data** refers to raw facts or observations - such as numbers, text, images, or clinical measurements - that are collected and used for analysis. It is the actual data collected that tells what was measured and the result (e.g., a patient's diagnosis, lab result, or prescribed medication).

- **Metadata** provides essential context about the data, such as its source, format, structure, creation date, author, or how it should be interpreted and reused. It describes the data that was collected by telling how it was measured, by whom, under what conditions, and how to interpret the result reliably (e.g., when and how the diagnosis was recorded, which coding system was used, or the data format). It is important to distinguish between metadata describing individual data elements (e.g. metadata of an EHR document) and metadata describing datasets as a whole. In the context of secondary use under the EHDS, HealthDCAT-AP is used to describe metadata of datasets, rather than of individual data entries.

This distinction is particularly important in the EHDS context, where metadata plays a critical role in enabling secondary use. By providing this structured, descriptive information, metadata allows data users to assess the relevance and quality of datasets before applying for access, supporting transparency, trust, and effective reuse of health data across borders. Importantly, this document uses the terms **metadata model** and *c***ommon metadata model**. A metadata model defines a structured framework and sets of rules for organising and representing descriptive information about data. In the context of the EHDS, a common metadata model refers to the harmonised metadata structure that will be adopted via implementing acts pursuant to Article 77(4) of the Regulation, to ensure semantic and technical interoperability across systems and countries. A comparison between data and metadata for different health-domain scenarios is provided in Table 18.

**Table 18.** Comparison of data and metadata for different types of health datasets - EHR-based and image-based datasets.

| Data (what) | Metadata (context about the data) |
|---|---|
| **EHR-based dataset** | |
| Patient ID | Data source (hospital/clinic) ; Population coverage |
| Diagnosis: Type-2 Diabetes | Coding system used: ICD-10 (E11); version of the codebook |
| Medication: Metformin 500mg | Dosage unit (mg); administration route (oral); prescribing physician ID (pseudonymised) |
| Visit date: 2024-12-01 | Timestamp format (e.g., ISO 8601); time zone; type of encounter (in-person, telehealth) |
| Blood pressure: 130/85 mmHg | Measurement method (manual/automated); measurement position (seated/standing); device model |
| Allergy: Penicillin | Allergy type (drug); source of record (patient-reported vs. confirmed); coding system (SNOMED CT) |
| Vaccination: COVID-19 (Pfizer) | Batch number; administration site; manufacturer; data entry timestamp |
| **Image-base dataset** | |
| X-ray image file (e.g., DICOM) | File format: DICOM; resolution (2048x2048); image compression method |
| Image modality: Chest X-ray | Modality code (e.g., DCM: CR); radiology procedure type |
| Imaging date: 2025-02-10 | Time of acquisition; time zone; imaging session ID |
| Diagnosis: No abnormality detected | Interpretation date; radiologist ID (pseudonymised); confidence score |
| Body position: Standing, posterior-anterior (PA) view | Positioning protocol; radiology technician ID (pseudonymised) |
| Radiation dose: 0.12 mSv | Dose measurement method; device model and calibration date |

In the context of the EHDS Regulation, particularly for the secondary use of health data, metadata plays a crucial role in ensuring that data can be understood, trusted, and reused.

- **Metadata enables interoperability**: By using standardised elements such as coding systems (e.g., ICD, SNOMED CT), timestamps, and measurement units, metadata allows

datasets from different systems, organisations, and countries to be combined and compared reliably.

- **Metadata supports data quality, transparency, and traceability**: It documents where the data came from, how it was collected or processed, and under what conditions, helping data users assess whether the dataset is suitable for their specific research or policy questions.

**Box 1. What does it mean to *"provide a description in the form of metadata"*? (Article 77)**

The requirement to provide a dataset description "in the form of metadata" highlights the need for harmonisation and structure. This means that descriptions must go beyond informal or unstructured formats and follow a recognised metadata schema.

This does not include:

- Free-text descriptions (e.g. in a Word document or email), informal documentation or notes, oral explanations , raw data files without contextual information

It does require:

- Structured information, organised into defined metadata fields (e.g. title, source, date of collection, coding system)
- Machine-readability, to enable automated processing and discovery (e.g. RDF, XML, JSON)
- Use of a formal metadata schema, such as HealthDCAT-AP, to ensure interoperability and consistency

In summary, the EHDS regulation expects a properly structured and machine-readable metadata description, not just any form of dataset documentation.

This clarification and the legal context provided at the beginning of this chapter, demonstrates why a metadata model is needed to describe datasets in the context of EHDS secondary use. Given the vast diversity and volume of health data types involved in the EHDS, managing this data effectively requires a common metadata model. Without such a model, it would be impossible to ensure that metadata is structured, consistent, and interpretable across different institutions, data holders, and countries. This common model is essential not only for regulatory compliance but also for enabling data users to efficiently search, understand, and assess health data for secondary use.

A metadata model is a structured framework that defines the rules and parameters for describing data. Specifically, it specifies:

- Which metadata elements must be included (e.g., date of creation, format, coding system)
- What each metadata element means (ensuring clarity and consistency in interpretation)
- What values are allowed or expected (e.g., predefined formats or coding standards)

- How the metadata should be organized and formatted (e.g., consistent data fields and hierarchies)

In essence, a metadata model is like a blueprint or rulebook for describing data in a way that is consistent, clear, and understandable to all stakeholders. By defining the rules on how to describe health data, a metadata model ensures that datasets are correctly interpreted, shared, and reused across different platforms, institutions, and borders.

**Box 2. What is "machine-readability" and why do we need it in the EHDS context?**

Machine-readability means that information is structured in a way that computers can easily read, understand, and process it, without needing human interpretation. Instead of writing descriptions in free text (which people can understand but computers cannot), machine-readable content uses standardised terms and formats that follow strict rules. For example, instead of saying: *"This dataset was created by a hospital and last updated in March 2024."* the information would be broken down into clearly labelled elements, such as: *Publisher type: hospital, Last updated: 2024-03-01.*

This kind of structure allows software to automatically search, filter, compare, and link datasets, without manual effort.

**Why does this matter for the EHDS?**

In the EHDS, data will come from many different sources across Europe—hospitals, public health authorities, research projects, registries, and more. Machine-readability makes it possible to:

- **Search and discover datasets efficiently**: A researcher can filter thousands of datasets based on country, disease area, time period, or data type, automatically.
- **Enable automated data catalogues**: Health Data Access Bodies can build national catalogues where dataset descriptions are indexed, updated, and accessed through software systems—not just viewed as static pages.
- **Support interoperability**: Machine-readable metadata allows datasets to be integrated, compared, or reused across borders, systems, and tools.
- **Improve transparency and traceability**: Machine-readable records help track where data came from, how it was generated, and under which conditions it can be reused

There are already commonly used standards and formats specifically in the health domain to support structured, interoperable, and high-quality data exchange and management. Among the most relevant are:

- **HL7 FHIR (Fast Healthcare Interoperability Resources)**[7]: FHIR is a widely adopted -in metadata for each data "resource" (e.g., Patient, Observation, Medication) and is extensively standard used to define and exchange health data in a machine-readable format. It includes built used in clinical care settings, EHRs, and health information

[7] HL7 FHIR Foundation

exchange systems. FHIR enables seamless data exchange, enhancing interoperability between different healthcare systems.

- **ISO/IEC 11179**[8]: This standard focuses on the consistent definition, naming, and administration of data elements within a metadata registry. It is widely used in health information systems and data repositories to ensure semantic interoperability. ISO/IEC 11179 is particularly important in contexts where precise definitions of data elements (such as lab test codes or value lists) are necessary for clarity and consistent interpretation across different platforms and organizations.

- **CDISC (Clinical Data Interchange Standards Consortium)**[9]: CDISC provides a set of metadata standards for clinical trial datasets, such as the Define-XML standard, which describes the structure and content of clinical trial data. CDISC standards are essential in the clinical research and pharmaceutical sectors, especially when preparing datasets for regulatory submissions to bodies like the European Medicines Agency (EMA) and the U.S. Food and Drug Administration (FDA).

While these standards and formats are widely used within specific domains of health data, they do not provide a comprehensive, unified metadata framework for the broad and diverse data types covered by the EHDS secondary use. The EHDS must accommodate a much broader and more diverse range of health-related data, including not only clinical and research data but also non-clinical and aggregated sources. For example, socio-economic, environmental, and behavioural determinants of health, as well as wellness data from apps like fitness trackers or sleep monitors, are not captured by clinical standards, which are primarily focused on medical or research contexts. Likewise, aggregated data such as healthcare resource use, health expenditure, and population-level health registries require a metadata approach suited to statistical reporting and policy analysis, rather than individual-level clinical records. Existing models are not designed to describe these kinds of datasets in a consistent or interoperable way. Therefore, a unified and flexible metadata model is needed to ensure consistency, interoperability, and usability across all relevant data domains in the EHDS.

Another key reason for adopting a common metadata model in the EHDS is to ensure interoperability with other sectoral data spaces under the European Strategy for Data[10]. This strategy aims to create a unified data market across diverse domains such as health, agriculture, energy, and mobility[11]. To enable meaningful data exchange and cross-sector innovation, metadata models must be compatible across these fields. A shared, adaptable metadata framework helps ensure that health data can be integrated with other types of data, for example, environmental or socio-economic, supporting broader EU goals such as the European Green Deal, AI development, and public health research.

## 6.3 DCAT-AP as basis for common metadata model development for EHDS2

The information provided in Sections 6.1 and 6.2 present the foundation for the required widely adopted and interoperable metadata model, essential to ensure consistency and

[8] ISO/IEC 11179-1:2023 - Information technology — Metadata registries (MDR)
[9] CDISC | Clear Data. Clear Impact.
[10] A European strategy for data | Shaping Europe's digital future
[11] Common European Data Spaces | Shaping Europe's digital future

functionality across dataset catalogues both within the health domain and across other sectors. The Data Catalog Vocabulary (DCAT)[12] offers a solid foundation for this purpose.

DCAT, developed by the W3C, is an Resource Description Framework (RDF), vocabulary, designed to support the publication and discovery of dataset catalogues on the web. It enables structured and machine-readable description of datasets, facilitating semantic interoperability. By using a common vocabulary, DCAT facilitates the linking and integration of datasets across catalogues, improving discoverability and reuse.

DCAT includes key classes to support consistent metadata modelling:

1. **dcat:Catalog:** This class represents a collection of metadata entries, each describing a dataset, data service, or other resource. It functions as the organisational structure that brings together and presents these entries within a unified catalogue.

2. **dcat:Resource:** A general-purpose class used to refer to any item included in a catalogue. While not typically used on its own, it serves as the foundational class for more specific resource types, such as datasets and services, allowing for flexibility and extensibility.

3. **dcat:Dataset:** This class describes a coherent collection of data, whether numerical, textual, visual, or otherwise. It could correspond to a database, a file, a report, or another structured data asset intended for use or reuse.

4. **dcat:Distribution:** A distribution refers to a specific accessible form of a dataset. For example, if a dataset is available in both CSV and JSON formats, each of those would be a separate distribution. It provides the technical details and links needed to retrieve the data.

5. **dcat:DataService:** When data is accessed dynamically, such as through an API or web service, rather than as a static file, this class is used to describe the interface through which the data can be queried or retrieved.

6. **dcat:DatasetSeries:** This class is used for datasets that form part of a logical sequence, such as regular statistical releases or time-based collections. It enables grouping datasets with shared themes or structures along a recurring dimension (e.g. time, geography).

7. **dcat:CatalogRecord:** This class contains metadata about the metadata record itself. It tracks administrative information such as when the entry was created or updated, and by whom, supporting cataloguing transparency and lifecycle management.

[12] Data Catalog Vocabulary (DCAT) - Version 3

Together, these classes provide the framework for building semantically rich, machine-readable catalogues that support effective data discovery, access, and interoperability.

**Box 3. Fundamentals of RDF**

In the previous section, we explained why machine-readable metadata is essential to the EHDS for secondary use. But **how** do we make metadata machine-readable and interoperable? The answer lies in **RDF,** the **Resource Description Framework**. RDF is a standard model used to structure information in a way that both humans *and* machines can understand. It represents data as subject–predicate–object relationships, known as **triples**, which together form a **graph-based model** that enables data to be linked and queried efficiently across systems. For example, we can express that *Dataset A* (subject) *was generated by* (predicate) *Hospital X* (object). Each triple forms a link in the graph, allowing complex networks of information to be built and explored.

This structured model enables:

- **Machine-readability**: Computers can read, interpret, and act on metadata automatically.
- **Interconnectivity**: Different datasets, catalogues, or metadata vocabularies can be connected, reused, and linked—even across countries or sectors.
- **Standardisation**: RDF uses well-defined vocabularies, helping everyone "speak the same language" when describing data.

**A practical example in the EHDS context:** Imagine a researcher is looking for datasets on breast cancer treatments in women aged 40–60, from 2018–2023, collected in hospitals across three EU countries.

- If metadata is not machine-readable, they might have to open and read dozens of documents manually, trying to understand what each dataset contains—and still risk missing some useful ones.

- If metadata is structured using RDF, the researcher can search the dataset catalogue automatically, using keywords, timeframes, locations, or medical codes. The system can match descriptions precisely, filter results, and even link related datasets—saving hours or days of manual effort and reducing the chance of errors.

RDF makes this kind of intelligent, cross-border data discovery possible. This kind of structure allows software to automatically search, filter, compare, and link datasets, without manual effort.

The DCAT-AP[13] is a specialised extension of DCAT designed to meet the specific needs of European data spaces. Developed and maintained by the European Commission under the SEMIC initiative for Interoperable Europe[14], DCAT-AP ensures that datasets adhere to EU-specific metadata requirements, fostering cross-border data sharing and interoperability. DCAT-AP builds upon the DCAT standard by introducing additional classes, properties,

[13] DCAT-AP 3.0
[14] DCAT-AP | Interoperable Europe Portal

controlled vocabularies, and metadata elements. It relies on EU Vocabularies managed by the Publications Office of the European Union, ensuring consistency in metadata descriptions across data portals. Therefore, DCAT-AP addresses common challenges such as:

1. **Discoverability**: Helping data users find and understand datasets maintained by various public administrations, particularly when accessing data from other Member States where language barriers or unfamiliar governmental structures may exist.
2. **Reusability**: Allowing data providers to increase the visibility of their datasets by publishing metadata online, even when making the underlying data available might be costly or demand is uncertain.
3. **Domain flexibility:** DCAT-AP's flexible framework also supports the development of domain-specific extensions, such as HealthDCAT-AP, allowing sectors like health, agriculture, or energy to build on the common specification while addressing their unique requirements.

Importantly, DCAT-AP specifies mandatory, recommended, and optional metadata elements for each dataset. Preserving mandatory cardinalities defined by DCAT-AP is essential for ensuring backwards compatibility and alignment with the broader Interoperable Europe framework. Methods such as SHACL (Shapes Constraint Language)[15] can be used to validate that metadata records meet these structural and semantic requirements, supporting metadata reusability and interoperability. This is particularly important in the context of the EHDS, where there is a strong need for rich metadata to support the discoverability, understanding, and reuse of datasets for secondary purposes. In this context, *rich metadata* refers to descriptive information that goes beyond basic technical attributes and includes provenance (e.g., who generated the data and how), data quality indicators**,** licensing terms**,** temporal and spatial coverage, and links to related datasets or services. These elements are essential for enabling meaningful, trustworthy, and cross-border use of health data.

DCAT-AP is widely recognised by semantic experts as an effective technical solution for implementing EU data spaces. By ensuring that key metadata elements are formatted consistently, DCAT-AP allows different catalogues and platforms to "speak the same language", simplifying data sharing, integration, and reuse.

**Key Features of DCAT-AP**

- A common set of constraints to standardise metadata descriptions across portals.
- A unified approach to dataset metadata, ensuring consistency across EU public sector data catalogues.
- Support for domain-specific extensions, allowing additional metadata elements tailored to specific data spaces (e.g., GeoDCAT-AP for geospatial data, HealthDCAT-AP for health data).
- Interoperability across data spaces, ensuring data portals across different domains can effectively exchange metadata.

Further information on the official DCAT-AP standard can be found in the latest version available at: https://semiceu.github.io/DCAT-AP/releases/3.0.0/. To support a comprehensive understanding of DCAT-AP, the SEMIC e-learning course "DCAT-AP for

[15] Shapes Constraint Language (SHACL)

Data Spaces" provides structured guidance on its application. Additional resources of interest include DCAT-AP Explained, which offers practical insights into its application.

## 6.4 Tailoring DCAT-AP to the Health Domain: Development of Health-DCAT-AP

HealthDCAT-AP is an extension of the DCAT-AP specification, developed to meet the specific metadata needs of the electronic health data ecosystem within the EHDS. While DCAT-AP provides a minimal, domain-agnostic framework for dataset description and interoperability across European data portals, HealthDCAT-AP introduces additional properties, controlled vocabularies, and recommendations tailored to the unique characteristics of health data. Since it is an extension, for applications to be compliant with HealthDCAT-AP, they must first conform to DCAT-AP. This requirement ensures semantic and structural interoperability across implementations and is essential to maintain alignment with the Interoperable Europe framework.

**What does it mean in practice?**
For DCAT-AP the three main classes that are mandatory when defining a dataset are: Catalogue, Dataset and Agent. Each of them has a set of mandatory properties that ensure that datasets have the minimum necessary information to be useful and understandable for users, thereby facilitating interoperability and data sharing.

• **Catalogue:** must have title, description, publisher and data.
• **Dataset:** must have title and description.
• **Agent:** must have the property name.

These are very relevant as DCAT-AP establishes fundamental rules regarding mandatory cardinalities that maintain consistency across implementations and extensions. When developing extensions to DCAT-AP:
- **Mandatory properties in DCAT-AP must remain mandatory in any extension**.

This ensures that extensions only enhance, rather than diminish metadata requirements, preserving interoperability while allowing for domain-specific needs, such as HealthDCAT-AP or the previously developed extensions for geospatial datasets (GeoDCAT-AP) or statistical data sets (StatDCAT-AP).

The HealthDCAT-AP 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.

**Key Features of HealthDCAT-AP (FAIR-aligned)**

- **Findability -** HealthDCAT-AP defines standardised and structured metadata elements tailored to health data, ensuring datasets are consistently and clearly described across catalogues. This enhances their discoverability across national metadata catalogues, HDABs, and the EU's central dataset catalogue. By supporting uniform metadata

descriptions, HealthDCAT-AP improves searchability and enables users to easily locate relevant datasets within the EHDS.

- **Accessibility -** The framework promotes transparent access conditions by clearly specifying how datasets can be accessed, including any legal, technical, or procedural constraints. Metadata is made available in a machine-readable format using RDF, enabling programmatic access and automated data retrieval in line with FAIR Accessibility principles.

- **Interoperability -** HealthDCAT-AP enhances semantic and technical interoperability by building upon DCAT-AP and leveraging widely adopted web standards such as RDF. It enables seamless integration and exchange of metadata between national catalogues, HDABs, and the EU-level infrastructure, allowing different systems and stakeholders to interpret and use the data consistently across borders and domains.

- **Reusability -** By including rich metadata, clear provenance information, and well-defined conditions for reuse, HealthDCAT-AP ensures that health datasets can be repurposed effectively. The extension introduces health-specific properties and controlled vocabularies aligned with EU regulations (e.g., GDPR, EHDS Regulation), thereby supporting legal compliance and long-term data usability. This structure is particularly valuable for enabling secondary use, including research, innovation, and integration into AI-driven solutions.

**Box 4. Development of HealthDCAT-AP extension**

The **HealthDCAT-AP extension** started to be developed in **January 2023** as part of the *HealthData@EU Pilot* project (Link), with the goal of improving interoperability of health data across Europe. Guided by a Technical Working Group and based on an analysis of existing metadata catalogues and DCAT profiles, the extension was aligned with FAIR principles and EU policy priorities. A draft version was published on GitHub for public feedback. As the profile was only piloted by a few partners, the **TEHDAS2** project now supports further refinement and validation with broader EHDS stakeholders.

Regarding HealthDCAT-AP extension, detailed documentation is available here. Additionally, *Deliverable 6.2 – HealthDCAT-AP: A DCAT Application Profile for the Description of Health Datasets – Recommendations on Further Development and Deployment for Possible EU-Wide Uptake* (Link) can be consulted for detailed insights into the design and implementation considerations of the HealthDCAT-AP metadata model.

The HealthDCAT-AP Specification is published using W3C ReSpec, ensuring a structured, machine-readable, and web-friendly format. This specification details how HealthDCAT-AP extends DCAT-AP to support metadata interoperability within the EHDS. It provides clear guidance on metadata elements, relationships, and best practices for dataset cataloguing. The specification follows an open, community-driven development process and is available for feedback and continuous refinement. Additional resources tailored to the EHDS are available on Sciensano's Health Information Portal, including the HealthDCAT-AP literacy

platform. These resources are continuously updated to support data holders in meeting EHDS metadata requirements and improving dataset discoverability and interoperability.

### 6.5 What HealthDCAT-AP Is Not

As HealthDCAT-AP is a newly developed metadata standard, particularly for many data holders who are the primary audience of this guideline, it is important to clarify what this extension is *not* intended to be. Understanding the boundaries and purpose of HealthDCAT-AP helps prevent confusion and ensures correct implementation within the broader context of data governance and interoperability under the EHDS.

- **Not a data standard or terminology:** HealthDCAT-AP is not a data standard itself. Rather, it provides a framework for *describing* datasets and their metadata. It does not replace existing medical terminologies such as SNOMED CT or ICD-10, or health data standards such as FHIR, CDISC or OMOP CDM. These standards and terminologies remain essential for annotating, standardising and structuring the data itself and should be referenced within the metadata. HealthDCAT-AP ensures that such datasets are discoverable and described in a harmonised manner, but it does not define how the data should be encoded or structured.

- **Not a data management or storage solution:** HealthDCAT-AP does not serve as a data management or data governance tool. Its purpose is to support the cataloguing of health datasets for improved discoverability and interoperability. It does not cover the operational aspects of dataset lifecycle management. Data holders remain responsible for managing their data infrastructure, complying with EHDS governance requirements, and meeting data protection and legal obligations, such as those under the GDPR.

- **Not a substitute for other metadata models:** HealthDCAT-AP is not intended to replace existing domain-specific metadata models or other data governance practices used in the health sector. Instead, it acts as a common framework - a *metadata lingua franca* - that enables interoperability across diverse systems and metadata standards. HealthDCAT-AP is designed to complement and coexist with standards such as HL7 FHIR, CDISC, and ISO/IEC 11179, supporting cross-system data exchange while enabling a unified approach to metadata description across the EHDS.

## 7 HealthDCAT-AP: Explanation for data holders

Building on the common metadata model introduced in the previous chapter, this chapter provides a detailed explanation of individual properties of HealthDCAT-AP. It offers practical guidance on how to use each property to ensure compliance with the metadata requirements for secondary use of health data under the EHDS Regulation.

### 7.1 What information should be provided in the metadata? Legal context

Under the EHDS Regulation, health data holders making datasets available for secondary use are required to describe these datasets using structured metadata. These obligations stem from the EHDS Regulation and are further supported by horizontal legislation such as the Data Act (Regulation (EU) 2023/2854), which sets essential metadata and interoperability

requirements across European data spaces. Under Article 57 of the EHDS Regulation, each HDAB must establish and maintain a national dataset catalogue that includes "d*etails about the source and nature of electronic health data, in accordance with Articles 77, 78 and 80, and the conditions for making electronic health data available*."[16]

Article 77(2) further specifies that:

"*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*."

These provisions ensure that data users can understand what the dataset contains, how it was generated, and under what terms it may be accessed. While the Data Act does not introduce metadata obligations specific to health, it reinforces the importance of structured metadata across data spaces by setting interoperability requirements (Article 33), which are complementary to those in the EHDS Regulation.

According to Article 33(1) of the Data Act: "Participants in data spaces offering data or data services to other participants shall comply with the following essential requirements to facilitate interoperability: (a) The content of the dataset, usage restrictions, licences, data collection methodology, data quality, and data uncertainty shall be sufficiently described, where applicable, in a machine-readable format to allow the recipient to find, access, and use the data; (b) Data structures, data formats, vocabularies, classification systems, taxonomies, and code lists, when available, shall be described in a consistent and publicly accessible manner."

Therefore, by analysing this legal framework, we can understand what are the mandatory elements that must be included in the metadata. Based on the EHDS Regulation (Articles 57 and 77) and the Data Act (Article 33), metadata describing electronic health datasets should include the following key topics:

1. **Source of the dataset:** Where and how the data was generated or collected.
2. **Scope of the dataset:** What population, timeframe, or health domains the data covers.
3. **Main characteristics and nature:** Key features, types of data included (e.g. clinical, administrative), and data modality.
4. **Conditions for data access and use*:** Legal, ethical, or organisational conditions, including licensing terms and usage restrictions.
5. **Data collection methodology:** How the data was collected or derived, including instruments or systems used.
6. **Data quality:** Information about completeness, accuracy, timeliness, and limitations of the data.
7. **Licences and usage terms*:** Applicable data licences and terms of reuse.
8. **Data structures and formats*[17]:** File formats, schemas, or APIs available for accessing the data.
9. **Vocabularies and classifications:** Controlled vocabularies, taxonomies, and coding systems used (e.g. ICD-10, SNOMED CT).

[16] EHDS Regulation, Article 57: *"(1) Health data access bodies shall carry out the following tasks: (j) making public, through electronic means: (i) a national dataset catalogue that includes details about the source and nature of electronic health data, in accordance with Articles 77, 78 and 80, and the conditions for making electronic health data available;"*

[17] * Information found in Class Dataset Distribution

10. **Machine-readability:** Metadata must be structured to be programmatically accessible (e.g. using RDF).

## 7.2 HealthDCAT-AP Categorising Data by Access Level

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) → **PUBLIC**

2. Non-personal electronic health data available as non-open data [protected or restricted data] → **RESTRICTED**

3. Personal electronic health data [sensitive data] → **NON-PUBLIC**

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[18] maintained by the Publications Office. Therefore, the following rationale applies:

If a dataset **does not contain any 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 **does not contain any 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[19].

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.

HealthDCAT-AP adapts the DCAT-AP metadata elements into groups with headings mandatory, recommended, optional and associated cardinalities to reflect the sensitivity level of health data, aligning with the requirements 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 in the following section for each group of properties.

[18] http://publications.europa.eu/resource/dataset/access-right
[19] Regulation - 2022/868 - EN - EUR-Lex

**Box 5. What are cardinalities in DCAT-AP?**

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:

- **Mandatory properties with cardinalities 1..1 or 1..*:** These properties **MUST** be included in the metadata by the sender, and the receiver **MUST** be able to process them. 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.

- **Recommended properties with cardinalities 0..1 or 0..*:** These properties **SHOULD** be provided by the sender **if the information is available**, and the receiver **MUST** be able to process them. Including these properties is strongly encouraged if available as they provide valuable additional information.The cardinality 0..1 allows for zero or one occurrence, while 0..* permits zero to many occurrences.

- **Optional properties with cardinalities 0..1 or 0..*:** These properties **MAY** be provided by the sender, and the receiver **MUST** be able to process them. Their use is fully discretionary and depends on the relevance or availability of the information. 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.

## 7.3 HealthDCAT-AP properties explained

As previously introduced, this section focuses on the dcat:Dataset class, detailing how its properties should be used to describe datasets within the EHDS framework for secondary use. This section groups the HealthDCAT-AP properties according to the type of information they describe. This is supported by comprehensive usage notes, practical instructions, and templates to guide data holders in completing the metadata fields across different scenarios. Finally, this section provides a detailed explanation of the cardinalities related to the various access restriction types that may apply to datasets in the EHDS context for secondary use.

### 7.3.1 Information on the dataset history and content

These properties describe what the dataset is about, how it was created, and what it contains. They help the data user understand the scope, origin, structure, population, time frame, and legal context of the dataset.

**Table 19. Properties of the Dataset class related to history and content.** Cardinalities are shown for public, restricted, and non-public (sensitive) datasets. M – Mandatory, R – Recommended, O – Optional.

| Property | Cardinalities | | | Question to answer |
|---|---|---|---|---|
| | Public | Restricted | Non-public | |
| Title | **M** | **M** | **M** | *What is the name of this dataset?* |
| Description | **M** | **M** | **M** | *What information does this dataset contain?* |
| Purpose | **O** | **O** | **M** | *For what intended objective was this dataset created?* |
| Provenance | **O** | **O** | **M** | *What activities or processes led to the creation of this dataset?* |
| Creator | **O** | **O** | **O** | *Who originally created this dataset?* |
| Generated by | **O** | **O** | **O** | *What processes or activities created this dataset?* |
| Keywords | **R** | **R** | **M** | *What keywords best describe the content of this dataset?* |
| Has code values | **O** | **O** | **R** | *Which standard health codes best describe the main topics of this dataset?* |
| Language | **O** | **O** | **R** | *Which language(s) is used in the dataset?* |
| Geographical coverage | **O** | **O** | **M** | *Which country or region does the dataset refer to?* |
| Population coverage | **O** | **O** | **R** | *Which population group is represented in this dataset?* |
| Number Unique individuals | **O** | **O** | **R** | *How many distinct individuals are represented in this dataset?* |
| Number Records | **O** | **O** | **R** | *How many total records or entries are in this dataset?* |
| Max and min typical age | **O** | **O** | **R** | *What is the typical age range of individuals in the dataset?* |
| Has personal data | **O** | **O** | **R** | *Does this dataset contain personal data? If so, which type of personal data?* |
| Temporal | **O** | **O** | **R** | *Which time period does this dataset cover?* |
| Temporal resolution | **O** | **O** | **R** | *What is the frequency or granularity of the data in time?* |
| Spatial resolution | **O** | **O** | **O** | *What is the geographic precision of the data?* |
| Accrual periodicity | **O** | **O** | **R** | *How often is the dataset updated or collected?* |
| Analytics | **O** | **O** | **R** | *Are there any summary statistics, or dashboards available for this dataset?* |
| Legal Basis | **O** | **O** | **R** | *What legal basis allows the use of this dataset for secondary purposes?* |
| Qualified attribution | **O** | **O** | **O** | *Who contributed to the dataset, and what was their role?* |
| Retention period | **O** | **O** | **O** | *How long will this dataset be retained?* |
| Identifier | **M** | **M** | **M** | *What is the unique ID assigned to this dataset?* |

#### 7.2.1.1 Title

Write a short, descriptive sentence summarising the dataset's content. The title should be understandable to someone unfamiliar with the dataset and help them identify its relevance (e.g., *"Hospital Discharge Records – Region X, 2015–2022"*).

- ❑ Avoid internal codes or overly technical terms.
- ❑ Use consistent naming across datasets from the same source.
- ❑ You can repeat the title in multiple languages using language tags (e.g., `@en`, `@de`)

#### 7.2.1.2 Description

Use this field to provide a clear, detailed summary of what the dataset contains and its key characteristics. Include what types of data are present (e.g. patient records, prescriptions, hospital visits), the reason for data collection (e.g. billing, clinical care, research), and any relevant context that helps the reader understand its scope and purpose.

- ❑ Use plain language, avoid technical jargons.
- ❑ Focus on what, why, and how the data was collected.
- ❑ Consider the most important information someone would need to understand the dataset you are describing.

#### 7.2.1.3 Purpose

Use this field to explain why the dataset was originally collected. This helps data applicants understand the context, original intent, and potential relevance of the data for secondary use.

- ❑ Be clear and concise - this is about the *original* purpose of data collection, not how it might be reused.
- ❑ Include the setting or system where the data was collected (e.g. EHR system, insurance registry, clinical trial).
- ❑ Mention the main use case (e.g. patient care, reimbursement, disease surveillance).
- ❑ Use simple language. For example: "This data was collected through the hospital's electronic health record system for routine patient care."

#### 7.2.1.4 Provenance

Provide a clear, concise statement describing the dataset's origin, creation process, and any significant modifications or transformations. Include information about the entities involved in its production, modification, and maintenance, as well as key milestones in the dataset's history that may impact its interpretation or application. This should also cover the standards followed and any relevant links to metadata catalogues. Include any relevant information about its creation, such as its original data source (e.g., registry, survey, electronic health records). Also, note any transformations or processing that occurred (e.g., data cleaning, harmonization), the entities involved (e.g., data creators, modifiers), and any steps in the data processing chain, including the calculation of new variables. If any of this information is unavailable or not tracked, state so and explain why.

This property helps future users understand how the data was generated, its quality, and the context behind its creation. The goal is to provide all necessary details so that users can interpret the dataset correctly and use it appropriately.

- Focus on the key information: Include details that will help users understand the dataset's origin, transformations, and processing steps. Mention the original data source (e.g., registry, survey, EHR), any transformations (e.g., cleaning, harmonization), and the entities involved (e.g., creators, modifiers).
- Select the most useful information - think about the details that will help others understand your dataset. Include the amount of detail that helps a user understand how the dataset was created and how it might affect the way the data is interpreted. Avoid going into unnecessary details that aren't important for understanding the data itself.
- If certain provenance details are unavailable (e.g., creator or modification history), explain why or what is known. If this information is not available, acknowledge it and provide as much context as possible with the information you do have available.

#### 7.2.1.5 Creator

This property indicates who produced the dataset, typically the person, group, or organization primarily responsible for the creation and initial collection of the data. When filling out this property, focus on the entity responsible for data collection, especially in cases where there are separate entities involved in the processing or use of the data. It's important to clearly identify the original creator, as this is essential for understanding the provenance and integrity of the dataset.

- If the dataset was collected by one organization but processed or used by another (e.g., a hospital providing data to a research institute), indicate the original data collector.
- Avoid listing entities that are not directly responsible for the data's creation, even if they played a role in processing or using the data later.

#### 7.2.1.6 Generated by

This property describes the activity or process behind the creation of the dataset. Focus on the activity that generated the data, not the technology or tools used. This activity could be a specific project, mission, ongoing survey, or any other relevant process (e.g., routine hospital data collection, research study, etc.).

- Be clear about the type of activity that led to the dataset's creation. For example, if the dataset was generated as part of a hospital's routine data collection, describe it as such. If it resulted from a specific research project, mention that project.
- Avoid focusing on the technologies or tools used (e.g., sequencing), as these can be applied across various activities. Instead, emphasize the context of the data collection activity itself.
- If multiple activities contributed to the dataset, use multiple `prov:wasGeneratedBy` properties to describe these contexts at different levels of granularity.

The technical working group recommend using a controlled vocabulary for this property, which represents a preliminary list of activities that could have generated the data types described in Article 51 of the EHDS Regulation. It is intended as a starting point and will require careful validation and refinement in future work. More information in Section 8.5 .

#### 7.2.1.7 Keywords

In this property, list short, specific words or phrases that describe the content of your dataset. These keywords help others find your dataset when they search for related topics. Choose terms that accurately reflect what the dataset is about.

- If you're familiar with research practices, this will be similar to picking keywords for a scientific paper. If not, start by looking at your dataset title and ask yourself: *What words would someone type into a search box to find this dataset?* Use those as your keywords.

#### 7.2.1.8 Has Code Values

This property is used to semantically annotate your dataset by associating it with relevant codes from established health classification systems, such as ICD-10 or SNOMED CT. These can be codes actually used within the dataset, but also codes that are not used directly but are strongly related to the dataset's content. By linking your dataset to meaningful standardised codes, you improve its discoverability, particularly in catalogues that support search by standard classifications. For example, if a user searches for a specific ICD-10 code and that code reflects your dataset's topic, your dataset will appear in the results thanks to this annotation.

- To fill this property, start by reviewing the **keywords** and **health themes** you have already used to describe your dataset. Then, identify relevant standard codes (e.g., from ICD-10, SNOMED CT) that correspond to those topics.
- Use **Wikidata URIs** where possible to enable machine-readable, interoperable annotations. *Example:* For a dataset about breast cancer, you could use the Wikidata entry for "Malignant neoplasm of breast (ICD-10 C50)" https://www.wikidata.org/wiki/Q128581
- Focus on the **most representative codes** that reflect your dataset's subject matter. You don't need to list all possible codes — just those that best describe its scope.

This property complements your descriptive metadata (like keywords and health themes), offering a **semantic layer** that supports automated search and reasoning in data catalogues.

#### 7.2.1.9 Language

- In this property, indicate the language or languages used in the dataset. This helps users understand what to expect when accessing your data. Use this field to reflect the actual language of the data content – for example, the language used in variable names, values, labels, or free-text fields. This helps users assess whether they can work with the dataset as it is, or if translation or interpretation may be needed.

#### 7.2.1.10 Geographical coverage

In this property, indicate the geographical area represented by the dataset**.** This helps users understand which region or population the data refers to and is especially useful when comparing or combining datasets. Using a controlled vocabulary in this property is

recommended as it supports consistency and improves searchability and interoperability across datasets.

- Be as precise as possible and use standardised terms, geographic identifiers from controlled vocabularies, or geographic coordinates. Use the EU Vocabularies Named Authority Lists for continents, countries, and places that are included in those lists. For example, for Belgium: http://publications.europa.eu/resource/authority/country/BEL

- If a specific location is not available in the Named Authority Lists, use GeoNames URIs to ensure accurate and consistent referencing.

#### 7.2.1.11 Population Coverage

Use this property to describe the population represented in the dataset. This is especially relevant for datasets derived from or focused on specific population groups, such as patients, age cohorts, or residents of a particular region. The Population Coverage should reflect who the data is about, based on the people from the geographic area specified in the Geographical Coverage property, from whom the data was collected and the dataset constituted.

- While Geographical Coverage indicates where the data originates (i.e., the area covered), Population Coverage explains who within that area the data represents. For example, if Geographical Coverage is Belgium, Population Coverage could be "residents aged 65 and over," "all hospitalised patients," or "children under 5."

#### 7.2.1.12 Number of Unique individuals

Use this property to indicate the number of unique individuals represented in the dataset. This provides users with a sense of the dataset's scale and its potential value for secondary use. This number should reflect the total (or approximate) count of distinct individuals whose data is included, helping potential data users quickly understand whether the dataset meets the scale requirements for their analysis or research purposes.

- If the exact number is not available, providing a well-founded estimate is appropriate and still valuable. Consider what would be most informative to a potential data user, for example: “Approximately 10,000 unique patients over 3 years.”

- This property is distinct from the number of records or data points - it focuses on individuals, regardless of how many times each person appears in the data.

#### 7.2.1.13 Number of Records

Use this property to indicate the total number of records in the dataset. This helps users understand the volume of data available and assess its suitability for specific types of analysis or research. Provide the total count of records, or an approximate average if the exact number is not available. This value reflects the total entries in the dataset, such as consultations, prescriptions, lab results, hospital admissions, etc. depending on the dataset’s structure and content.

- This property complements the Number of Unique Individuals property. While Number of Unique Individuals refers to how many distinct persons are represented in the dataset, Number of Records refers to how many total data entries are available. For example, 10,000 individuals may each have multiple medical records, resulting in 100,000 total records.

#### 7.2.1.14 Maximum and minimum typical age

Use the Maximum and Minimum Typical Age properties to define the age range of the population represented in the dataset. These properties help users understand the typical age coverage of individuals in health-related datasets, offering valuable insights for research, analysis, and population studies.

Minimum Typical Age: Specify the minimum typical age of the population represented in the dataset. This could reflect the age group typically included in the dataset or the lower bound of the age range for the individuals represented.

Maximum Typical Age: Specify the maximum typical age of the population represented in the dataset. This provides the upper bound of the age range for individuals within the dataset.

- If your dataset covers a broad range of ages, consider using these properties to narrow down the typical age ranges. For example, if your dataset primarily focuses on adults aged 18–65, this should be reflected in both the **Minimum** and **Maximum** age properties.

#### 7.2.1.15 Has personal data

Use this property to identify and describe the types of personal data included in the dataset. This helps data users understand the sensitivity of the dataset and supports more accurate filtering during dataset discovery. Indicate whether the dataset includes variables that contain personal information. This may refer to the presence of direct identifiers (e.g., names, contact information) or indirect identifiers and sensitive attributes (e.g., age, gender, health conditions, or education level), without specifying the exact variables. To ensure consistency with legal requirements and semantic clarity, refer to the categories of personal data defined in the Personal Data Categories (PD) vocabulary[20], which aligns with the GDPR.

- Carefully review your dataset's variables and list those that constitute personal data. Examples include gender, age, location, nationality, education, health record data, and contact information. Providing this information is essential for transparency, risk assessment, and compliance with legal and ethical standards, especially when handling sensitive health data.

#### 7.2.1.16 Temporal

The Temporal property defines the time period covered by the dataset. This is an essential property for datasets where the timing of data collection or observation is critical. Specifying the temporal period enables users to assess the dataset's relevance for time-sensitive

[20] Personal Data Categories (PD)

research, historical analysis, or trend tracking. This property is crucial for ensuring that users understand the temporal context of the dataset, helping them evaluate whether it meets their time-based research or analysis needs.

- **Start Date**: Specify the **start date** of the time period the dataset covers. This should reflect the beginning of the data collection or the relevant timeframe for the dataset.
- **End Date**: Specify the **end date** of the time period the dataset covers. This marks the conclusion of the data collection or the relevant time span for the dataset.

❑ If your dataset spans multiple years or is regularly updated, indicate the exact time range (e.g., from January 1, 2020, to December 31, 2020). If the dataset includes ongoing data collection, you can specify the end date as "ongoing" or "present" if applicable.

#### 7.2.1.17 Temporal resolution

The Temporal Resolution property defines the finest time interval resolvable in the dataset. This is a critical property for datasets that involve time-based data, as it helps users understand how frequently data points were recorded or observed. By specifying the temporal resolution, you provide clarity on the granularity of the dataset's time-based information.

- Definition: Specify the minimum time period resolvable in the dataset. This refers to the smallest time interval between consecutive data points or observations. Examples include hourly, daily, or monthly intervals.
- Standard Format: To express the temporal resolution, you can use the ISO 8601 [21]duration format. For instance:
  - "P1D" refers to a one-day interval (daily data).
  - "PT1H" refers to a one-hour interval (hourly data).
  - "P1M" refers to a one-month interval (monthly data).

❑ If your dataset contains events or measurements taken at different intervals (e.g., daily and monthly), provide the resolution for the most frequent interval. If the resolution varies within the dataset, consider providing a range (e.g., daily to monthly).

#### 7.2.1.18 Spatial resolution

The Spatial Resolution property defines the minimum spatial separation resolvable in the dataset. This is essential for geospatial datasets, as it indicates the finest level of spatial detail or precision represented in the data. You should specify the smallest spatial separation or distance that can be reliably distinguished in the dataset. This is typically measured in meters and represents the level of precision for geographic coordinates or other spatial data.

❑ If your dataset represents geographic features, and can reliably distinguish details down to 10 meters, the spatial resolution would be "10 meters."
❑ If your dataset includes multiple layers or types of spatial data with varying resolutions (e.g., satellite images at different scales), specify the resolution for the finest detail or, if appropriate, provide a range of resolutions.

[21] ISO - ISO 8601 — Date and time format

#### 7.2.1.19 Accrual periodicity

The Accrual Periodicity property defines the frequency at which the dataset is updated. This property helps users understand how often new data is added in the dataset. You should specify how often the dataset is updated, using standard terms from the EU Vocabularies Frequency Named Authority List[22], such as:

**DAILY** (updated every day)
**WEEKLY** (updated every week)
**MONTHLY** (updated every month)
**ANNUALLY** (updated once per year)

- The use of a controlled vocabulary is important because using inconsistent descriptions like "updated every day," "day to day," or "on a daily basis" creates interoperability issues, as these variations may be interpreted differently across systems. By using a controlled vocabulary, the data is standardized, improving discoverability and ensuring consistent interpretation.

- While Temporal Resolution property, previously described, focuses on the finest time interval that can be distinguished within the dataset (i.e., the smallest unit of time for which data is available), Accrual Periodicity refers to how often the dataset itself is updated (i.e., how often new data is added or existing data is updated). These properties complement each other, helping users understand both how often the dataset is updated and the level of time granularity within the data.

#### 7.2.1.20 Analytics

This property allows you to link your dataset to useful resources that help others better understand the dataset, without giving them direct access to the data.

**What type of resources can you include?**
You can include links to:

- **Dashboards** – interactive tools that show charts, maps, or summaries based on your dataset.
- **Technical reports** – documents that describe your dataset, such as data quality, completeness, or usability.
- **APIs** – services like the Beacon API that allow users to ask questions about your dataset (e.g., "do you have any data on condition X?") without revealing individual-level information.

- These resources help users explore and understand your dataset without seeing the actual data. They show information *about* the data, not the data itself.
- Example: The European Centre for Disease Prevention and Control (ECDC) provides a dashboard with interactive maps and charts on infectious diseases reported by EU/EEA countries: Surveillance Atlas of Infectious Diseases

- This property is optional. Only use it if you already have, or plan to have, these kinds of tools linked to your dataset.

[22] http://publications.europa.eu/resource/authority/frequency

#### 7.2.1.21 Legal Basis

This property allows you to specify the legal or regulatory foundation under which the data was originally collected and processed. It helps users understand that the dataset was collected lawfully and transparently, especially important when it includes personal data. Mention the relevant law, regulation, or legal provision that authorised the collection and processing of the data. For example, you can refer to national legislation, the General Data Protection Regulation[23] (GDPR), or other sector-specific legal frameworks.

- This refers to the legal basis at the time of initial collection, not the legal basis for secondary use under the EHDS Regulation.
- If your dataset contains personal data, referencing the correct GDPR Article (e.g. Article 6(1)(e) for public interest) is highly relevant.

#### 7.2.1.22 Retention Period

This property allows you to specify how long the dataset will be available for secondary use. It defines the time frame during which users can request access to the dataset. After this period, the dataset may no longer be reusable, but the metadata must remain available. Indicate the start and end dates during which the dataset is available for secondary use. If no end date is set, leave this property blank, it's only mandatory if a retention limit applies. Even if the dataset is removed, the metadata (including this property) must remain published for transparency and traceability.

- This property is important if your dataset is only available for reuse for a limited period (e.g. due to data protection, contractual, or project constraints).

#### 7.2.1.23 Qualified attribution

This property allows you to name individuals or organisations who played a significant role in the creation or maintenance of the dataset. Use it when someone other than the main publisher or creator should be credited for their contribution. List the names of contributors and specify their role. Roles can include author, co-author, editor, contributor, or stakeholder. Only include information that reflects a significant contribution and brings added value to someone consulting the metadata.

- This property is especially useful for datasets developed collaboratively or involving domain experts. Common roles include:
  - **Author** – main creator of the dataset
  - **Co-author** – contributed equally to dataset creation
  - **Editor** – revised or curated the dataset
  - **Contributor** – added meaningful content or expertise
  - **Stakeholder** – provided institutional or strategic support
- Keep the description focused and relevant, including only what adds meaningful context for the user.

[23] Regulation - 2016/679 - EN - gdpr - EUR-Lex

#### 7.2.1.24 Identifier

This property allows you to assign a unique and persistent identifier to your dataset. Use it to ensure that your dataset can always be reliably found and referenced, even if it moves across platforms. Provide a Persistent Uniform Resource Identifier (PURI) that links directly to the metadata or dataset description in the primary catalogue. If your dataset is already described in your organisation's internal catalogue or FAIR data point, you can use the PURI from that system.

- A PURI is a stable and permanent web link that will not change over time, helping users and systems reliably retrieve your metadata about this dataset, ensuring that your dataset description remains accessible, citable, and interoperable across health data catalogues.

### 7.3.2 Contact information

These properties provide contact details for the people or organisations responsible for the dataset. This helps users know who to reach out to for more information or support.

**Table 20. Properties of the Dataset class related to contact information.** Cardinalities are shown for public, restricted, and non-public (sensitive) datasets. M – Mandatory, R – Recommended, O – Optional.

| Property | Cardinalities | | | Question to answer |
|---|---|---|---|---|
| | Public | Restricted | Non-public | |
| Contact point | **R** | **R** | **M** | *Who can be contact for more information about this dataset?* |
| Publisher | **O** | **M** | **M** | *Who is responsible for making this dataset available?* |
| HDAB | **M** | **M** | **M** | *Which HDAB oversees access to this dataset?* |

#### 7.2.2.1 Contact point

This property allows you to provide up-to-date contact information for the person or organisation responsible for the dataset. It helps users know who to reach out to with questions, clarifications, or access requests, serving as the main "helpdesk" for the dataset. Provide the name, organisation, and a reliable contact method that you wish to associate with the dataset.

- The contact point should be someone who can respond to queries about the dataset.
- It may be the same as the publisher, but this is not always the case. Some examples:
  - For large institutions, the publisher might be the organisation (e.g. a university hospital), while the contact point should be an individual or unit able to handle specific dataset-related questions.
  - For research datasets, where project staff may no longer be available after the project ends, you can use a general contact at the organisation level (e.g., university or research office email) to ensure continuity.
- The contact point can be a person or an organisation, depending on your context. What matters is that it remains functional and responsive over time.

#### 7.2.2.2 Publisher

This property allows you to specify the organisation responsible for making the dataset available. The publisher is typically the data holder - the entity accountable for ensuring the dataset can be accessed and used. Provide the name of the organisation that manages and maintains access to the dataset. This property is mandatory under HealthDCAT-AP and also required by the Data Governance Act as it ensures clarity about who is ultimately responsible for the dataset's availability.

- ❑ In small organisations, the publisher and contact point may be the same. In large institutions, they are often different roles.

- ❑ Do not confuse the publisher with other roles. Here's how to distinguish them:
  - **Creator** - Who generated the data?
  - **Publisher** - Who is responsible for making the data available upon receiving a request?
  - **Contact point** - Who should be contacted to answer any question about the dataset?

#### 7.2.2.3 HDAB

This property allows you to specify which HDAB is responsible for managing access to the dataset. Under the EHDS Regulation, HDABs act as national authorities that oversee access to health data for secondary use, ensuring compliance with legal and ethical requirements. Provide the name of the HDAB that is responsible for this dataset in your Member State. To ensure consistency across metadata records, a common EU-wide register of HDABs should be established and used for this property in the future.

- ❑ If your Member State has only one HDAB, that's the one to list here.
- ❑ If there are multiple HDABs, check with your national authority or follow national guidance to determine which HDAB is responsible for your dataset.
- ❑ This property helps ensure that data users know who oversees access requests and whom to contact for permits or procedures.

### 7.3.3 Information on Data Standards in use

These properties describe the standards, classifications, and terminologies used in the dataset. They help users assess how the data is structured and whether it is compatible with other datasets.

**Table 21. Properties of the Dataset class related to data standards in use.** Cardinalities are shown for public, restricted, and non-public (sensitive) datasets. M – Mandatory, R – Recommended, O – Optional.

| Property | Cardinalities | | | Question to answer |
|---|---|---|---|---|
| | Public | Restricted | Non-public | |
| Conforms to | O | O | R | *Which data model or specification does this dataset follow?* |
| Has coding system | O | O | R | *Which coding systems or terminologies are used in this dataset?* |

The health data landscape relies heavily on data standards to ensure semantic and technical interoperability. Standards are essential for enabling data integration, consistent interpretation, and efficient reuse across systems and countries. Documenting the standards your dataset follows is crucial for users to assess data quality, reusability, and compatibility with their own systems. It also supports automated discovery and filtering of datasets based on specific standards. There are three properties in HealthDCAT-AP that allow you to describe the data standards used in your dataset.

#### 7.2.3.1 Conforms to

This property allows you to specify which standards, schemas, or regulations your dataset conforms to. It ensures clarity about how the data is structured or interpreted, which supports interoperability and reuse. List any data standards used in your dataset (e.g. HL7 FHIR, OMOP CDM, CDISC).

- To maximise interoperability, these standards should be identified using URIs from controlled vocabularies, such as Wikidata. For example, instead of writing "FHIR", use the Wikidata URI: https://www.wikidata.org/wiki/Q19597236
- Using free text (e.g. just writing “FHIR”) should be avoided, as it limits the ability to filter or retrieve datasets based on shared standards.

Chapter 7 will show concrete examples of how using Wikidata links improves search and interoperability capabilities across catalogues.

#### 7.2.3.2 Has Coding System

This property allows you to specify the coding systems used within your dataset. 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. Specify the standardised coding systems used in the dataset (e.g. ICD-10-CM, SNOMED CT, DRGs).

Your dataset uses ICD-10 to classify diseases? Indicate ICD-10 here.
Your dataset uses SNOMED CT to encode clinical terms? Indicate SNOMED CT here.

- Use Wikidata concept URIs to ensure the property is machine-readable and supports automatic dataset discovery. Examples:
  - ICD-10: http://www.wikidata.org/entity/P494
  - SNOMED CT: http://www.wikidata.org/entity/Q1753883

- While **Conforms To** property describes the health data standards the dataset follows (e.g. HL7 FHIR, OMOP CDM, CDISC), **Coding System** focuses on the terminologies and classifications used within the dataset (e.g. ICD-10, SNOMED-CT).

### 7.3.4 Information on Relations with other datasets and resources

These properties show how the dataset is connected to other datasets or publications. They help users understand the dataset's context, origin, and possible updates or related resources.

**Table 22. Properties of the Dataset class describing relationships with other datasets or resources.** Cardinalities are shown for open, restricted, and sensitive datasets. M – Mandatory, R – Recommended, O – Optional.

| Property | Cardinalities | | | Question to answer |
|---|---|---|---|---|
| | Public | Restricted | Non-public | |
| Relation | O | O | R | *Is there another dataset or resource that is related to this dataset?* |
| Source | O | O | R | *Was this dataset derived from another dataset or document?* |
| In series | O | O | O | *Is this dataset part of a larger collection or series of datasets?* |
| Referenced by | O | O | R | *Are there publications or resources that refer to this dataset?* |
| Qualified relation | O | O | O | *How exactly is this dataset related to another resource?* |
| Landing page | O | O | R | *Where can users go to learn more about this dataset?* |

#### 7.2.4.1 Relation

This property allows you to indicate a general link between this dataset and another resource (e.g., another dataset, document, or service). Use it when there is a connection, but none of the more specific relation properties below apply.

- This is the most generic way to describe a connection.
- Use it when you want to indicate a related resource but cannot define the relationship more precisely.

When the relationship is known and semantically clear, use one of the more specific properties instead:

#### 7.2.4.2 Source

This property allows you to identify the dataset or document from which this dataset was derived, helping users trace back to the original data source.

- Use it when your dataset was created by processing, aggregating, or modifying another dataset.

#### 7.2.4.3 In series

This property allows you to specify that the dataset belongs to a series or collection of datasets. It improves navigation and understanding of datasets released over time or grouped together.

- Use it when the dataset is part of a **recurring publication**, **periodic release**, or **thematic data collection**.

#### 7.2.4.4 Referenced by

This property allows you to indicate any publication, report, or document that refers to or cites the dataset.

- Link to scientific articles, project reports, or policy documents that have used or cited this dataset.

#### 7.2.4.5 Qualified relation

This property allows you to describe a relationship with a specific role, for example, “updates,” “complements,” or “is replaced by.”

- Use when the nature of the relationship is explicit and meaningful.
- Helps clarify the functional role between related datasets or versions.
- Ideal for describing dataset evolution or any dependencies you consider relevant for the applicant reading the metadata to understand the dataset.

**Choosing the right relation property:**

- Was this dataset created by processing, transforming, or deriving from another dataset or document? Use the **Source** property. This helps users trace the origin of the dataset, giving context and transparency.
- Is this dataset part of a recurring release or thematic collection? Use the **In Series** property. Relevant for datasets published regularly (e.g., annual statistics) or grouped in a project series.
- Has this dataset been cited or used in a publication, report, or other work? Use the **Referenced By** property. This shows where and how the dataset has been used.
- Is there a clearly defined relationship with another dataset, like an update or a complement? Use the **Qualified Relation** property. Best for expressing specific roles, such as “updates,” “is replaced by,” or “complements.” Useful for managing versions or dependencies.
- Is there a general connection to another dataset or resource, but it is not possible to describe the exact type of relationship? Use the **Relation** property. This is the most generic way to show a connection. Use it when no other relation property fits.

#### 7.2.4.6 Landing page

This property allows you to link to a webpage that provides more information about the dataset. Use it to provide a direct link to a web page where users can learn more about the dataset and access its distributions. This can be a page on your organisation's website, a FAIR data point, a national catalogue, or any other resource where the dataset is already described.

- ❑ Make sure the landing page includes clear and useful information (such as dataset description, contact, licence, and how to access the data).
- ❑ This page should be specific to the dataset you are describing.

### 7.3.5 Information on the variables

This property provides information on how the data is distributed across variables, helping users assess the dataset's structure and whether it suits their analysis needs.

**Table 23. Properties of the Dataset class describing information on the variables.** Cardinalities are shown for open, restricted, and sensitive datasets. M – Mandatory, R – Recommended, O – Optional.

| Property | Cardinalities | | | Question to answer |
|---|---|---|---|---|
| | Public | Restricted | Non-Public | |
| Sample distribution | O | O | M | *Is there a detailed description, sample or preview of the dataset structure that users can view before requesting access to the full dataset?* |

#### 7.2.5.1 Sample distribution

This property allows you to provide a safe, representative overview of the dataset's structure and variables. This helps users understand how the dataset is organised, such as which variables it includes, what coding or classifications are used, and the data formats, without without disclosing any sensitive or personal information.

In this property you can provide:

- **A data dictionary** describing the dataset's structure and content (variables, formats, classifications, taxonomies, code lists, etc.).
- **A proxy dataset** that simulates the structure and characteristics of the original data but does not contain any personal data or any information that could reasonably lead to re-identification.

Importantly, any example or sample provided under this property must be demonstrably non-personal data. Under the EHDS Regulation, data with any residual risk of re-identification is still considered personal data and must not be shared as a sample. Therefore, only marked proxy datasets may be used.

The data dictionary should clearly and comprehensively describe 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. In the context of TEHDAS2, we are proposing as technical solution to use CSVW, an extension of the widely used CSV format (See Chapter 8). An example of a data dictionary using CSVW is provided in Figure 1 below.

The ultimate goal of this property is to give data users a reliable and **safe-to-share representation** of the dataset's structure and characteristics. This enables them to assess the relevance of the data for their intended use, while ensuring that no personal or sensitive information is disclosed.

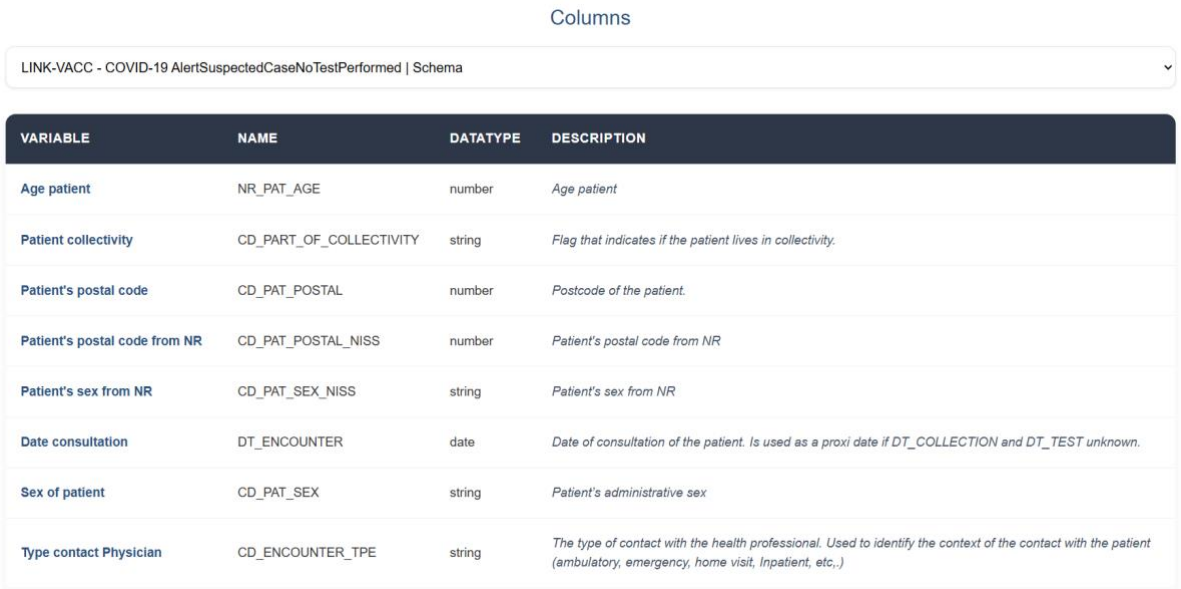


**Figure 1.** Example of a data dictionary using CSVW for the LINK-VACC dataset[24]. The full-example can be consulted in the Health Information Portal demo catalogue (Link).

[24] HealthDCAT-AP – LINK-VACC dataset

### 7.3.6 Information on Access Conditions

These properties explain how users can access the dataset, under what legal conditions, and what distributions are available. This helps users know how they can use the data.

**Table 24. Properties of the Dataset class describing information on access conditions.** Cardinalities are shown for open, restricted, and sensitive datasets. M – Mandatory, R – Recommended, O – Optional.

| Property | Cardinalities | | | Question to answer |
|---|---|---|---|---|
| | Public | Restricted | Non-public | |
| Distribution | **M** | **M** | **M** | *How is the dataset made available for access or download, and in what formats?* |
| Applicable legislation | **M** | **M** | **M** | *What laws or regulations apply to the access of this dataset?* |
| Access rights | **M** | **M** | **M** | *What type of access is allowed for this dataset: open, restricted, or closed?* |

#### 7.2.6.1 Distribution

This property is mandatory in DCAT to ensure that every dataset include at least one distribution, which gives users a way to access the dataset, or, in the case of sensitive data, to access information about how the dataset can be obtained. Each distribution should include key technical and legal information, such as:

- **Access URL** (e.g., the HDAB landing page or the actual download location)
- **Format** (e.g., CSV, JSON, RDF)
- **Byte size**
- **Rights** (conditions for access or reuse)
- **Applicable legislation**, which must reference the **European Health Data Space Regulation ELI** for EHDS datasets

For personal electronic health data (sensitive data) in the EHDS, since access to the dataset is made through a data access application, this distribution should link to an HDAB landing page that provides access to the instructions on how to request access to the dataset. This ensures that even when the data cannot be downloaded directly (e.g., for sensitive health datasets), there is a clear and documented path to access.

- ❑ If the dataset is sensitive and not directly downloadable, the distribution must still be provided, it just takes the form of a link to the HDAB.
- ❑ If there are multiple ways to access the dataset, you can include multiple distributions, each reflecting a different access method.
- ❑ The distribution complements the landing page at the dataset level — the landing page gives general information about the dataset, while the distribution points to specific access methods or files.

#### 7.2.6.2 Applicable legislation

This property allows you to specify the legal frameworks that govern the dataset's creation, management, and use. This is a mandatory property in HealthDCAT-AP. Its purpose is to

make clear which laws apply to the dataset, especially to indicate whether the dataset falls within the scope of the EHDS Regulation.

If a dataset is in scope of the EHDS Regulation, you must include the European Legislation Identifier (ELI) for the EHDS Regulation in this property. You can also include references to other relevant legislative acts that apply, such as the GDPR, the INSPIRE Directive, National laws or regulations on health data or data reuse

- ❑ If your dataset is managed or accessed under the EHDS framework, always include the EHDS Regulation ELI here (you can find this on the EUR-Lex portal).

- ❑ You may include more than one piece of legislation if the dataset is governed by multiple legal acts (e.g., both EU and national laws), examples:
  - European Health Data Space Regulation: https://eur-lex.europa.eu/eli/reg/2025/327/oj/eng
  - Data Governance Act: http://data.europa.eu/eli/reg/2022/868/oj
  - High Value Dataset Act: https://eur-lex.europa.eu/eli/reg_impl/2023/138/oj

- ❑ If you're unsure what laws apply, consult your organisation's legal or compliance team, especially when it concerns personal health data or cross-border data sharing.

#### 7.2.6.3 Access Right

This property allows you to indicate whether the dataset is publicly accessible, has restrictions, or is not publicly available. This is a mandatory property in HealthDCAT-AP, and is crucial for helping users understand whether and how they can access the dataset. You must choose the appropriate value from the Access Rights Authority List (from the EU Publications Office vocabulary). These standardised values ensure consistency and interoperability across datasets.

For datasets that fall under the EHDS Regulation (such as personal electronic health data), the correct value is typically:

- `NON_PUBLIC` – The dataset is not publicly available and access is subject to restrictions (e.g., through an HDAB request process)

When the dataset does not have sensitive information (such as personal electronic health data) other possible values include:

- `PUBLIC` – The dataset is openly accessible to everyone
- `RESTRICTED` – The dataset has access restrictions but may be available under specific conditions

- ❑ **Always use predefined values** from the Access Rights Authority List – avoid free-text values.
- ❑ If your dataset contains **personal electronic health data**, the correct value is **`NON_PUBLIC`**.

### 7.3.7 Information on Versioning

These properties provide details about the version history of the dataset, including when it was created or modified. This helps users track updates and ensure they are working with the correct version.

**Table 25. Properties of the Dataset class describing information on versioning.** Cardinalities are shown for public, restricted, and non-public (sensitive) datasets. M – Mandatory, R – Recommended, O – Optional.

| Property | Cardinalities | | | Question to answer |
|---|---|---|---|---|
| | Public | Restricted | Non-public | |
| Version | O | O | O | *What is the current version number or label of this dataset?* |
| Has version | O | O | O | *Are there other versions of this dataset available?* |
| Version notes | O | O | O | *What has changed in this version of the dataset?* |
| Issued | O | O | O | *When was this version of the dataset originally published?* |
| Modified | O | O | O | *When was this dataset last updated?* |

**Understanding dataset versions:** Use versioning properties to describe how this dataset relates to previous or newer versions in a series. These properties help users track dataset evolution, understand changes over time, and cite the correct version for reuse.

#### 7.2.7.1 Version

This property allows you to declare the version number or identifier of the current dataset. Use it to provide a clear and unique version name or number (e.g., "3.0", "v1.2-beta"). This helps users reference or cite the exact dataset version used in their work.

#### 7.2.7.2 Has version

This property allows you to point to another dataset that is a version, edition, or adaptation of the current one. Use it to link to a newer version or related adaptation. For instance, if this is version 3.0 and version 4.0 has been released, you can use has version to point users to the newer dataset.

#### 7.2.7.3 Version notes

This property allows you to describe what changed in this version of the dataset. Use it to briefly explain the key updates, improvements, or differences compared to previous versions. This helps users understand whether this version meets their needs.

**Choosing the right property**

- What version is this dataset? → Use Version
- Is there a newer version of this dataset available? → Use Has Version
- What changed in this version compared to earlier ones? → Use Version Notes

#### 7.2.7.4 Issued

This property allows you to specify the date of formal issuance or publication of the dataset. Use it to state the official release date of the dataset - this is when the dataset became officially available.

#### 7.2.7.5 Modified

This property allows you to indicate the most recent date when the dataset was changed or updated. Use it to reflect any significant updates or modifications to the dataset content, such as:

- New data entries added
- Corrections or reclassifications
- Updates to coding systems
- Changes in structure or variables

❑ Tip: This should not include minor technical adjustments to the metadata record itself, but rather reflect changes in the dataset content.

### 7.3.8 Information for metadata management and search functionalities

These properties support the categorization, semantic annotation, and filtering of datasets when HealthDCAT-AP is applied in a data catalogue. They are essential for improving findability and helping users locate datasets that are most relevant to their research questions. You can already see how some of these properties are integrated in the search functionalities of the EU Dataset Catalogue at HealthData@EU Central Platform (Link).

**Table 26. Properties of the Dataset class describing information for metadata management and search functionalities.** Cardinalities are shown for public, restricted, and non-public datasets. M – Mandatory, R – Recommended, O – Optional.

| Property | Cardinalities | | | Question to answer |
|---|---|---|---|---|
| | Public | Restricted | Non-public | |
| Type | O | O | M | *What type of dataset is this?* |
| Health Category | M | M | M | *To which Article 51 category does this dataset relate to?* |
| Theme | M | M | M | *What general subject does the dataset cover?* |
| Health theme | O | O | M | *What specific medical or health-related topic does the dataset focus on?* |
| Trusted Data Holder | O | O | R | *Is the data holder officially recognised as a Trusted Data Holder under the EHDS?* |
| Quality annotation | O | O | R | *Is there any documentation on the quality of the dataset?* |
| Legal Basis | O | O | R | *What is the legal framework that allows this dataset to be processed or accessed?* |

#### 7.2.8.1 Type

This property identifies the classification of the dataset using a controlled vocabulary. Use it to specify if the dataset contains personal electronic health data or falls under another category (e.g., statistical data, geospatial data). Use values from the Dataset-Type Controlled Vocabulary[25] maintained by the Publications Office. For datasets containing personal electronic health data, use the mandatory value PERSONAL_DATA.

- Use only values from the official controlled vocabulary.
- Always use PERSONAL_DATA for datasets containing personal electronic health data, as required by the EHDS Regulation.

#### 7.2.8.3 Health Category

This property indicates the EHDS data category the dataset belongs to, based on Article 51 of the EHDS Regulation. Use the official controlled vocabulary to ensure standardised classification across all datasets. This property helps users filter and discover datasets relevant to their needs.

- Choose the category or categories that best match the dataset's content, using the controlled list based on Article 51.

#### 7.2.8.4 Theme

This property categorises the dataset using the Dataset Theme Controlled Vocabulary from the Publications Office[26]. For datasets under the EHDS Regulation, the theme must be set to: http://publications.europa.eu/resource/authority/data-theme/HEAL

- Always use the "HEAL" value for datasets within EHDS scope.
- This supports consistency and interoperability across EU data spaces.

#### 7.2.8.5 Health theme

The Health Theme property complements the free-text keywords by introducing concepts in a machine-readable and actionable way. Instead of using plain text, this property requires a link to a defined concept, making the dataset easier to discover and process by automated systems.

To ensure consistency and semantic alignment, we propose using Wikidata[27] as the centralised knowledge base. Wikidata offers a large-scale, multilingual, open, and ontologically structured environment that supports both human and machine understanding. It enables alignment across datasets by linking shared biomedical concepts. For more details and implementation guidance on using Wikidata for Health Themes, consult Chapter 8.

[25] http://publications.europa.eu/resource/dataset/dataset-type
[26] http://publications.europa.eu/resource/dataset/data-theme
[27] Wikidata

- Use the **Wikidata URI** for each concept (not the label or free-text), by using Wikidata's search to find the correct concept: https://www.wikidata.org/
- Match your keywords to their corresponding Wikidata concept.
  Example: If you selected the keyword **"Diabetes"**, the corresponding health theme would be: https://www.wikidata.org/wiki/Q12152
- Prefer specific terms where possible to improve precision in categorisation.

#### 7.2.8.6 Trusted Data Holder

This property indicates whether the health data holder is recognised as a **Trusted Health Data Holder** under the EHDS Regulation (Article 71). Trusted Health Data Holders are authorised to make data available for secondary use under specific conditions defined in the Regulation. Use this property to **declare the status** of the data holder, whether or not they are recognised as a Trusted Health Data Holder.

- Provide a **yes/no** value based on the designation status.
- If marked "yes", ensure the organisation is officially recognised under Article 71 of the EHDS Regulation – more information on how to become a Trusted Data Holder will be provided in the future at both European and Member-state level

#### 7.2.8.7 Has 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. This includes a statement related to quality of the dataset, including rating, quality certificate, feedback that can be associated to the dataset.

In this property, 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 – More information on the relation with the data quality and utility label being developed in QUANTUM project on Chapter 9.

#### 7.2.8.8 Legal Basis

Use this property to indicate the legal or regulatory basis under which the data was originally collected and processed, especially relevant for personal data under the GDPR.

This legal basis refers to the initial data processing, not the secondary use regulated by the EHDS Regulation. Including it promotes transparency and helps users understand if the dataset meets compliance requirements.

- Specify the relevant GDPR article or national law that justifies data collection (e.g., public interest, consent).
- Use this property for datasets containing personal data to clarify lawful collection.

# 8 Proposed solutions for HealthDCAT-AP implementation

## 8.1 Scope

As explained previously, one of the key objectives of this task in TEHDAS2 was to refine and validate the HealthDCAT-AP metadata model to ensure it meets the needs of the health data community and supports the implementation of the EHDS. To achieve this, a series of structured activities were undertaken with a strong focus on stakeholder engagement and technical discussion. A total of seven dedicated Technical Working Groups (TWGs) took place, each tasked with addressing specific elements of the HealthDCAT-AP model that required further clarification, improvement, or validation. These sessions brought together domain experts, technical implementers, and representatives from national and European bodies to share feedback and build consensus on the proposed solutions.

The TWG discussions covered the following key topics:

- **TWG1 Lineage and Provenance**: Explored the use of properties such as Data lineage, Provenance, QualifiedAttribution, and GeneratedBy. The goal was to improve the usage notes and assess the need for controlled vocabularies.
- **TWG2 Health Theme**: Focused on the suitability of using Wikidata as the source for controlled vocabularies representing health themes.
- **TWG3 Publisher Type**: Aimed to develop a controlled vocabulary for the Publisher type property and discourage the use of ambiguous categories like "other."
- **TWG4 Property Cardinalities**: Reviewed and validated the cardinality rules for HealthDCAT-AP properties to ensure consistency and clarity in metadata implementation.
- **TWG5 Sample Distributions**: Discussed how to best represent sample distributions, validated the proposed approach, and refined usage guidance and cardinalities.
- **TWG6 Data Dictionaries and Codebooks**: Evaluated CSV on the Web (CSVW) as a potential standard to support structured documentation of data dictionaries and codebooks.
- **TWG7 Open Questions**: Provided space to address additional outstanding issues raised throughout the process.

In addition to these sessions, we integrated feedback received during the public consultation on *Milestone 5.1 – Draft Guideline on Data Description*[28], which offered valuable insights into practical challenges faced by data holders. Together, these activities resulted in a set of targeted solutions to improve the clarity, usability, and technical robustness of HealthDCAT-AP. The following sections present the key topics discussed and the corresponding refinements proposed to the model.

**Note:** These proposals represent the outcome of TEHDAS2 work and do not constitute formal guidance, regulatory interpretation, or endorsement by the European Commission.

[28] Milestone 5.1 - Draft guideline on Data Description

## 8.2 Cardinalities

One of the key topics discussed in the technical working groups was the cardinality of metadata properties, specifically, whether certain properties should be considered mandatory, recommended, or optional. This dialogue allowed experts to provide input on the importance of each property and the rationale behind its proposed status in HealthDCAT-AP. It also offered an opportunity to reflect on how best to represent these properties to support clarity, usability, and interoperability. This section highlights the most debated properties in terms of cardinality and the reasoning behind the decisions.

**Note:** These cardinality proposals go beyond the minimum legal requirements set out in the EHDS Regulation and aim to support best practices for implementation.

### 8.2.1 Keywords and Health Themes

**Cardinality: Both properties set as mandatory**

As part of the review of property cardinalities, we are proposing that both dcat:keyword and healthdcatap:healthTheme be made mandatory in HealthDCAT-AP. These two properties were discussed together, as they play complementary roles in enhancing dataset discoverability and interoperability.

The technical working group first focused on clarifying the purpose of each property:

- dcat:keyword is intended for free-text, human-readable tagging. It is already widely used and plays a crucial role in improving dataset searchability in catalogues. There was consensus among experts that this property should be mandatory.
- healthdcatap:healthTheme is designed for structured, machine-readable annotation using controlled vocabularies such as Wikidata. It allows datasets to be semantically described in a way that supports automated discovery, reasoning, and cross-catalogue integration.

Although healthTheme is a newer property, its complementary function was well understood during the discussions. It expands the value of dcat:keyword by introducing a semantic layer that helps create a machine-actionable framework for thematic classification.

**Remaining Challenges highlighted:** The use of Wikidata as a source for controlled vocabulary health themes raised some concerns, particularly regarding the consistency and appropriateness of available concepts. While there was consensus on the importance of health Theme property, it was acknowledged that further work is needed to mitigate the challenges of relying on Wikidata. These issues are explored in detail in the dedicated section on the Health Theme property and controlled vocabularies (see Sections 8.4 and 8.5 ).

**Note:** While Wikidata has been explored as a potential vocabulary source, it is not officially maintained or endorsed. Its use therefore remains pending further validation or institutional governance.

### 8.2.2 Provenance

**Cardinality: Mandatory**

The property dct:provenance is proposed as mandatory in HealthDCAT-AP, as it plays a fundamental role in helping data users understand the origin, creation process, and context of a dataset. This information is essential for assessing the fitness of the dataset for secondary use in health research and policymaking.

dct:provenance is intended to provide a concise, human-readable statement that describes how the dataset came into existence. It should ideally cover:

- The origin of the data
- The methodology or process used to create it
- Any significant transformations or updates
- The entities involved in its creation or curation

Experts emphasized that this kind of contextual information is critical for data users to assess whether a dataset is appropriate for their specific use case. It is particularly important in the health domain, where the validity of secondary use often depends on understanding the primary data collection process. While the majority of the working group participants supported making this property mandatory, some concerns were raised: 1) In specific cases, detailed provenance information may be unavailable and 2) there is a risk that making the field mandatory could prevent the publication of otherwise useful datasets if data holders cannot provide the required information. A solution was found in allowing a statement on the absence of provenance information and explaining why is still valuable. Since this is a free-text field, data holders can simply indicate, for example:

"*No detailed provenance information is available for this dataset due to its age and lack of documentation.*"

This approach maintains the value of transparency without excluding datasets from being catalogued. The group acknowledged the property as mandatory, under the condition that guidance is provided for data holders on how to fill it in when full provenance information is unavailable. This ensures that the field remains meaningful while accommodating real-world data availability constraints.

### 8.2.3 Contact Point

**Cardinality: Mandatory**

The property dcat:contactPoint is proposed as mandatory, reflecting its importance in ensuring that potential data users have a clear and reliable point of contact for inquiries about the dataset.

The discussion in the working group emphasized that this property often overlaps with the publisher information. In several Member States it is already common practice to use the publisher, typically the organization responsible for the dataset, as the default contact point. Defining the contact point as mandatory reinforces the need for a reachable and accountable representative for each dataset. However, a concerned raised was that, in research contexts, the contact person may no longer be available once a project ends, making it difficult to maintain up-to-date individual contact details. A suggested solution was to allow the use of

an organizational contact (e.g., a university or department) instead of a named individual. This ensures continuity, as organizations are better positioned to redirect inquiries to the appropriate person. While this approach may impose a small additional burden on data holders, particularly when assigning a consistent contact point, the consensus was that the benefit of ensuring communication pathways outweighs the challenge.

### 8.2.4 Sample Distribution

**Proposed Cardinality: Mandatory**

**Box 6. Why is sample distribution property needed?**

The sample distribution property provides variable-level metadata, typically in the form of a **data dictionary** or a **synthetic preview**. This metadata is essential to support dataset discoverability, appropriate use, and compliance with legal interoperability obligations under the broader European data framework.:

1. **Alignment with the Data Act:** The inclusion of this information is required by Article 33 of the Data Act (Regulation (EU) 2023/2854), which establishes that participants in European data spaces must provide adequate metadata describing dataset content, structure, data formats, and classification systems in a machine-readable format. EHDS should build on and align with this broader legal framework. HealthDCAT-AP, as part of the broader European data framework, must therefore align with these interoperability obligations, fostering legal and technical consistency across data spaces.

2. **Supporting discoverability and appropriate reuse:** There was broad consensus in the TEHDAS2 working group that **metadata at the variable level is critical for secondary use**. Without this level of detail, data users are unable to properly assess the relevance and applicability of a dataset for their intended purpose. Variable-level metadata allows data users to evaluate the suitability of datasets **before submitting access requests**, avoiding wasted time, effort, and money on data that turn out to be unfit for purpose. Without this information, datasets risk being treated as “black boxes”, which limits their usability and increases the likelihood of misinterpretation. he general feedback from the public consultation on Milestone 5.1 further supports this: One of the most frequently raised challenges was the lack of metadata at the variable level. Multiple respondents to the public consultation on Milestone 5.1 cited the lack of variable-level information as a major barrier to secondary use, requesting for more granular information, including value sets, data distributions, and structured variable descriptions.

3. **Reducing burden on data holders and streamlining procedures:** Without this information in the metadata record, data holders are likely to receive repeated requests from applicants seeking data dictionaries, codebooks, or other clarifications. This increases their workload by requiring them to respond individually to queries about the dataset’s structure and content. By providing this information upfront through a structured data dictionary, they can significantly reduce repetitive communication and avoid inefficiencies or delays in the data access process.

The Sample Distribution property was the most discussed in terms of cardinality, not because of disagreement over its usefulness, but due to implementation challenges:

- Lack of tooling to automate the generation of data dictionaries, especially in legacy or non-tabular formats (e.g. images, free text);
- Heterogeneous portfolios of datasets, often undocumented or stored in distributed systems;
- Resource constraints, especially for public hospitals or smaller research entities;

While these challenges were acknowledged, the majority of stakeholders agreed that t while these implementation barriers are real, they can be progressively addressed through adequate support and capacity-building, without lowering the ambition of the model. The working group expressed concern that making the property only recommended could slow progress toward improved metadata quality and maintain existing limitations in dataset discoverability. The final position, supported by a slim majority (30 votes for mandatory, 28 for recommended, 18 optional), recommended a mandatory status, coupled with transitional support and progressive implementation mechanisms. There was widespread consensus in the discussions, and in prior public consultation feedback, that sample distributions, including data dictionaries and/or anonymised samples, are essential to support high-quality data reuse – this was the main rationale for recommending a mandatory status within the model, although we recognize that major efforts would be necessary for the successful implementation on a large-scale**.**

**Implementation challenges and future work**

While there is broad consensus on the importance of providing sample distributions and variable-level metadata, the practical implementation of this property remains challenging.

A key concern relates to the **significant effort required from data holders**, particularly those who manage large and heterogeneous collections of datasets. For many, especially those lacking prior experience with research data dissemination, creating detailed data dictionaries may be a new and resource-intensive task. This is particularly true for datasets that were not originally collected for secondary use, and for which documentation may be incomplete or non-existent. In such cases, even identifying the necessary metadata can be a time-consuming process. Participants also raised the challenge of applying this property to non-tabular data types, such as images or free-text records. The concept of a “data dictionary” is typically associated with structured, tabular formats, raising open questions about how to adapt this requirement for other data modalities within the EHDS. Another major concern is **scalability**. Some institutions may be responsible for dozens or even hundreds of datasets. Without automation or appropriate tooling, manually creating and maintaining metadata records for each one is simply not feasible. There is a risk that, if this property is made mandatory without supporting infrastructure, it could result in poor compliance or overburden already limited institutional capacities. However, despite these challenges, the working group agreed that these are not unsolvable technical problems. Rather, they would require **investment, coordination, and time**. The discussion made clear that the main barrier is not the technical complexity per se, but the current limitations in tooling and resourcing capacity across different contexts. To address these issues, several possible pathways were discussed:

- First, there was agreement on the need for **scalable and user-friendly tools** that can assist data holders in generating data dictionaries. Ideally, such tools could build on existing internal cataloguing systems or provide simple, harmonised ways to generate structured metadata (e.g. via CSVW), depending on national contexts and available infrastructures..

- There was also a strong suggestion to provide transitional **flexibility** in the early years of EHDS implementation through a progressive validation model: In early stages of EHDS implementation, allow for incremental metadata enrichment, where datasets can be published with minimal fields initially, and gradually improved over time. Feedback agents or automated assistants may support this process.

- Finally, participants emphasised that this property should not be treated as a standalone requirement, but rather as part of a broader, collaborative implementation effort. Achieving widespread provision of variable-level metadata will require joint action from the European Commission, national authorities, HDABs, and data holders Achieving this goal may require, in addition to technical standards, capacity-building efforts such as training, funding opportunities, and appropriate governance mechanisms, adapted to national implementation contexts.

In conclusion, while the implementation of this property is undeniably demanding, it is the opinion of TEHDAS2 participants that omitting this element could reduce the effectiveness of the EHDS in supporting high-quality secondary use and structured data discovery.. The long-term vision is one in which all datasets include structured, accessible, and informative sample distributions, and that vision can only be reached with clear, mandatory expectations, paired with appropriate transitional measures.

The discussion group recommended a mandatory status for this property within the model, as a means to drive long-term improvements in metadata quality and reuse across Europe. Without this effort, there is a real risk that the needed tools, support systems, and methodologies will never be developed, perpetuating current gaps in metadata quality and limiting the reuse potential of health data in Europe. At the same time, it strongly acknowledged the **critical need for short-term flexibility** and emphasized the importance of progressive, sustained support to help data holders gradually reach this implementation goal.

**Note:** Ultimately, the final decision on the status of this property, whether mandatory or recommended, and the corresponding implementation approach, including possible support measures, will rest with the European Commission and the Regulatory Committee as part of the adoption of the implementing act on dataset description.

## 8.3 Use of CSVW for data dictionaries

To enable variable-level metadata in a consistent, reusable, and machine-readable way across the EHDS, a common approach for describing data dictionaries is needed. This approach must fulfil two essential criteria:

1. **Flexibility and adaptability**, to accommodate the diversity of data types encompassed by the 17 categories listed in Article 51 of the EHDS Regulation
2. **Ease of use and accessibility**, to reflect the varying levels of technical maturity and resources among data holders across Europe.

**CSV on the Web (CSVW)** was identified as an example of a solution that meets both requirements. CSVW is a W3C standard that enables structured metadata to be attached to CSV files, allowing tabular data to be semantically described in a lightweight, machine-readable format. It supports incremental adoption, integrates easily into existing workflows, and can be used both in high-resource environments and low-complexity contexts where manual editing may still be the norm. While not a mandatory or legally binding under the EHDS Regulation, CSVW is proposed here as an open and pragmatic solution to support interoperability and usability within the EHDS.

Several alternative standards were considered during the discussion, including DDI-RDF, GSIM, among others. These standards offer powerful frameworks for describing complex datasets and are well-established in domains such as official statistics, survey-based research, and clinical systems. However, for the broad and diverse context of the EHDS, they were found to be less suitable as a common baseline solution for two main reasons:

- **Domain specificity**: Many existing standards are tightly linked to particular types of data or sectors (e.g. statistical surveys, clinical records), making them difficult to generalize across all 17 EHDS data categories. For instance, DDI-RDF and GSIM are primarily designed for statistical data and are widely used in the official statistics and survey research domains. They are structured around concepts such as populations, sampling frames, and statistical processes, which are not applicable to many of the other data categories in the EHDS, such as real-world data from wearable devices or environmental health data.

- **Complexity and implementation burden**: Their adoption typically requires specialised expertise, tooling, and a level of institutional maturity that may not be present in all data holding organisations, particularly those without existing semantic infrastructures. Tools like DDI and SDMX require the use of XML-based or RDF-based authoring environments, ontology mapping tools, and knowledge of semantic web technologies. Many data holders, especially those managing administrative or registry data, do not have the capacity or resources to implement such frameworks. Some of these standards also come with rigid data models and mandatory components that make it hard to adopt them incrementally. This creates a high barrier for smaller institutions or those with less mature data infrastructure. Additionally, feedback during the discussion highlighted that, in real-world scenarios, metadata is often still managed in spreadsheets or flat files. Solutions that diverge too far from this reality would not be usable without a major shift in capacity and tooling, which is not feasible in the short term.

CSVW offers a middle ground: it provides a semantically rich, standardised way to describe tabular data without requiring a complete overhaul of current workflows and heavy technical or organisational burden. Metadata can be authored directly alongside CSV files using familiar tools like Excel or basic text editors, while still enabling machine readability and integration into broader metadata infrastructures.

**Note:** Despite the proposal of CSVW by the TEHDAS2 technical working group, the use of other domain-specific standards like the ones described above is not excluded. They may continue to be referenced in national or sectoral guidance where appropriate.

**What information should be in the data dictionary?**

A key theme of the discussion was the level of detail and the minimum information elements that should be included in a data dictionary. Participants broadly agreed that setting a clear

minimum requirement is essential to ensure consistency and usability, while allowing for extensibility where more detailed metadata is available.

An initial exploratory exchange with the group gathered views on the essential information a data user needs to understand a dataset's variables. This included input from organisations that already provide variable-level metadata, as well as from data users who regularly face gaps in documentation. The group highlighted the following as core metadata elements that should be present in any CSVW-based data dictionary (Table 27).

**Table 27.** Feedback from the technical working group on metadata attributes essential and important to be included in a CSVW data dictionary.

| Attributes mentioned by the participants | Definition |
|---|---|
| **ESSENTIAL INFORMATION** | |
| Name/Label **of the variable/column** | Short descriptive name of the individual variable, e.g. Gender. The "human understandable" name for a variable. ; Could support presentations, visualization and analytics etc. |
| Description/Definiton **of the variable/column** | Detailed description of the variable ; Full and detailed definition of the column. |
| **Variable** data type | Examples: Number, Integer, Data field, Date, Date-time ; Data type as specified in FHIR Primitive Types ; Technology agnostic data type for the column |
| Code value **of the variable** | Both possibility to link to the international coding system and provide the national codes and values ; Complete list of values and allowed values for this column if not extensive (else link to complete list in Coding system) ; - A linked ValueSet (OID, URI etc) where you can find more detailed information about the codes and values. Used for enumerated variables based on a classification, code list, value set or questionnaire (for example EQ-5D) which can be linked. |
| **Information on the vocabulary used** | Customised list of allowed values (e.g. “M” and “F” for Gender) ; for instance binary one means male, 2 means female ; Annotation of the list of allowed values (e.g.: M=male;F=female). When not described in a URL. |
| **ADDITIONAL INFORMATION** | |
| **Variable** identifier | Unique name string for a variable |
| Technical name **of the variable/column** | The name of the variable as it is in the dataset. This can be used to connect metadata with data. If the variable does not have a separate "human understandable" name and a technical name, the same information should be entered in both fields. |
| Ontology **or** RDF type ID | A unique identifier that captures the type of the variable. Semantic types such as schema.org or ontology terms enhance the findability of the data in repositories. |
| **Field Size** | The size (length) of the variable value, e.g. 8 digits, 5,3 (for floating numbers)… |
| **Max Allowed Value** | Upper limit of the allowed value |
| **Min Allowed Value** | Lower limit of the allowed value |
| **Coding System** | Link to comprehencive coding systems not possible to show in Valid Values, e.g. ICD or ATC |
| **Measurement Type** | What the variable measures. E.g. time, age, weight and heart rate |
| **Code for missing value** | "Enter the code used for the missing information in the data in the field. Example: -1" |
| **Precision** | Maximum decimal places. |
| **Conforms to** | This property is used to refer to a standard used / mapped to (for example FHIR, OMOP, SNOMED CT, dictionaries). |

| | |
|---|---|
| **Presentation/Column Order** | The sequence number of the column in a logical order used for sorting the columns in the data set. Can be used for presenting the columns in a logical order. |
| **Data Representation Format** | Data representation format (eg date format YYYY-MM-DD) |
| **Value** domain type | Example: Continuous or Discrete ; Examples of continuous variables: time, length, energy consumption Examples of discrete variables: gender, appointment recipient ; To unambiguously specify if the data associated with the variable being defined should be treated as a continuous variable, discrete/polychotomous variable or an ordinal variable. |
| **Missing variable data** | If there are such factors, this is a space to enter information that can impact interpretation of the information (for example, if a large amount of information is missing from the variable, coverage of the variable and relevant temporal changes) ; A measure of non-response. (if the measure is stable over time and across the entire dataset. Otherwise, this is presented in the detailed quality description for the dataset.) |
| **Computed Value** | If a field is computed based on values from other fields, annotate the calculation rule (e.g BMI= WEIGHT/(HEIGHT*HEIGHT)). |
| **Collection Form Name** | Optional, if the field is collected in certain forms (e.g. in Case Report Forms from a clinical trial). |
| **Quality Description** | Other relevant information about column-specific quality beyond non-respons and comparability ; Overall textual description of the quality of the data that the variable represents, e.g. completeness in the form of completeness (encoding quality) and/or code quality. |
| **Measurement Unit** | Which unit of measure the variable is based on. E.g. year, month, kg, gram, beat/minute |
| **Replaces** | Link to the variable that is replaced by this variable. |
| **Is replaced by** | Link to the variable that has replaced this variable. |
| **Question in questionnaire** | The question / text in the questionnaire or reporting form for each variable. |
| **Degree of identification** | The data manager's classification of contributions to risks for identification of individuals. Useful information for the researcher when the project needs to consider measures for data minimisation and for the HDAB. |
| **Linked Column** | Indicates whether the column is a linked column, i.e. whether it is used to link to other datasets |
| **Linking Description** | Describes how the Linked Column is used in different links |
| **Geographical Data Quality** | Geographical Data Quality describing how the quality varies over different geographical regions |
| **Temporal Data Quality** | Temporal Data Quality describing how the quality varies over time |

Importantly, while this list reflects TEHDAS2 consensus on critical elements, it serves as input for future discussions. Final decisions on mandatory attributes will be made by the European Commission and relevant governance structures during implementation. The definition of mandatory elements for data dictionaries falls outside the scope of this task. This work should be undertaken in future efforts, in close collaboration with data holders and domain-specific communities, to ensure that the unique characteristics and methodologies of each data type are adequately addressed. It is essential that any commonly agreed set of mandatory elements remains limited to the most essential information, specifically elements that are universally applicable across all dataset types, such as variable names and descriptions. The goal of adopting CSVW, an open standard, is to offer a simple, open, and practical standard for creating data dictionaries, supporting the EHDS objectives on interoperability and openness, rather than imposing a rigid or overly prescriptive framework. This does not exclude the use of other suitable formats where appropriate. The data

dictionary should be viewed as a complementary component of the metadata record, which already provides a substantial amount of descriptive information about the dataset.

Overall, the discussion confirmed both the relevance of CSVW and the necessity of a common, scalable approach to building data dictionaries. However, participants also underlined that successful implementation will require further work. This includes:

- Defining and formalising the minimum metadata elements expected for all datasets;
- Providing clear guidance, examples, and templates for data holders;
- Developing or supporting tools to help automate the creation of data dictionaries, especially for organisations with limited resources;
- Aligning this effort with the broader metadata architecture of the EHDS and ensuring interoperability with existing standards where needed.

These next steps are essential to support data holders in creating high-quality, variable-level metadata and to unlock the full potential of health data for secondary use across Europe.

## 8.4 Wikidata as semantic framework

To enable interoperability, discoverability, and effective data reuse within the EHDS, it is essential to establish a common semantic layer. A key aspect of this semantic infrastructure is the use of **machine-readable keywords** to describe datasets in a consistent, structured, and accessible manner. These keywords must be linked to a **centralised ontology** that enables cross-referencing between datasets, facilitates automated reasoning, and ultimately supports seamless integration across the 17 data categories outlined in Article 51 of the EHDS Regulation. Currently, metadata records in HealthDCAT-AP include a `healthTheme` property, intended to describe the dataset's thematic scope. However, for this property to serve its purpose in supporting semantic interoperability, it must be populated using a **linked, multilingual, and machine-actionable vocabulary** that spans the full spectrum of EHDS data categories, from clinical data and genomic records to administrative, environmental, and socio-economic datasets. To meet this need, TEHDAS2 Technical Working Group proposes Wikidata as a foundational semantic framework for the EHDS.

**Note:** These recommendations were developed by TEHDAS2 partners in the context of stakeholder and technical discussions. The final decisions regarding the use of Wikidata or other semantic frameworks within the EHDS lie with the European Commission and the Regulatory Committee, subject to legal, technical, and governance assessments before formal adoption.

### Why do we need a centralised ontology for EHDS[29]?

The health domain has several taxonomies and ontologies that are well-established and commonly used. Well-known systems like MeSH, SNOMED-CT, UMLS, ICD, and LOINC provide comprehensive vocabularies for various aspects of clinical and biomedical data. However, in the context of EHDS, which must cover a broad range of health data types (17

[29] Turki, Houcemeddine, et al. "Wikidata: A large-scale collaborative ontological medical database." *Journal of biomedical informatics* 99 (2019): 103292 https://doi.org/10.1016/j.jbi.2019.103292

categories defined in Article 51), from clinical and genomic data to environmental and socio-economic factors, these existing ontologies present limitations regarding its applicability in the EHDS for secondary use:

- **Ontologies can lack important concepts or relations:** Even in mature biomedical ontologies, key concepts or connections may be missing. If a specific disease, exposure, or relation is absent, it becomes impossible to describe the corresponding dataset accurately. This limits both metadata completeness and dataset discoverability.

- **Incompatibility of ontologies:** Different ontologies often cover different domains and are built on incompatible systems or logic. For instance, a term might be coded differently in SNOMED-CT than in MeSH, making integration across datasets difficult. This fragmentation is one of the biggest semantic challenges in the health domain today.

- **Domain-specificity and complexity:** Most existing ontologies were developed for specific professional or clinical communities. While they are detailed and precise, they are often too complex for general use by the diverse range of EHDS data holders, especially those with lower semantic maturity. These stakeholders require a system that is intuitive, accessible, and adaptable across domains and not just for clinicians or medical researchers.

  Example: MeSH (Medical Subject Headings) is widely used for indexing medical literature, but it is not suitable as a standalone solution for EHDS metadata. It is primarily focused on biomedical publications, insufficient for describing non-clinical categories such as environmental or administrative data and lacks flexibility for integrating with datasets not linked to academic publications.

- Additionally, there is the challenge of structural limitation. Most ontologies in the biomedical domain are curated and verified by a closed group of domain experts. While this helps ensure accuracy, it introduces additional problems: missing concepts or relations are slow to be added, and incompatibility remains unresolved between ontologies serving different specialisations. This results in a **fractured landscape** of isolated, partial databases, each valuable but disconnected from the broader knowledge ecosystem.

**Practical Example in EHDS:** A data holder in one Member State may describe a dataset using SNOMED-CT, while another uses MeSH for the same topic. A user searching for datasets on a specific condition using MeSH terms may completely miss the relevant datasets described using SNOMED-CT. No semantic link exists, even though both datasets refer to the same real-world concept. This breaks dataset discoverability, one of the key goals of the EHDS.

These considerations illustrate why the EHDS requires:

- A single, centralised, and voluminous biomedical semantic knowledge base.
- One that is easily editable, verified, and interlinked with existing standards.
- Capable of covering all 17 EHDS data categories for secondary use

- Simple enough to be adopted by data holders with varying levels of technical maturity.

In other words, it requires **an ontology that can integrate and align existing taxonomies, not replace them**. One that **supports semantic interoperability by acting as a bridge between fragmented systems**.

**How can WikiData support the semantic framework for EHDS?**

Wikidata is a collaboratively maintained, machine-readable knowledge base designed to support linked data.

**Data Organisation in Wikidata**

- **Items**: Represent concepts in the ontology (e.g. *trachoma*).
- **Properties**: Define relationships between items. These fall into different types:
  - **Taxonomic relations**: e.g. instance of (P31), subclass of (P279), part of (P361)
  - **Non-taxonomic relations**: e.g. drug used for treatment (P2176)
  - **Database matching relations**: e.g. PubMed ID (P698)
  - **Ontology alignment relations**: e.g. MeSH ID (P486), Disease Ontology ID (P699), SNOMED-CT ID (P5806)

**Example: *Trachoma* in Wikidata**

The concept *trachoma* in Wikidata (trachoma - Wikidata) illustrates the suitability of this solution to tackle the challenges identified:

- It includes multilingual labels and definitions.
- It links to external ontologies:
  - MeSH ID: D014128
  - PubMed ID: 11838089
  - SNOMED-CT ID: 29369001
  - ICD-10 Code: A71
- These references allow any data described using one of these ontologies to be semantically linked through the central *trachoma* concept in Wikidata.

This demonstrates how **Wikidata functions as a semantic bridge**, facilitating cross-ontology search, tagging, and integration.

**What are the challenges in using WikiData?**

While Wikidata offers substantial potential, the TEHDAS2 technical working group identified important challenges:

1. **Multiple Entries for the Same Concept:** There can be duplicate or near-duplicate entries for the same medical concept, making it hard to determine the authoritative item to use. This reflects the current lack of governance over item creation and merging.

2. **Lack of Central Governance:** While Wikidata operates as a community-driven, open platform, it lacks formal governance or validation processes specifically aligned with public health authorities. This decentralised model can introduce risks such as inconsistencies, incomplete metadata, or misalignments with official health standards and requirements.

3. **Quality Assurance Concerns:** Some stakeholders worry about the reliability of crowd-sourced content. Although many quality mechanisms exist (e.g. property constraints, community monitoring), confidence is not yet universal.

4. **Usability Challenges:** Non-expert users may struggle to find the correct entry among several similarly named items. Without training or clearer guidance, adoption by data holders with limited semantic expertise remains a barrier.

5. **Complex Property Landscape:** Understanding and correctly applying the diverse types of properties (e.g. taxonomic vs. associative vs. external identifier) requires training and documentation. Moreover, any proposed use of Wikidata within the EHDS would need to be carefully assessed for compatibility with existing EU-controlled vocabularies (e.g. EuroVoc, EU Vocabularies Authority Lists) as well as established public health terminologies already in use across European Commission systems.

**Considerations for implementation**

Despite these challenges, the structural flexibility and open nature of Wikidata make it a suitable solution for the EHDS, provided appropriate governance and support structures are established. To make Wikidata (or a similar centralized, flexible, machine-readable platform) an operational component of the EHDS, several steps are necessary:

1. **Governance framework**: A European-level governance model must be established, potentially under the leadership of the European Commission. This model should mirror successful initiatives such as DCAT-AP, involving structured collaboration and oversight mechanisms.

2. **Community engagement**: Health data communities working with the different Article 51 data categories must be actively involved. They can curate domain-specific entries, validate terminology, and ensure the platform remains relevant and accurate.

3. **Support and usability improvements**: A user-centered approach must be taken to address current usability issues, for example, the difficulty data users experience in identifying the right concept among similar entries. With governance and community input, guidance and tools can be developed to facilitate accurate tagging and discovery.
4. **Sustainability and scalability**: Infrastructure and funding models need to be considered to support long-term development, tool integration, and training for data holders at all maturity levels.

Overall, it is acknowledged that the EHDS requires a shared, centralized semantic framework to ensure data interoperability, discoverability, and reuse. While existing health ontologies fall short of this goal, Wikidata presents a viable, scalable, and adaptable alternative that can integrate the needs of both clinical and non-clinical domains. However, implementing WikiData or a similar solution requires establishing a **strong governance and maintenance**

**structure**, involving **active collaboration with health data communities**, and developing **tools and support systems** to facilitate widespread adoption.

## 8.5 Controlled Vocabularies

In a multilingual and multidisciplinary environment like the EHDS, spanning 24 official EU languages and a wide spectrum of data holders, it is crucial to maintain a common set of terms with shared definitions. Controlled vocabularies serve this purpose by:

- Ensuring terminological consistency: Each concept is described using a single, agreed-upon term, avoiding ambiguity and duplication.
- Supporting machine-readability: Vocabularies provide unique identifiers for each term, enabling automated tools to process and interpret metadata unambiguously.
- Enabling cross-border and cross-domain interoperability: Consistent use of terms facilitates dataset discovery, comparison, and reuse across Member States and sectors.

In HealthDCAT-AP, the metadata model proposed for dataset descriptions in EHDS, several properties can or should use controlled vocabularies instead of free-text values. These include properties related to dataset themes, formats, access rights, and provenance, among others. Controlled vocabularies are structured terminological resources that provide unique identifiers and definitions for each concept. They are structured collections of concepts from a specific domain, each with a unique identifier and clear definition and, optionally, a set of synonyms and translations. This ensures a shared understanding of what each term represents, which is essential for accurate and reliable metadata.

### 8.5.1 Controlled vocabulary for Was Generated By property

Within the context of TEHDAS2 and the Technical Working Groups supporting the EHDS implementation, the need for **controlled vocabularies for certain metadata properties** was discussed and recognised. A particular focus was placed on the `wasGeneratedBy` property in HealthDCAT-AP.

This property is used to indicate the activity responsible for generating a dataset. Having a controlled vocabulary for this property would significantly enhance interoperability within the EHDS by ensuring that datasets referring to the same kind of origin are consistently annotated, besides facilitating dataset discovery, filtering, and aggregation based on activity types and enabling structured links between dataset purpose and the types of data generated, in line with Article 51 of the EHDS Regulation.

During the Technical Working Groups, a **collaborative mapping exercise** was carried out to develop a first version of such a controlled vocabulary. This involved:

- Collecting feedback through surveys from various domain experts and data holders
- Identifying typical activities that generate health data in different domains
- Mapping each activity to the relevant categories of health data listed in Article 51

The result is a preliminary controlled vocabulary of activities, each associated with one or more types of data it can generate. This vocabulary is presented in Table 28.

**Note:** The future applicability of this preliminary vocabulary will depend on its validation by competent authorities and its potential alignment with existing vocabularies maintained by EU bodies (e.g. Eurostat, EMA, ECDC). Further refinement may be necessary to ensure consistency with sectoral standards and to support its integration into future EHDS governance and implementation frameworks.

**Table 28.** Controlled vocabulary for wasGeneratedBy property, gathered from the input of the technical working group.

| Activity | Description of the data-generating activity | Relevant EHDS Categories |
|---|---|---|
| **Administrative processes** | Generation of data during healthcare system operations, e.g. appointments, coding, payments. | e, c |
| **Infectious disease monitoring** | Surveillance and reporting of disease outbreaks and infections. | d, b |
| **Prescribing or dispensing medicines** | Issuing or providing medication to patients by healthcare professionals or pharmacies. | a, e |
| **Automatically generated** | Data passively collected by sensors, apps, or devices without active input from users. | h, i |
| **Laboratory tests** | Data produced by clinical or diagnostic testing in a lab setting. | a, d, f |
| **Probability survey** | Data collected through a statistically representative sampling process. | b, p |
| **Biobank/sample collection** | Acquisition and storage of biological samples from individuals for current or future research. | f, g, q |
| **Measurements** | Collection of physiological, anthropometric, or other health indicators (e.g., blood pressure, BMI). | a, f, h |
| **PROM (Patient-Reported Outcome Measures)** | Self-reported data provided by patients on their health status and quality of life. | a, p |
| **Census data** | Large-scale demographic and socio-economic data collection conducted by public authorities. | b, c |
| **Use of medical devices** (updated) | Clinical or home use of health technologies that generate or record data during patient care. | h, a, n |
| **Public Health Surveillance** | Systematic collection and analysis of health data to inform public health action. | b, d, k |
| **Claims, insurances and reimbursement** | Data generated through processing health-related financial transactions and coverage validation. | c, e |
| **Medical registry** | Structured, ongoing data collection on a specific disease, procedure, or treatment population. | l, o, k |
| **Quality Registry** | Collection of healthcare quality indicators to monitor care delivery and outcomes. | j, l |
| **Clinical trial** | Structured, controlled investigation into medical treatments, interventions or technologies. | m, f, g |
| **Models and simulations** | Use of computational or statistical models to simulate clinical or population outcomes. | c, m |
| **Geospatial monitoring** | Data collection using satellite or location-based systems to monitor environmental or health exposures. | b, d |
| **Cohort** | Longitudinal tracking of a defined group over time to observe health outcomes. | p, q, f |
| **Municipal health data repository** (updated) | Centralised local government data collection on public health and services. | b, c, e |
| **Research project** | Targeted scientific activity to answer specific health-related research questions. | m, p, q |
| **Dedicated research database** | Purpose-built repository developed as part of a research activity for future reuse. | q, m, p |

| | | |
|---|---|---|
| **National Health Registries** | Government-led databases tracking diseases, treatments, or health trends at national level. | k, o, l |
| **Routine records (non-health)** | Administrative or system-generated records not primarily intended for health, but used secondarily. | b, c |
| **eHealth application** | Data created through interaction with digital health apps by individuals. | i, h |
| **National Medical Quality Registries** | Nationally coordinated registries collecting data on treatment quality, outcomes, and standards. | j, l |
| **Sample collections** | Systematic gathering of human material (e.g. blood, DNA) for storage and analysis. | f, g, q |
| **Non-medical application** | Data collected through platforms or tools not designed for medical use (e.g., fitness apps). | i, b |
| **Surveillance** | Ongoing, systematic monitoring of public or population health indicators. | d, b, c |
| **Health survey** | Collection of structured self-reported data about health behaviours, symptoms, and determinants. | b, p |
| **Observational Data** | Data generated by non-interventional observation of patients or populations. | m, p, q |
| **Healthcare visit** | Patient interaction with a healthcare professional for diagnosis, treatment, or consultation. | a, e, j |
| **Hospital records database** | Institutional system aggregating clinical and operational hospital data. | a, e, h |
| **Patient admission, care and discharge** | Process of hospitalizing, treating, and releasing a patient, generating clinical and administrative data. | a, e, j |

While already quite detailed, this list is intended as a starting point. Further validation and refinement by domain-specific communities, such as clinical data holders, public health authorities, and environmental data providers, is required to ensure that it offers comprehensive coverage across the 17 categories of health data, maintains consistency and an appropriate level of granularity for the diverse use cases within the EHDS, and remains practically applicable for data holders with varying levels of metadata expertise.

### 8.5.2 Controlled vocabulary for Publisher Type property

In addition to the controlled vocabulary for the generated by property, a preliminary controlled vocabulary was also developed for the publisher type property. The goal of this vocabulary is to provide an harmonised list of organization types responsible for making datasets described under Article 51 available. This contributes to greater consistency and clarity when describing the source of datasets across different Member States and sectors. The common vocabulary suggested by the technical working group, based on input from participants and analysis of existing practices, is presented in Table 29.

**Table 29.** Controlled vocabulary for Publisher Type property, gathered from the input from the technical working group.

| Publisher type group | Value |
|---|---|
| **Government and Public Sector Organizations** | Administrative institution |
| | National authority |
| | Other government agency |
| | Public health institute |
| | Public health organisation |

| | |
|---|---|
| | Public health registry |
| | Regional authority |
| | Pathology registry |
| | Quality registry |
| | Statistics agency |
| **Research and Academic organizations** | Biobank |
| | National Cancer Institute |
| | Other public institutes (not strictly only dedicated to health) that collect health data |
| | Research institute/organisation |
| | Research Infrastructures |
| | University |
| **Healthcare Providers** | Inpatient institution/Hospital |
| | Mental health organisation |
| | Municipality or other area |
| | Outpatient institution |
| | Primary care organisation |
| **Private Sector Entities** | Health insurance company/organisation |
| | Laboratory |
| | Other type of company that somehow collects health data |
| | Health app/technology manufacturer |
| | Pharmaceutical companies |
| | Pharmacy |
| | Private company |
| | Private health insurance |
| | Software manufacturer |
| **Non-Governmental Organizations** | Not-for profit organisation |

# 9 Considerations for implementation

## 9.1 Need for tools, training and support

According to Article 60 of the EHDS Regulation, health data holders are responsible for communicating to the Health Data Access Body (HDAB) a description of the datasets they hold, in line with the requirements to be defined in forthcoming implementing acts. These acts are expected to formally establish HealthDCAT-AP as the common metadata model for describing datasets intended for secondary use under the EHDS, subject to final decisions by the European Commission and the EHDS Regulatory Committee. Throughout the activities conducted in the TEHDAS2 project, and in dialogue with stakeholders and data holders across Member States, it has become clear that this obligation represents a significant technical, organisational, and operational challenge, particularly for data holders with limited prior experience in structured metadata or semantic technologies. Although the

responsibility lies with data holders, the successful implementation of secondary use under the EHDS will require broad, coordinated support at both national and European levels. This includes strong collaboration between governments, policymakers, HDABs, data holders, and data users. To ensure a harmonised and inclusive adoption of HealthDCAT-AP, the TEHDAS2 community has identified an urgent need for centrally coordinated tools, training programmes, and support mechanisms, adapted to the diverse needs and capacities of the actors involved.

- **Data holders** must receive **targeted training** to understand the EHDS legal framework, their specific obligations, and the architecture of the ecosystem for secondary use of health data. Training initiatives must be tailored to the wide spectrum of data holders covered by the EHDS Regulation, both in terms of the different types of health data involved, and the highly different levels of digital and metadata maturity across the EU health data landscape. In addition to training, data holders will require **intuitive, user-friendly tools** to support the generation of HealthDCAT-AP-compliant metadata records. These tools should be promoted or provided at the European level to ensure consistency in metadata creation processes and avoid fragmentation in documentation practices, interoperability gaps, or divergent national approaches. In parallel with European-level efforts, several initiatives are emerging, and will continue to emerge, at Member State level to support national implementation of the EHDS. These may include the development of dedicated portals or platforms offering legal, technical, and procedural guidance to data holders, particularly regarding metadata creation and the application of HealthDCAT-AP. For example, the Health Information Portal [30]developed by Sciensano (Belgium) illustrates a national initiative aimed at supporting data holders through accessible documentation and resources. While such tools are not developed or endorsed at EU level, national knowledge-sharing platforms will play an important role in complementing European efforts, supporting consistent training and guidance across Member States, and facilitating the effective participation of data holders in the EHDS ecosystem

- **HDABs** must also be trained to effectively manage national dataset catalogues based on HealthDCAT-AP. They will play a critical role in guiding and assisting data holders, especially during the initial phases of implementation. HDABs are also responsible for validating metadata records for compliance with the minimum mandatory elements. While the long-term goal is to automate compliance checks using tools such as the DCAT-AP validator[31], the TEHDAS2 community emphasises that in the early stages of implementation, **HDABs should prioritise support over rejection**. In other words, instead of rejecting non-compliant metadata records outright, HDABs should provide feedback and assistance to enable data holders to correct and complete their descriptions.

Without such tailored support mechanisms, there is a real risk that a significant number of data holders may struggle to fulfil their obligations, potentially undermining the establishment of a comprehensive and high-quality metadata catalogue across the EHDS.

[30] HealthDCAT AP | European Health Information Portal
[31] HealthDataEU · GitLab

## 9.2 Addressing data/organization-specific concerns

Feedback received through the TEHDAS2 public consultation and activities has highlighted a range of data-type-specific and organisation-specific concerns regarding the implementation of HealthDCAT-AP. These concerns reflect the inherent challenge of applying a general-purpose metadata model, which is designed for broad applicability, across the highly diverse and complex health data ecosystem of the EHDS.

Importantly, addressing these concerns should not be solely the responsability of individual data holders. Decisions on how to interpret and apply the common metadata model in specific contexts should be guided by further clarification at national or EU level, to ensure consistency and alignment with the overall EHDS framework.

### 9.2.1 Data-Type Specific Considerations:

HealthDCAT-AP is not intended to replace existing domain-specific metadata standards or formats. Many data types listed in Article 51 of the EHDS Regulation, such as clinical trials, genomic data, or registries, are already supported by mature, community-defined standards. Rather than replacing these, HealthDCAT-AP functions as a high-level, harmonised metadata model to enable consistent dataset discovery and catalogue integration across the European Health Data Space. To support this approach, there is a clear need for official mappings between domain-specific standards and HealthDCAT-AP. These mappings should be developed collaboratively with key communities (e.g. clinical trials networks, genomics consortia) and should help ensure alignment without sacrificing the specificity and rigour of existing practices. Such cooperation will be critical for achieving semantic interoperability and a coherent implementation of the metadata model across Member States.

### 9.2.2 Organisation-Specific Considerations:

In addition to technical concerns, many organisations raised questions related to how to apply HealthDCAT-AP in the context of their specific data architecture and internal processes. Health data is organised in a wide variety of ways, often differing not only from organisation to organisation, but also across Member States. Some organisations already have well-established methods for structuring and cataloguing datasets, and will need to evaluate how these can be adapted to comply with EHDS requirements. A recurring point of discussion within the TEHDAS2 community has been the definition of a "dataset" under the EHDS framework. For data providers managing large, integrated data warehouses, such as national health registries, biobanks, or public health agencies, deciding how to meaningfully define and break down their data into coherent datasets for secondary use is particularly challenging. Before creating HealthDCAT-AP records, these organisations will need to assess how their data can be structured and segmented for inclusion in the dataset catalogue. Two guiding considerations should inform this process:

- **Dataset granularity and discoverability**: Dataset descriptions should be created at a level that is meaningful for potential users. For example, in the case of cohort datasets with multiple collection events or subpopulations, it is often more useful to define datasets at the level of a collection event. A practical principle is to organise datasets around a common aspect, such as population, collection period, or data collection method, and then use HealthDCAT-AP's relation properties to link related datasets and provide an overview of their structure.

- **Feasibility of data provision and legal compliance**: Under Article 60, data holders are expected to provide datasets to the secure processing environment (SPE) within a defined timeframe (typically three months). Overly large or complex datasets, such as entire data warehouses or national registries, may hinder timely and legally compliant responses to access requests. Additionally, data minimisation principles must be respected, making it crucial to define datasets in a way that ensures both targeted use and manageable delivery. Further guidance on how to address these concerns can be found in the HealthData@EU Pilot Deliverable 6.2 *HealthDCAT-AP – A DCAT Application Profile for the description of health datasets "Recommendations on further development and deployment for possible EU-wide uptak.* Link

## 9.3 Handling of IP-protected information in the catalogues

The protection of IP rights, trade secrets, and other legal restrictions related to metadata was one of the recurring concerns raised during the TEHDAS2 public consultation and stakeholder engagement activities. The balance between transparency and intellectual property protection is a broader issue that cannot be fully resolved at the metadata level alone and will require further guidance at policy and implementation levels. However, this topic was explored in the context of TEHDAS2 activities, particularly with regard to how metadata can support transparency while respecting intellectual property constraints.

According to Article 52 of the EHDS Regulation, health data holders are required to inform the HDAB of any electronic health data containing content or information that is protected by intellectual property rights, trade secrets, or subject to regulatory data protection. This obligation includes identifying the specific parts of the dataset that are affected and providing a justification for the need for protection. This information must be communicated either when the dataset description is submitted to the HDAB or, at the latest, following a specific request from the HDAB. To support this requirement, the HealthDCAT-AP metadata model has the mandatory property dct:rights, that enables data holders to include a rights statement describing any legal conditions or limitations on the use of the dataset or its distributions. It is a critical mechanism to allow compliance with the obligations under Article 52, including in cases where parts of the data are protected by IPR or similar restrictions. This requirement also applies to sample distributions. In cases where a dataset includes information protected by intellectual property rights that cannot be shared in a sample distribution, data holders are encouraged to provide an alternative form of metadata that still supports discoverability. One practical solution is to include a redacted data dictionary within the sample distribution, indicating which parts of the dataset are affected by IPRs and clearly stating the justification for their restricted status. This enables potential users to understand the nature and scope of the dataset while respecting legal limitations.

Once this information is made available, it becomes the responsibility of the HDABs to take appropriate action. Article 52 establishes that HDABs must adopt all specific, appropriate, and proportionate measures, whether legal, organisational, or technical, to safeguard intellectual property rights, trade secrets, or regulatory protections. These measures may vary depending on the sensitivity of the information and the nature of the dataset, but they are fundamental to building trust with data holders and ensuring lawful and secure secondary use of health data under the EHDS.

## 9.4 Relation with quality and utility label (QUANTUM)

The provision of a data quality and utility label by health data holders is foreseen in Article 78 of the EHDS Regulation. The QUANTUM project (Grant Agreement No. 101137057, https://quantumproject.eu/) is currently developing such label, dependent on future Commission and regulatory committee decisions. This topic falls outside the scope of this guideline. However, in response to several questions raised during the TEHDAS2 Public Consultation, some clarifications are provided below.

A key point to understand is that the data quality and utility label and HealthDCAT-AP serve different, but complementary, purposes:

- The HealthDCAT-AP metadata model is focused on the **cataloguing and description of datasets** to enable discoverability, interoperability, and reuse across the EHDS.
- The **QUANTUM label**, by contrast, aims to provide an evaluation of the dataset through defined data-level metrics, such as completeness, accuracy, timeliness, and utility for specific use cases.

This distinction is essential: HealthDCAT-AP describes metadata, while QUANTUM assesses data quality and fitness for use. They are not interchangeable but are both necessary components in enabling high-quality, trusted data reuse.

For each dataset intended for secondary use, both the HealthDCAT-AP metadata record and the data quality and utility label will need to be provided by the data holder. While they operate independently, certain elements, such as data provenance or variable-level descriptions provided in data dictionaries, may be relevant to both. However, there is currently no automated information exchange mechanism between the two.

To facilitate linkage between these two layers of information, HealthDCAT-AP includes a dedicated property for this purpose: qualityAnnotation. This property is being refined in the context of the QUANTUM project and will be used to reference the quality and utility label associated with a dataset. The definition and practical implementation of this link is an ongoing effort under a collaborative process between the QUANTUM project and TEHDAS2.

Further guidance and updates on how to harmonize HealthDCAT-AP metadata with the QUANTUM label will be made available as this work progresses in QUANTUM project.

# 10 Annex 1

## Summary of feedback through public consultations

A draft version of this document was in public consultation in January 2025. This document was commented in total for 110 times. The number of responses may contain some duplicates as there was no individual identification and verification required to respond to the surveys. Some respondents have also responded both from data holder's and data user's perspective. The responses came from 13 different countries from the EU countries and the European Economic Area countries. Responses from Slovenia, Romania, Lithuania, Greece and Croatia and international organisations were largely missing. The respondents were primarily from three main types of organisations, listed in order of prevalence: Public organizations, academic or research organisations and private organisations.

As part of the public consultation, we asked for input on how the HealthDCAT-AP could be tailored to better support health datasets, which aspects are most challenging to describe, perceived strengths and weaknesses, implementation challenges, needed improvements for clarity and usefulness, and the kind of additional support that would facilitate adoption. Additionally, the survey was used to collect input on the cardinalities for each property discussed. The results are presented in Figures 2 - 4.

It was highlighted that metadata coverage should be expanded, especially at the variable level, and that structured fields should be preferred over free text to enhance consistency. Stronger alignment with established health data standards, clearer distinctions between datasets and catalogues, and better handling of data versioning were also mentioned. To address this, the final guideline introduces detailed guidance and examples for generating variable-level metadata (e.g. using CSVW), strengthens references to health data standards, and adds a new chapter explaining metadata fundamentals, including distinctions between data, metadata, datasets, and catalogues. Versioning is now better explained, and usage notes for each property offer practical, clearer guidance. Additionally, several property-specific and dataset-specific challenges were raised. These were addressed through improved, tailored guidance within the property descriptions. Challenges that were too specific to address in a general guideline are acknowledged, with future collaboration with data holders and HDABs encouraged – this approach is described in the final chapter focused on considerations for implementation.

It was identified that HealthDCAT-AP's strengths include comprehensiveness, flexibility, and machine-readability. However, weaknesses such as insufficient support for column-level metadata and technical complexity were noted. In response, an introductory chapter now explains the purpose and foundation of HealthDCAT-AP, including an overview of RDF and DCAT-AP. This section also links to e-learning resources and provides context for variable-level metadata and regulatory documentation. Implementation concerns were also raised, including the lack of harmonised data dictionaries, handling of IP and legal restrictions, and the resource burden of metadata maintenance. While implementation is beyond the scope of the guideline, the final chapter offers preliminary proposals such as new dictionaries and clarifies how to indicate IP and legal considerations using HealthDCAT-AP. To improve clarity, it was requested that the guideline offer clearer terminology, better support for machine-readability, step-by-step instructions, and practical examples, particularly with controlled vocabularies. These needs were met by restructuring the document, adding an introductory chapter, and providing detailed, user-friendly guidance for each property, including improved legal and governance context. Finally, the need for additional support, such as documentation, training, and tools for metadata generation, was largely emphasized. While addressing this feedback is beyond the scope of the current guideline, it is essential for understanding the needs of data holders and shaping future

efforts and developments. These insights have been discussed within the TEHDAS2 community and are reflected in the final chapter, which provides considerations for implementation and outlines potential future work that responds to the needs identified through the consultation.

This feedback significantly informed the final version of the guideline and will guide future implementation efforts under the EHDS.

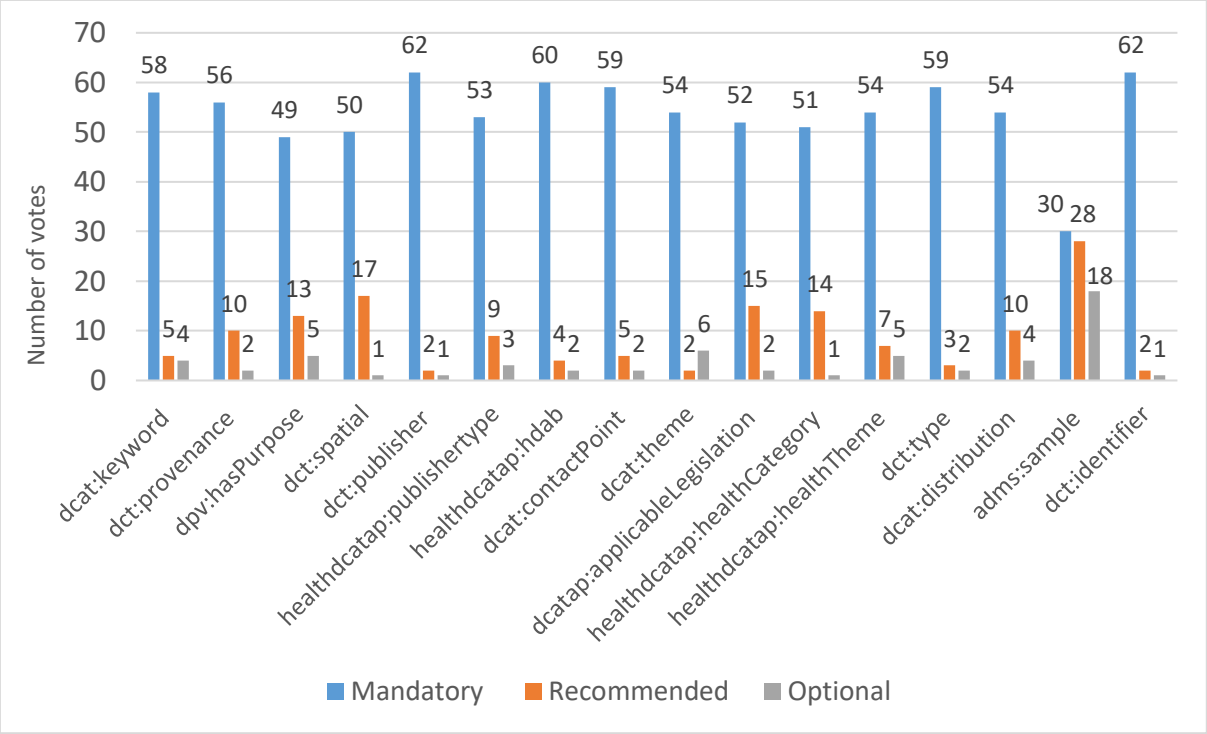


**Figure 2.** Results from Public Consultation regarding HealthDCAT-AP mandatory properties. For all proposed mandatory properties, the large majority of responses supported their mandatory status. The only exception was the adms:sample property, which generated a more divided opinion, with 30 votes in favour of making it mandatory, 28 votes for recommended and 18 votes suggesting it should remain optional. This was addressed in the content of the deliverable.

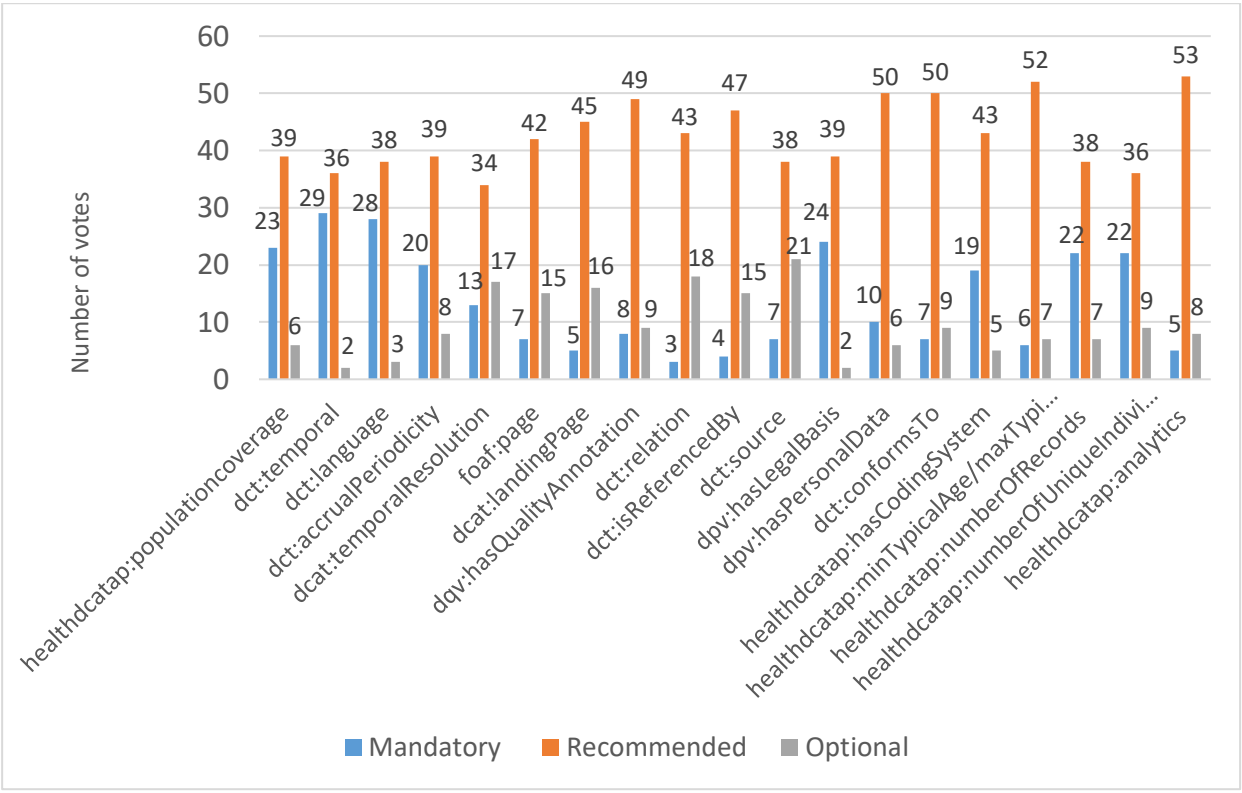


**Figure 3.** Results from Public Consultation regarding HealthDCAT-AP recommended properties. For all proposed recommended properties, the vast majority of responses agreed that they should be classified as recommended.

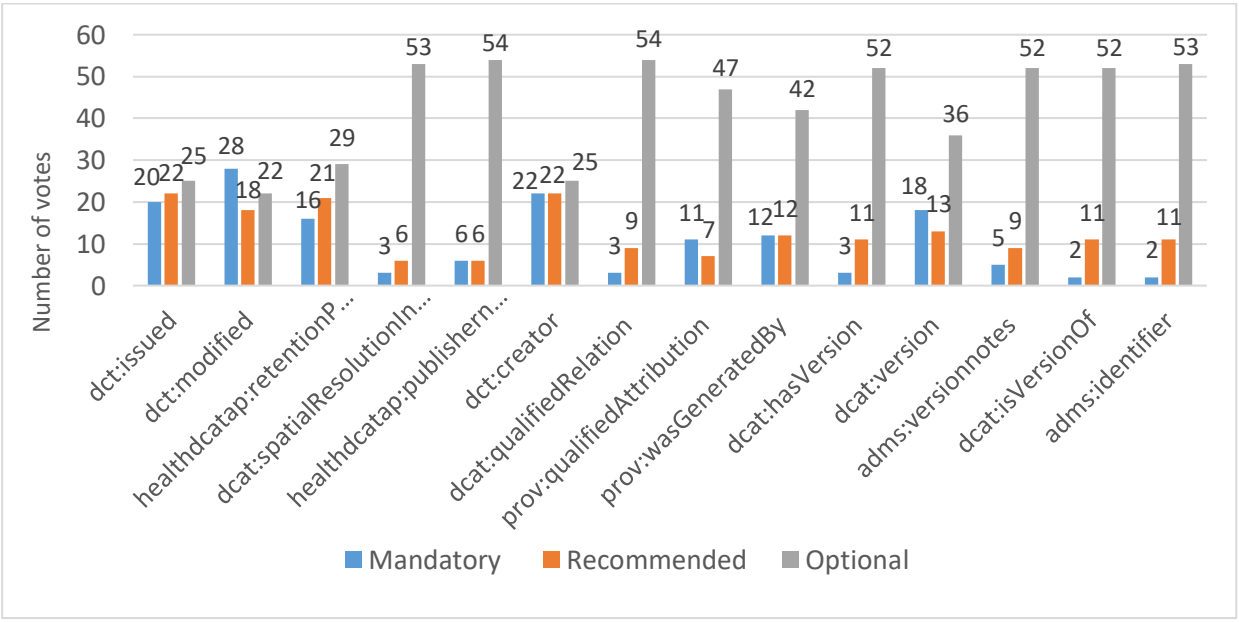


**Figure 4.** Results from Public Consultation regarding HealthDCAT-AP optional properties. Some optional properties received more divided responses, particularly dct:issued, dct:creator, and healthdcatap:retentionPeriod, though the majority still supported their optional status. However, for dct:modified, most respondents indicated that it should be mandatory. The role of this property was explored in detail in the deliverable.