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"

Pascal Derycke, Truls Korsgaard, Charles Andrew Vande Catsyne, Hans Aage Huru, Nienke Schutte
30 September 2024

Co-funded by
the European Union

Disclaimer

The content of this deliverable represents the views of the author(s) only and is his/her/their sole responsibility; it cannot be considered to reflect the views of the European Commission and/or the Consumers, Health, Agriculture and Food Executive Agency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use of its contents.

Copyright Notice

Contents

1. Executive summary

This document presents recommendations for the further development and potential EU-wide adoption of HealthDCAT-AP, an extension of the DCAT Application Profile for dataset catalogues in Europe[1]. HealthDCAT-AP is specifically designed to describe health datasets and dataset access services, ensuring they are consistently represented and easily discoverable across various platforms.

The principal objective of the Health Data Catalog Application Profile (HealthDCAT-AP) is to establish a standardised, interoperable metadata schema tailored to the health domain. This schema is intended to facilitate the discovery, sharing, and reuse of health datasets across the European Union. By aligning with the broader DCAT-AP (Data Catalog Vocabulary Application Profile) standard, HealthDCAT-AP ensures that health-related datasets are uniformly described, promoting interoperability and enabling seamless data exchange among researchers, public health institutions, policymakers, and other stakeholders.

HealthDCAT-AP introduced specific extensions to the DCAT-AP model to meet the unique requirements of the health sector. This standardisation is crucial for the European Health Data Space (EHDS) framework, enabling effective health data sharing across Europe. By enhancing the accessibility and availability of health data, HealthDCAT-AP plays an essential role in helping the EHDS achieve its objectives of improving public health outcomes, advancing research, and informing policy-making across the EU, all while upholding stringent data protection and privacy standards.

This document is the second and final deliverable of Work Package 6 of the EHDS2 pilot project consortium. The first deliverable, the draft specification of HealthDCAT-AP, is available at HealthDCAT-AP Draft Specification. The development of HealthDCAT-AP will continue transparently and publicly within the framework of the Second Joint Action Towards the European Health Data Space (TEHDAS2). TEHDAS2 lays the groundwork for the harmonised implementation of secondary health data use within the European Health Data Space (EHDS), advancing the EU's vision of a connected and interoperable health data ecosystem.

2. Introduction

2.1. Background

Background on DCAT-AP and its relevance to the EHDS

The EHDS Regulation mandates health data holders to create and manage high-quality metadata records for their datasets.

EHDS Regulation Article 60) Duties of health data holders

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

---

[1] DCAT Application Profile for dataset catalogues in Europe

These records are crucial for effective data discovery, which is supported by dataset catalogues designed as Web platforms. These platforms must expose metadata records and ensure that their data discovery systems can perform searches ranging from simple full-text queries to more complex faceted search or linked data queries. The governance of these datasets catalogues within the EHDS is overseen by Health Data Access Bodies, making the support for a seamless data discovery experience essential.

DCAT-AP (Data Catalog Vocabulary - Application Profile) is a metadata standard designed to describe datasets in a way that facilitates their discovery, accessibility, and interoperability. By providing a structured common vocabulary - a lingua franca - for the dataset catalogues, DCAT-AP *ensures that datasets can be easily shared and reused across various platforms. It* enables data users to search for datasets efficiently, understand their content and purpose and ensure their reuse.

Enhancing Findability and Reuse according to the FAIR data principles

DCAT-AP supports the implementation of the FAIR data principles. The FAIR data principles stand for Findability, Accessibility, Interoperability, and Reusability, which are guidelines to ensure that data and metadata are FAIR[2]. The EHDS Regulation endorses the FAIR data principles to ensure health data sharing and responsible reuse across the EU. The Regulation lays the foundation for an ecosystem of federated national dataset catalogues alongside a central catalogue.

[EHDS Regulation Article 79) EU Dataset Catalogue](#)

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

The EHDS also provides the necessary metadata properties that align with what a data user might search for. What is the dataset about? Who created the dataset? Who is responsible for it? Who can access the dataset, and under what conditions? When was the dataset published? Where was the dataset collected? Where can the dataset be accessed or downloaded? Why was the dataset created? How is the dataset structured (e.g., data model, file format, schema)? Etc. For instance, DCAT-AP includes essential metadata properties such as dataset types (e.g., geospatial or statistical data), access conditions (e.g., open data or protected data), and dataset themes, ensuring efficient categorisation and discoverability.

[EHDS Regulation Article 80) Minimum specifications for datasets of high impact](#)

The Commission may, by means of implementing acts, determine the minimum specifications for datasets of high impact for secondary use, taking into account existing Union infrastructures, standards, guidelines and recommendations.

Interoperability and machine readability of DCAT-AP: DCAT-AP is inherently designed to support interoperability, machine readability, and machine actionability, using the RDF (Resource Description Framework) to enable seamless interaction between health data

---

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

platforms and AI applications. This dual nature allows datasets to be more accessible not only to people but also to machines such as AI applications, provided that dataset catalogues offer the necessary interfaces and care about interoperability for machine interaction.

Its linked data approach (RDF-based structure), coupled with the use of persistent identifiers (HTTP URIs), promotes the concept of openness, enables the creation of interconnected knowledge graphs, which enhance the discoverability of datasets and reduce the need for duplicated descriptions. Cross-domain compatibility, federation, and semantic interoperability ensure that metadata records are harvestable, indexable, and manageable across federated catalogues. They reduce the need to duplicate descriptions, as they encourage the creation of global knowledge graphs pointing to single sources of truth. The standard also includes rules for managing duplication of metadata records across federated catalogues. By leveraging open controlled vocabularies to describe concepts and relations, DCAT-AP models information in a way that machines can process as both readable and actionable data. This capability extends the semantic web, increasing the discoverability and reuse of datasets catalogued with DCAT-AP.

DCAT-AP in the EU Data Spaces: DCAT-AP helps ensure that metadata standards across EU data spaces are aligned, which fosters the seamless sharing of data across borders and domains. This is essential to the goals of the European Data Strategy[3], which aims to create a single European data market.

Example of the interconnection between EOSC EU Node **and EHDS:** The DataCite Metadata Schema, utilised within the European Open Science Cloud (EOSC), offers a standardised framework for describing and cataloguing research data, publications, and other scholarly outputs. This standardisation is vital for ensuring the discoverability, accessibility, and reuse of research outputs, thereby supporting open science and promoting a connected and interoperable research environment across Europe. The envisioned alignment of DCAT-AP and DataCite standards within the RDF framework is intended to facilitate the interconnection between the European Health Data Space (EHDS) and the EOSC through DataCite-to-DCAT-AP mapping. While DCAT-AP is used to catalogue datasets that can be employed in research, the DataCite schema provides a detailed description of digital objects generated by research and assigns Persistent Identifiers (PIDs) for their citation. These PIDs can be seamlessly integrated into DCAT catalogues, thereby enhancing both the discoverability and understanding of datasets. For instance, this interconnection enables machines to efficiently list all research outputs associated with a specific dataset. This is particularly beneficial when handling sensitive health data that cannot be shared directly, as it allows researchers to discover 'proxy' information related to the sensitive data. This approach enhances researchers' understanding of the dataset's relevance and potential applicability in their studies.[4]

The interconnection of DCAT-AP and DataCite standards enhances the interoperability and findability of datasets, reinforcing the connection to FAIR data principles and supporting open science and health data reuse across Europe.

---

[3] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS A European strategy for data COM/2020/66 final
[4] Insights on DCAT-AP and this topic were presented in the oral presentation "Working with the Future European metadata catalogue" at the EUPHA conference in November 2023, which is available in Annex 2.

The governance of these dataset catalogues in scope of the EHDS is overseen by **Health Data Access Bodies**, in line with the **EHDS regulation**. These bodies are responsible for ensuring that metadata records meet high-quality standards for data discovery."
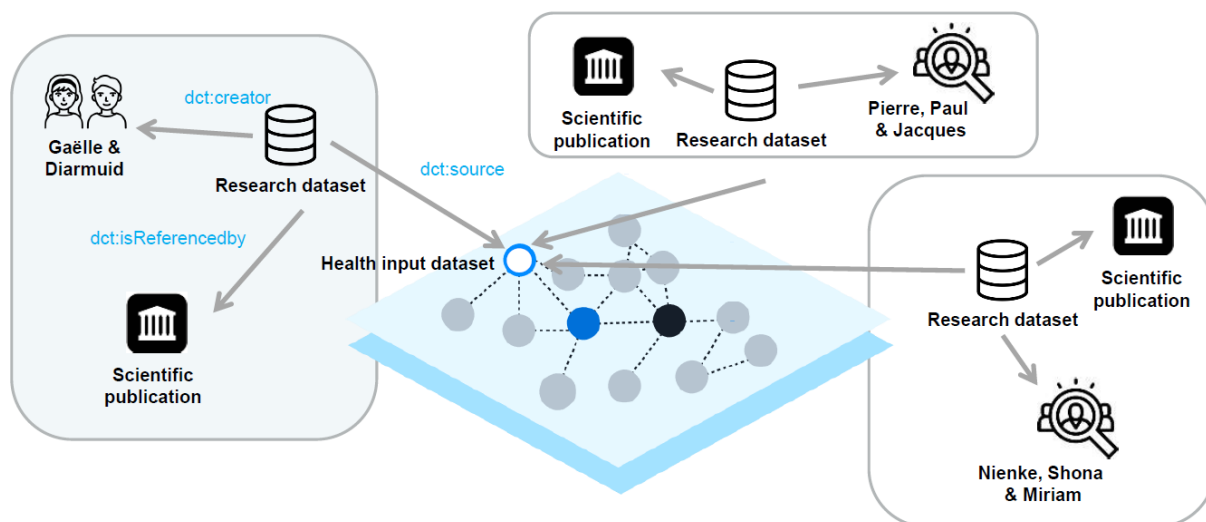
"European Health Data Space – Working with the future European metadata catalogue" - EUPHA conference 9 November 2023, 9:00-10:00, Dublin

2.2. Objectives of HealthDCAT-AP

The primary objectives of HealthDCAT-AP are to extend the DCAT-AP standard to better support the discovery, accessibility, and interoperability of health-related datasets within the European Health Data Space (EHDS). HealthDCAT-AP aims to standardise the minimum metadata elements and characteristics that health data holders are required to provide for describing their datasets, as specified under the EHDS Regulation. By doing so, it ensures that both non-sensitive and sensitive health data can be shared responsibly across the EU, enhancing the discoverability and usability of health data for secondary purposes, such as research, public health policy, and innovation.

HealthDCAT-AP extends the DCAT-AP framework by integrating additional health-specific metadata elements, properties, and guidelines tailored to the unique requirements of health data management and sharing. This extension addresses the specific challenges of managing, sharing, and discovering health-related datasets within the EHDS while ensuring compliance with EU data protection regulations, including the General Data Protection Regulation (GDPR)[5] and EHDS Regulation. Additionally, HealthDCAT-AP remains fully compatible with the broader DCAT-AP framework used across multiple domains in the EU Data Spaces.

---

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

2.3. Use cases

2.3.1. HealthData@EU: A Federated Infrastructure for Health Dataset Catalogues within EU Data Spaces

2.3.1.1 Legal Context for Health Data Catalogues

Over the past decade, the state of dataset catalogues in Europe has evolved significantly, driven by various EU legal acts aimed at improving data sharing, transparency, and interoperability across the European Union. Today, the overarching European Data Portal https://data.europa.eu includes nearly 2 million datasets, of which nearly 27,000 records are labelled as health-related datasets. This ecosystem of dataset catalogues is constantly evolving, fuelled by the continuous publication of legal acts:

- The Open Data Directive (Directive (EU) 2019/1024)[6] pushes for public sector data to be made easily accessible and reusable, encouraging the creation of open dataset catalogues across EU countries.

- The INSPIRE Directive (Directive 2007/2/EC)[7] plays a crucial role in developing standardised and interoperable spatial dataset catalogues, enabling better sharing of environmental information across Europe.

- Meanwhile, the Data Governance Act (DGA, Regulation (EU) 2022/868)[8], effective from September 2023, promotes secure data sharing by introducing data intermediaries and data altruism, further connecting dataset catalogues across different sectors.

- The Data Act (DA, Regulation (EU) 2023/2854)[9] is expected to boost this progress by setting clear rules for data access and use, particularly in business contexts, and by promoting data interoperability. This will further drive the development and federation of dataset catalogues.

- Additionally, the European Interoperability Framework (EIF)[10] provides guidelines to ensure that dataset catalogues from different countries and sectors can work together seamlessly.

These legal acts collectively support Europe's strategy to build a secure and innovative data economy, with dataset catalogues playing a central role in this digital landscape. The system of interoperable dataset catalogues, through metadata, serves as a general proxy to data.

As part of this broader strategy, the EU is also developing a broad range of European Data Spaces - sector-specific environments where data can be shared and accessed securely

---

[6] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast)

[7] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)

[8] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act)

[9] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act)

[10] ISA² - Interoperability solutions for public administrations, businesses and citizens

across borders and industries. These data spaces will aim to bring together data from various sources, making it easier for businesses, governments, and researchers to collaborate and innovate. The creation and integration of dataset catalogues within these spaces are essential for ensuring that data is accessible, interoperable, and valuable across the entire European Union.

Recognising the value of data has been a long-standing effort in data management, embodied by the EMODnet (European Marine Observation and Data Network) motto: 'Create once, use many times.'"

EMODnet principles:

- Collect data once and use it many times
- Develop standards across disciplines as well as within them
- Process and validate data at different levels. Structures are already developing at national level but infrastructure at sea-basin and European level is needed
- Provide sustainable financing at an EU level so as to extract maximum value from the efforts of individual Member States
- Build on existing efforts where data communities have already organised themselves
- Develop a decision-making process for priorities that is user-driven
- Accompany data with statements on ownership, accuracy and precision, and
  Recognise that marine data is a public good and discourage cost-recovery pricing from public bodies.

The EHDS Regulation adopts principles similar to **EMODnet**, focusing on collecting data once, reusing it many times, and applying cross-disciplinary standards to ensure maximum value from health datasets across the EU. While these principles provide a common foundation for EU data spaces, each thematic data space will have to establish its own governance framework to ensure effective stakeholder engagement. This approach is particularly crucial for health data, as its sensitive nature demands careful curation and reuse under the guidelines set by the GDPR and national laws of member states.

2.3.1.2 Role of the Health Data Access Bodies of the EHDS

Chapter 4 of the European Health Data Space (EHDS) Regulation is as this regard is a valid example as it focuses on the governance and rules for accessing and using health data within the EHDS. This chapter outlines the requirements for the establishment of Health Data Access Bodies in each EU member state. These bodies are responsible for granting access to health data for secondary use, such as research, policy-making, and innovation, while ensuring that data protection and privacy are strictly maintained. Chapter 4 of the EHDS Regulation outlines the conditions for health data access, emphasising data minimisation and the use of anonymisation or pseudonymization to protect individual privacy, while ensuring that health data can be reused for secondary purposes such as research, policy-making, and innovation.. Additionally, this chapter addresses the interoperability standards that must be adhered to, ensuring that health data from different countries can be effectively shared and used across the EU. As this regard, the Health Data Access Bodies are responsible for cataloguing health data in scope of Art. 33 and serve as reference national health clearinghouses.

2.3.1.3 Harmonisation across Data Spaces

In the development of the European Health Data Space (EHDS), it is recognised that while governance structures can be customised to fit health data specific requirements and improve the operational efficiency of the EU Health Data Space, certain standards must remain consistent across all EU Data Spaces. In particular, standards for interoperability or security are critical and should not be compromised. They should be defined not only for EU Health Data Space but also beyond. The common EU data spaces will form a unified, federated data infrastructure across Europe, ensuring data openness, transparency, and sovereignty. By aligning with established standards and guidelines, the implementation process for the EU data Spaces will become simpler and more efficient, as it leverages the collective expertise to implement the data spaces. This harmonisation not only simplifies the process, but also promotes collaboration and consistency, ensuring compliance with best practices and efficient use of resources.

For instance, the Article 8 (i.e.: single information point) of the Data Governance Act Regulation (EU) 2022/868[11] ensures a unified and secure approach across the European data ecosystem, preventing implementation of ad-hoc solutions and ensuring that all data spaces operate with the same level of compatibility.

[Regulation (EU) 2022/868 (DGA - Data Governance Act)](#)

> *DGA Article 8 2. « The single information point shall make available by electronic means a searchable asset list containing an overview of all available data resources including, where relevant, those data resources that are available at sectoral, regional or local information points, with relevant information describing the available data, including at least the data format and size and the conditions for their re-use. »*

The "single information points" under Article 8 of the Data Governance Act refer to the establishment by Member States of a single information point to act as the primary interface for re-users seeking to re-use data held by public sector bodies. For instance, to fulfil their reporting obligations in regard to the HVD regulation, Member States have to maintain a [High-value Dataset National single information point](#).

[EHDS Regulation Recital 86)](#)

> *The EU dataset catalogue should minimise the administrative burden for the health data holders and other database users, be user-friendly, accessible and cost-effective, connect national dataset catalogues and avoid redundant registration of datasets. Without prejudice to the requirements set out in [Regulation (EU) 2022/868](#), the EU dataset catalogue could be aligned with the data.europa.eu initiative. Interoperability should be ensured between the EU dataset catalogue, the national dataset catalogues and the dataset catalogues from European research infrastructures and other relevant data sharing infrastructures.*

[EHDS Regulation Art.57 Tasks of health data access bodies](#)

> *1. Health data access bodies shall carry out the following tasks:*
> *(j) making public, through electronic means:*
> *The national dataset catalogue referred to in point (j)(i) of this paragraph shall also be made available to single information points under Article 8 of Regulation (EU) 2022/868. [[Data Governance Act COM/2020/727 final](#)].*

---

[11] [Regulation (EU) 2022/868 (DGA - Data Governance Act)](#): The DGA applies from 24 September 2023, establishing important guidelines for the re-use within the EU of certain categories of data held by public bodies, as well as a framework for the collection and provision of data brokerage services.

According to EHDS Art. 37 and 55, the Health Data Access Bodies (HDABs) of the healthData@EU infrastructure will have to maintain similar end-points to expose health metadata records.

[EHDS Regulation Section 5 Health data Quality and utility for secondary use](#)

> *Article 77 Dataset description and dataset catalogue*
> *3. The dataset catalogue shall be made available to single information points*
> *established or designated under Article 8 of [Regulation (EU) 2022/868](#).*

The same requirement also applies to the EU Dataset catalogue (EHDS Recital 60).

[EHDS Regulation Recital 86)](#)

> *The EU dataset catalogue should minimise the administrative burden for the health*
> *data holders and other database users, be user-friendly, accessible and cost-effective,*
> *connect national dataset catalogues and avoid redundant registration of datasets.*
> *Without prejudice to the requirements set out in Regulation (EU) 2022/868, the EU*
> *dataset catalogue could be aligned with the data.europa.eu initiative. Interoperability*
> *should be ensured between the EU dataset catalogue, the national dataset*
> *catalogues and the dataset catalogues from European research infrastructures and*
> *other relevant data sharing infrastructures.*

Harvesting metadata from National single information points is a standard procedure for populating metadata records in a catalogue broker like the data.europa.eu [12].

[The European Single Access Point on data.europa.eu: Harvesting guidelines](#)

> In the framework of the Data Governance Act (DGA), the European Commission shall establish a European Single Access Point (ESAP), which will be integrated into data.europa.eu.
>
> As a searchable electronic European register, the ESAP will collect, partially mirror, and render the data provided by national single information points (NSIPs). NSIPs will assist potential re-users in finding information on what protected data (e.g., personal, or commercially confidential data) can be reused under specific conditions. They are to be established by the EU Member States by 24 September 2023.
>
> For the NSIPs' information to be successfully collected into the ESAP, their metadata will need to be structured and provided in a specific way. To facilitate the implementation of the ESAP and provide clarity on the metadata requirements, the European Commission prepared a set of guidelines for the Member States.

2.4. Roadmap

The HealthDCAT-AP specification's development began with the kick-off of the EHDS2 Pilot project in October 2022. By January 2023, the project focused on defining the requirements, starting with a comprehensive landscape analysis of existing health metadata catalogues and models, as well as an examination of existing DCAT Application Profiles, such as GeoDCAT, StatDCAT, and JRC-DCAT-AP. This phase also included the release of a DCAT-AP Sandbox catalogue, representing the "AS-IS" state. In June 2023, the project advanced by training and onboarding a Technical Working Group (TWG) of health experts, including workshops on DCAT-AP rationales provided by DIGIT/SEMIC and establishing rules for extending DCAT-AP. From September to December 2023, the focus shifted to defining the

---

[12] [European Single Access Point: Harvesting guidelines for Member States](#)

HealthDCAT-AP domain model, utilising iterative and interactive sessions alongside EU surveys to gather continuous feedback from the Technical Working Group. The draft HealthDCAT-AP specification, formatted according to the ReSpec template, was published in February 2024. The project will then enter the implementation and validation phase from March until August 2024. This phase includes the creation of a HealthDCAT-AP Sandbox catalogue (TO-BE), workshops (Task 6.3 of the work package) to support users in creating HealthDCAT-AP records, exercises in mapping to HealthDCAT-AP, development of compliant endpoints, and a functional analysis of implementing HealthDCAT-AP in the healthdata.europa.eu portal (Ref: Annex 1).

Oct. 2022 > KICK-OFF OF EHDS2 PILOT PROJECT (from inception to implementation)

Jan. 23 > DEFINITION OF THE REQUIREMENTS

• Landscape analysis of existing health metadata catalogues and health metadata models

• Analysis of existing DCAT Application Profiles: geoDCAT, statDCAT, JRC-DCAT-AP

• Release of a DCAT-AP Sandbox catalogue (AS-IS)

Jun. 23 > TRAINING AND ONBOARDING OF A TECHNICAL WORKING GROUP OF HEALTH EXPERTS

• Workshop on DCAT-AP rationales provided by DIGIT/SEMIC

• Rules for extending DCAT-AP

Sep. to Dec. 23 > DEFINITION OF THE HEALTHDCAT-AP DOMAIN MODEL

Iterative and interactive sessions and publication of surveys to collect feedback and content on a continuous basis from the Technical Working Group

Feb. 24 > DRAFT HealthDCAT-AP

• Publication of the draft HealthDCAT-AP according to the ReSpec template
 https://HealthDCAT-AP.github.io

March until August 2024 > IMPLEMENTATION & VALIDATION of HealthDCAT-AP CONCEPTS

• HealthDCAT-AP Sandbox catalogue TO-BE (http://healthdataportal.eu/)

• Workshops on creating HealthDCAT-AP records with support of tutors and a HealthDCAT-AP editor.

• Mapping to HealthDCAT-AP exercises and development of HealthDCAT-AP compliant endpoints

• Functional analysis of implementing HealthDCAT-AP in the healthdata.europa.eu portal

2.5. Methodology

The methodology for the design of the HealthDCAT-AP specification was grounded in an iterative approach, ensuring continuous refinement and alignment with stakeholder feedback. From September to December 2024, the Technical Working Group engaged in bi-weekly reviews and discussions, focusing on the identification of key metadata elements grouped by functional categories such as Data Discovery, Data Access, Data Provenance, Data Ownership, Temporal and Spatial Coverage, Population Coverage, Data Analytics, Data Quality, Variables, and Data Categorisation. A persona-based approach was adopted to evaluate use cases, focusing on the needs of different types of data users (e.g., researchers, policymakers). For example, use cases followed the structure: 'As a researcher, I want to easily discover relevant datasets so that I can conduct cross-border health studies. Additionally, requirements were derived from the FAIR data principles and aligned with EU policies, including the European Health Data Space Regulation, the Digital Governance Act, the HVD Implement Regulation, and the Data Act. Technical considerations were also addressed, encompassing Keyword, Faceted, Full-Text Search, Semantic Search, Natural Language Processing (NLP) and GenAI, Geospatial Search, and Metadata Management. The iterative process also involved the implementation of a Sandbox environment, where real DCAT metadata records were gathered and tested, comparing the current (AS-IS: DCAT-AP) and the future (TO-BE: HealthDCAT-AP) states. Feedback from health experts, collected via EU forms, was integral to refining these requirements and ensuring the HealthDCAT-AP specification met the practical and regulatory needs of the health domain. (Ref: "Technical working group on the transition from existing metadata templates to HealthDCAT-AP – Working group minutes")

Definition of the requirements:

Identification of the metadata elements by functional groups:

- Data Discovery

- Data Access

- Data Provenance

- Data Ownership

- Temporal Coverage

- Spatial Coverage

- Population Coverage

- Data analytics

- Data quality

- Variables

- Data categorisation

Review of use cases

- AS « persona » I WANT … SO THAT …

Requirements derived from the FAIR data principles

AS a metadata catalogue I WANT TO … SO THAT …

Requirements derived from EU Policies

- European Health Data Space Regulation

- Digital Governance Act
  (NSIP requirements)

- HVD Implement Regulation

- Data Act

Requirements derived from technical considerations

- Keyword, Faceted, Full-Text Search

- Semantic Search

- Natural Language Processing (NLP) and GenAI

- Geospatial Search

- Metadata Management

Implementation of a sandbox

- Gathering and testing real DCAT metadata records (AS-IS vs TO-BE)

2.6. Structure of this document

The **Introduction** chapter provides an essential foundation for understanding the HealthDCAT-AP initiative, offering a clear view of its context and objectives. It outlines the broader European landscape, focusing on data interoperability and standardisation in the health sector. The section also introduces the methodology, structure of the document, and presents a roadmap for HealthDCAT-AP development.

The **Terminology** chapter defines key terms and concepts used throughout the document, ensuring a common understanding of the vocabulary around metadata, datasets, catalogues, and interoperability.

The chapter 4 outlines the **scope of HealthDCAT-AP**, focusing on how it fits within the broader DCAT-AP family and health-related standards.

The **HealthDCAT-AP model** chapter provides a detailed overview of the conceptual model, with a focus on how health-specific metadata elements are structured and managed. It offers a clear vision of the overall design of HealthDCAT-AP, highlighting its strategic

approach to addressing the unique needs of describing health datasets while ensuring alignment with broader interoperability frameworks.

The **Uptake Strategies** chapter outlines the strategies for adopting HealthDCAT-AP by various stakeholders, including data holders, Health Data Access Bodies (HDABs), and data users. It provides a guide on how different organisations can integrate HealthDCAT-AP into their systems, facilitating data access and discovery.

2.7. Draft version of HealthDCAT-AP

The draft version of HealthDCAT-AP is available on GitHub at https://HealthDCAT-AP.github.io/. This draft adheres to the W3C ReSpec documentation style[13], ensuring a structured approach to technical documentation. By using this format, HealthDCAT-AP aligns with other application profiles such as DCAT-AP, GeoDCAT-AP, and DCAT-AP HVD, serving as the authoritative reference for health dataset catalogue implementers.

GitHub serves as the primary platform for feedback, allowing users to contribute to the draft's refinement by submitting issues, commenting on specific sections, or proposing enhancements through pull requests. This interactive feature encourages community collaboration, enabling implementers, stakeholders, and other users to contribute to the specification's development.

2.8. Mapping to DCAT-AP, HealthDCAT-AP

In the context of data sharing, mapping a metadata model to DCAT-AP is a relatively common exercise because organisations are free to adopt any data management and metadata standard that meets their needs, or even develop a custom model. However, to ensure interoperability and facilitate the exchange of metadata between dataset catalogues and across different domains, DCAT-AP serves as a common interchange metadata standard.

The mapping process consists in aligning metadata elements from the chosen standard with DCAT-AP properties. While mapping an organisation's metadata model to DCAT-AP is essential for consistent metadata sharing, it often presents significant challenges. To mitigate these challenges, it is necessary to optimise the mapping process to prevent information loss and ensure that all necessary data is accurately represented.

Examples of metadata mapping to DCAT-AP:

ISO 19139 (Geographic MetaData XML (gmd) encoding) to GeoDCAT-AP

SDMX (Statistical Data and Metadata eXchange) to StatDCAT-AP

CKAN (Catalogue system's metadata model) to DCAT-AP

EHDS Regulation Article 79) EU Dataset Catalogue

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

---

[13] https://respec.org/docs/

Within the HealthData@EU infrastructure, HealthDCAT-AP will serve as the common interchange metadata standard between the national dataset catalogues of Health Data Access Bodies and authorised participants, and the EU dataset catalogue. Considering that mapping to DCAT-AP is a complex process often fraught with technical issues, semantic challenges, and potential information loss, it is, therefore, recommended to adopt and standardise the use of HealthDCAT-AP across the entire HealthData@EU infrastructure.

However, specific requirements for particular health domains may arise and must be addressed. In such cases, DCAT allows for extensions to meet these needs. It is crucial that stakeholders ensure the interoperability of HealthDCAT-AP remains intact. Extending HealthDCAT-AP at the national level, the authorised participant level, or customising it at the data holders' level is a viable option, and will likely facilitate the integration and adoption of HealthDCAT-AP within the HealthData@EU infrastructure.

3.Terminology used in this document

EHDS Regulation Article 2 Definition

(y) 'dataset catalogue' means 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;

(f) 'interoperability' means the ability of organisations, as well as of software applications or devices from the same manufacturer or different manufacturers, to interact through the processes they support ▌ , involving the exchange of information and knowledge, without changing the content of the data, between those organisations, software applications or devices;

(u) 'health data user' means a natural or legal person, including Union institutions, bodies, offices or agencies, which has been granted lawful access to ▌ electronic health data for secondary use pursuant to a data permit, a health data request approval or an access approval by an authorised participant in HealthData@EU;

(w) 'dataset' means a structured collection of electronic health data;

(ad) 'data quality' means the degree to which the elements of electronic health data are suitable for their intended primary and secondary use;

(aa) 'data quality and utility label' means a graphic diagram, including a scale, describing the data quality and conditions of use of a dataset;

(t) 'health data holder' means any natural or legal person, public authority, agency or other body in the healthcare or the care sectors, including reimbursement services where necessary, as well as any natural or legal

person developing products or services intended for the health, healthcare or care sectors, developing or manufacturing wellness applications, performing research in relation to the healthcare or care sectors or acting as a mortality registry, as well as any Union institution, body, office or agency, that has either:

(i) the right or obligation, in accordance with █ applicable Union or national law and in its capacity as a controller or joint controller, to process personal electronic health data for the provision of healthcare or care or for the purposes of public health, reimbursement, research, innovation, policy making, official statistics or patient safety or for regulatory purposes; or

(ii) the ability to make available non-personal electronic health data through the control of the technical design of a product and related services, including by registering, providing, restricting access to or exchanging such data;

---

(64) The establishment of one or more health data access bodies, supporting access to electronic health data in Member States, is essential to promoting the secondary use of health-related data. Member States should therefore establish one or more health data access bodies to reflect, inter alia, their constitutional, organisational and administrative structure. However, one of those health data access bodies should be designated as a coordinator in the event there is more than one health data access body. Where a Member State establishes several health data access bodies, it should lay down rules at national level to ensure the coordinated participation of those bodies in the European Health Data Space Board (the 'EHDS Board'). That Member State should, in particular, designate one health data access body to function as a single contact point for the effective participation of those bodies, and ensure swift and smooth cooperation with other health data access bodies, the EHDS Board and the Commission. Health data access bodies could vary in terms of organisation and size, spanning from a dedicated fully fledged organisation to a unit or department in an existing organisation █ .

Controlled vocabulary[14]: A controlled vocabulary is a predefined, standardised set of terms and phrases used to ensure consistency in naming and categorising concepts within a dataset. It restricts the use of alternative terms or synonyms to avoid ambiguity and maintain uniformity in data description. Controlled vocabularies are commonly used in metadata, taxonomies, and classification systems to improve data discoverability, interoperability, and accuracy in search and retrieval across datasets or systems. Examples include thesauri, ontologies, and code lists.

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

Data dictionary: A data dictionary is a centralised repository of metadata that provides definitions, descriptions, and details about the structure, fields, and variables within a dataset. It typically includes information such as data types, allowed values, relationships between fields, and the meaning of each element. A data dictionary helps data users understand the content and structure of a dataset, facilitating proper use and interpretation of the data.

---

[14] Controlled vocabularies - EU Vocabularies - Publications Office of the EU (europa.eu)

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

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

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

Tabular data: Tabular data refers to data organised in a structured format of rows and columns, where each row represents a single record or entity, and each column represents a specific attribute or variable. This structure is commonly found in spreadsheets or relational databases, making it easy to store, query, and analyse. Tabular data is often used for structured datasets where relationships between variables are well-defined. (See: Metadata vocabulary for tabular data)

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

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

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

4. Scope of HealthDCAT-AP

HealthDCAT-AP is an extension of the European DCAT-AP (Data Catalog Vocabulary Application Profile) designed to meet the specific metadata needs of the healthData@EU infrastructure. While it retains the core structure of DCAT-AP, HealthDCAT-AP introduces additional classes and metadata elements specifically adapted to the health sector. The purpose of HealthDCAT-AP is to streamline (meta)data exchange within the healthData@EU infrastructure and ensure interoperability with other EU Data Spaces.

EHDS Regulation Article 77) Dataset description and dataset catalogue

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

What is DCAT-AP? DCAT-AP and any DCAT Application Profiles like HealthDCAT-AP is a descriptive metadata standard designed to ensure the interoperability and exchange of metadata records across multiple dataset catalogues within the EU. By providing a common structure, it allows different catalogues to seamlessly exchange metadata, improving the discoverability and accessibility of datasets across sectors and borders. It ensures that datasets are described in a consistent manner, making it easier to search for, find, and access data across various platforms. By providing a common structure for metadata, DCAT-AP allows different dataset catalogues to interoperate, which is crucial for integrating datasets from different sources. It improves the discoverability of datasets by ensuring that key descriptive information (such as title, description, keywords, and access information) is available and formatted in a uniform way. It supports the sharing of data across borders and sectors within the EU by ensuring that all dataset catalogues can interpret and use the metadata in a compatible manner. While the adoption of DCAT-AP as a standard for EU dataset catalogues is encouraged through various legal acts, such as Article 33 of the Data Act, DCAT-AP itself is not explicitly mentioned in EU legal texts. This is because DCAT-AP is a technological framework, and the possibility of disruptive technical evolution is always present. However, it is widely recognised by experts as the best technical solution currently available for implementing EU data spaces.

Data Act - Article 33

Essential requirements regarding interoperability of data, of data sharing mechanisms and services, as well as of common European data spaces
1. Participants in data spaces that offer data or data services to other participants shall comply with the following essential requirements to facilitate the interoperability of data, of data sharing mechanisms and services, ...
(b)   the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists, where available, shall be described in a publicly available and consistent manner;

What DCAT-AP is not: Although DCAT-AP is a powerful tool for metadata exchange, it is not commonly used for internal data management due to its reliance on RDF (Resource

Description Framework), a graph-based structure. Many organisations still rely on relational databases with tabular formats, making it more challenging to integrate DCAT-AP without additional technological adjustments. They do not perceive the need for a standardised metadata model like DCAT-AP. Moreover the most popular database management systems are typically relational databases, which use a table-based structure with rows and columns. In contrast, DCAT-AP is based on RDF (Resource Description Framework), a graph-based structure that uses triples (subject-predicate-object) to represent data. IT departments are more accustomed to traditional database management and may be unfamiliar with linked-data technologies such as DCAT-AP. This fundamental difference means that organisations would need to implement an additional layer of technology to integrate DCAT-AP as a metadata standard within their existing data management systems (Ref: mapping to DCAT-AP). In summary, DCAT-AP may not align with the current technical infrastructure or needs of many organisations, especially those that do not prioritise interoperability or lack expertise in Semantic web technologies.

Impact of the EHDS Regulation: Entities subject to the EHDS Regulation - including natural or legal persons, public authorities, healthcare providers, research institutions, and others managing health data - must ensure that the data they hold is discoverable and interoperable within the healthData@EU infrastructure. These data holders are required to create, manage and make metadata about their datasets (as outlined in Article 33 of the EHDS Regulation) available in a standardised, machine-readable format in the National dataset catalogue. The use of HealthDCAT-AP supports compliance with the interoperability and data exchange standards necessary for the secure, cross-border sharing of health data across the EU.

EHDS Regulation Article 60) Duties of health data holders

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

EHDS Regulation Article 51) Minimum categories of electronic data for secondary use

1. Health data holders shall make the following categories of electronic health data available for secondary use in accordance with this Chapter:
(a) electronic health data from EHRs;
(b) data on factors impacting on health, including socio-economic, environmental and behavioural determinants of health;
(c) aggregated data on healthcare needs, resources allocated to healthcare, the provision of and access to healthcare, healthcare expenditure and financing;
(d) data on pathogens that impact human health;
(e) healthcare-related administrative data, including on dispensations, reimbursement claims and reimbursements;
(f) human genetic, epigenomic and genomic data;
(g) other human molecular data such as proteomic, transcriptomic, metabolomic, lipidomic and other omic data;
(h) personal electronic health data automatically generated through medical devices ;
(i) data from wellness applications;
(j) data on professional status, and on the specialisation and institution of health professionals involved in the treatment of a natural person;
(k) data from population-based health data registries such as public health

4.1. HealthDCAT-AP and health standards

The health data landscape encompasses a wide array of domains, including: Clinical data, epidemiological data, public health data, pharmaceutical data, genomic data, healthcare facility data, patient demographics and more... Article 51 of the EHDS Regulation expands the scope of health data to include additional domains that impact health, such as pathogen and environmental data. Many of these domains already rely on specific standards for describing, categorising, or modeling their data, often tailored to their unique purposes. Below is a brief, non-exhaustive overview of some of these standards and their focal areas:

- The **HL7** ([Health Level Seven International](#)) [wikidata:Q17054989](#) defines a set of international standards for the exchange, integration, sharing, and retrieval of electronic health information. HL7 standards provide a comprehensive framework for clinical and administrative data. Its primary scope is the exchange of individual clinical and administrative data elements (e.g., patient demographics, clinical observations). It is used to describe individual health records or transactions, not entire datasets. HL7's data models and messaging standards are HL7 V2, V3, and FHIR.

- **FHIR** ([Fast Healthcare Interoperability Resources](#)) [wikidata:Q19597236](#) is a standard describing data formats and elements for exchanging electronic health records. Developed by HL7, it is designed to enable fast and efficient exchange of healthcare information. It uses modern web technologies and focuses on interoperability. FHIR focuses on specific elements like patients, observations, medications, and other clinical data points rather than on metadata for datasets as a whole.

- **OpenEHR** [wikidata:Q838025](#) is an open standard specification in health informatics that describes the management and storage, retrieval and exchange of health data in electronic health records (EHRs). Part of the OpenEHR framework, [OpenEHR Archetypes](#) are formal models or templates that define the structure, meaning, and relationships of health-related data in an interoperable and standardised way.

- **ICD** ([International Classification of Diseases](#)) [wikidata:Q50018](#) is a globally recognised standard , maintained by the World Health Organization (WHO), for coding diseases and health conditions. It provides standardising classification codes for diseases and health conditions. ICD codes and descriptions can be used to standardise the classification of health-related datasets in HealthDCAT-AP.

- **LOINC** ([Logical Observation Identifiers Names and Codes](#)) [wikidata:Q502480](#) is a universal standard for identifying health measurements, laboratory observations, and clinical data. LOINC codes can be used to describe lab tests, measurements, and other clinical observations in HealthDCAT-AP.

- **SNOMED CT** (Systematized Nomenclature of Medicine Clinical Terms) wikidata:Q37616346 is a comprehensive clinical terminology that provides codes, terms, synonyms, and definitions used in clinical documentation and reporting such as diseases, clinical findings, and procedures. SNOMED CT can be utilised to describe clinical concepts and healthcare terms in HealthDCAT-AP.

- The **ISO/IEC 11179** wikidata:Q3146900 standard provides guidelines for metadata registries, including the registration and management of metadata for data elements. It offers a structured approach to define and manage metadata elements, which can be applied to health datasets. As DCAT does not provide recommendations on metadata management, ISO/IEC 11179 can serve as a complementary standard to provide the necessary guidelines for Health Data Access Bodies to manage metadata effectively. Together, DCAT and ISO/IEC 11179 can support the creation of interoperable and well-governed health data spaces.

- **OMOP** (Observational Medical Outcomes Partnership) wikidata:Q125499706 Common Data Model standardises the format and content of observational health datasets (i.e.: clinical observations, treatments, and outcomes data).

- **CDISC** (Clinical Data Interchange Standards Consortium) wikidata:Q571067 standards facilitate the exchange of clinical trial data and include models like CDASH (Clinical Data Acquisition Standards Harmonization) and SDTM (Study Data Tabulation Model).

- **SDMX** (Statistical Data and Metadata Exchange) wikidata:Q2713163 is an international initiative that aims to standardise the exchange of statistical data and metadata. Used by statistical organisations to describe and exchange entire datasets and their metadata. It is the foundation for StatDCAT-AP.

- **DICOM** (Digital Imaging and Communications in Medicine) wikidata:Q81095 is a standard for the handling, storing, printing, and transmitting information in medical imaging.

- **MeSH** (Medical Subject Headings) wikidata:Q199897 is a comprehensive controlled vocabulary used by the National Library of Medicine (NLM) to index and organise biomedical and health-related information in databases like PubMed and MEDLINE. MeSH terms facilitate precise and consistent search and retrieval of scientific and medical information by categorising content into hierarchical topics and subtopics. This system includes descriptors, qualifiers, and supplementary concept records to cover various aspects of medical knowledge, ensuring that researchers, healthcare professionals, and librarians can find relevant information efficiently.

- **GSIM** (Generic Statistical Information Model) wikidata:Q122873933 GSIM is a reference framework of internationally agreed definitions, attributes and relationships that describe the pieces of information used in the production of official statistics (information objects). The framework enables generic descriptions of the definition, management and use of data and metadata throughout the statistical production process.

- **WHO** classifications: https://www.who.int/standards/classifications

By integrating established health data standards, HealthDCAT-AP offers a comprehensive framework for describing health datasets in a machine-actionable and interoperable way. This ensures that health data, whether clinical, epidemiological, or genomic, can be efficiently exchanged across Member States, facilitating secondary uses like research and public health analysis. Integrating these standards into various HealthDCAT-AP properties facilitate improved interoperability by promoting standardised data exchange protocols, as well as harmonised data models and formats. These standards, developed to meet specific domain requirements, are highly effective within their respective contexts. It is important that these standards are in use in HealthDCAT-AP as soon as they are de facto recognised

as standard due to widespread community usage. During the implementation phase of the EHDS, reviewing the common standards used by health data holders to describe their datasets would be highly beneficial. This review will offer valuable insights for the effective governance of the health data space for instance to evaluate and further foster data harmonisation advancements.

HealthDCAT-AP has been designed to serve as a comprehensive framework for describing diverse health data that utilise various data models, exchange services, formats, and vocabularies, including thesauri and ontologies. It provides an integrated solution for managing data relevant to Article 33 within health dataset catalogues. The vocabulary of HealthDCAT-AP must be robust enough to describe the standards associated with the dataset in a machine-actionable way.

Examples of health Data Models, Services, Formats, and Ontologies for use in HealthDCAT-AP				
Data models	**Data services**	**Formats**	**Thesauri**	**Ontologies**
OMOP Common Data Model: A framework for transforming data from various sources (e.g., EHRs, claims data) into a common format. **OpenEHR:** Open standard for managing electronic health records (EHRs). **SDTM** (Study Data Tabulation Model) defines a standard structure for human clinical trial (study) data tabulations etc.	**FHIR API:** Standard APIs for exchanging healthcare data using RESTful principles. **OpenEHR API:** Facilitating the management and exchange of electronic health records. **DICOM Protocol:** Beyond being just a file format, DICOM also defines the communication protocol used to exchange medical images and associated information between medical devices, such as scanners, servers, workstations, and printers. This ensures that different systems and devices can communicate and interpret the data correctly.	**DICOM File Format:** The DICOM standard defines a file format for medical images, which includes both the image data (e.g., MRI, CT scans, X-rays) and associated metadata (e.g., patient information, imaging parameters). The file extension is typically .dcm. **FASTA** is a text-based format used for representing nucleotide sequences or peptide sequences (proteins) in bioinformatics. etc.	**ICD** is used globally for health management, epidemiology, and clinical purposes, providing codes for diseases, conditions, and procedures. **UMLS** integrates multiple health and biomedical vocabularies, providing a large compendium of healthcare-related terms and their relationships. **RxNorm:** A normalised naming system for generic and branded drugs **MESH** (Medical Subject Headings) etc.	**GO** (Gene Ontology) provides a controlled vocabulary to describe gene and gene product attributes across all species, focusing on biological processes, cellular components, and molecular functions. **OBO** (Open Biological and Biomedical Ontologies) **DOID** (Disease Ontology) etc.

Search for Medical data models:

Medical-Data-Models.org is a web-based platform designed to provide access to a comprehensive repository of data models that are used in clinical research and patient care.

4.2. HealthDCAT-AP vs DCAT-AP, GeoDCAT, StatDCAT, DCAT-AP HVD

The data covered by the EHDS Regulation is extensive and diverse, and it may already be described by publishers using other specialised DCAT Application Profiles, such as DCAT-AP, GeoDCAT-AP, StatDCAT-AP, and DCAT-AP HVD. For example, "*data known to influence health*" encompasses environmental data, which is already findable and accessible within existing geospatial infrastructures, with datasets typically described in compliance with INSPIRE and mapped to GeoDCAT-AP. It is important to recognise that the EHDS Regulation will apply to a wide array of datasets that are already catalogued in European data portals. For instance, data.europa.eu currently hosts around 27,000 records labelled as health-related datasets, which may need to be aligned with the EHDS-specific requirements. It is important to retain that HealthDCAT-AP is specifically tailored to address the diverse nature of data covered by the European Health Data Space. As such it includes various dataset types as well as various data access rights within one unique Data Space.

EHDS Regulation Recital 56)

The categories of electronic health data that can be processed for secondary use should be broad and flexible enough to accommodate the evolving needs of health data users, while remaining limited to data related to health or known to influence health. They can also include relevant data from the health system, for example electronic health records, claims data, dispensation data, data from disease registries or genomic data, as well as data with an impact on health, for example data on consumption of different substances, socio-economic status or behaviour, and data on environmental factors such as pollution, radiation or the use of certain chemical substances. The categories of electronic health data for secondary use ▌include some categories of data that were initially collected for other purposes such as research, statistics, patient safety, regulatory activities or policy making, for example, policy-making registries or registries concerning the side effects of medicinal products or medical devices. European databases that facilitate use or reuse of data are available in some areas, such as cancer (the European Cancer Information System) or rare diseases (for example, the European Platform on Rare Disease Registration and European reference networks (ERN) registries). The categories of electronic health data that can be processed for secondary use should also include automatically generated data from medical devices and person-generated data, such as data from wellness applications. Data on clinical trials and clinical investigations should also be included in the categories of electronic health data for secondary use when the clinical trial or clinical investigation has ended, without affecting any voluntary data sharing by the sponsors of ongoing trials and investigations. Electronic health data for secondary use ▌should ▌be made available preferably in a structured electronic format that facilitates their processing by computer systems. Examples of structured electronic formats include records in a relational database, XML documents or CSV files and free text, audios, videos and images provided as computer-readable files.

The High-Value Datasets Implementing Regulation (HVD IR) serves as a key example of how new regulatory frameworks can impose additional requirements on existing DCAT-AP records. The HVD IR introduces a set of rules specifically applicable to certain classes of datasets classified as high-value, which are organised into six thematic categories:

Geospatial, Earth Observation and Environment, Meteorological, Statistics, Companies and Company Ownership, and Mobility. Publishers whose datasets fall within the scope of the HVD IR are obligated to enhance their dataset descriptions to meet these new standards.

As the DCAT-AP HVD specification states, "Any improvement of the metadata will immediately flow throughout the European network of (Open) Data Portals and thus increase the level of metadata quality." This reflects the broader impact of the HVD IR, where enhancements in metadata quality are propagated across the entire network of data portals, thereby elevating the overall standard of metadata.

The improvements in metadata quality mandated by the HVD IR for existing DCAT-AP records are analogous to the enhancements required by HealthDCAT-AP for open health-related data. Both frameworks underscore the importance of high-quality, interoperable metadata and require, at least, one dataset Distribution.

EHDS Regulation Article 60)

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

EHDS Regulation Art. 77) Dataset description and dataset catalogue

> *4. … the Commission shall, by means of implementing acts, set out the minimum ▌ elements health data holders are to provide for datasets and the characteristics of those elements. Those implementing acts shall be adopted in accordance with the examination procedure referred to in Article 98(2).*

Regarding the scope of HealthDCAT-AP, we recommend the following approach: for EHDS Art.55, HealthDCAT-AP should align with the "minimum elements" established by DCAT-AP HVD. For non-public data (i.e., protected and sensitive health data), for high-impact datasets (as mentioned in EHDS Article 58), HealthDCAT-AP will introduce additional "minimum elements". For all protected and sensitive datasets that are subject to data application procedures managed by Health Data Access Bodies, data users will benefit from enhanced data descriptions. These enhancements might include mandatory variable descriptions (i.e., data dictionaries) or/and the availability of open synthetic datasets.

This approach would enables health dataset catalogues to harvest metadata records from other sources that have a HVD reporting endpoint, with the capability to filter datasets tagged as health-related*.

Moreover, the healthData@EU infrastructure will not only benefit from ongoing improvements introduced by the core DCAT-AP vocabulary but will also gain from the metadata quality enhancements mandated by the HVD IR.

EHDS Regulation Recital 58)

> *In order to increase the effectiveness of the secondary use of personal electronic health data, and to fully benefit from the possibilities offered by this Regulation, the availability in the EHDS of electronic health data described in Chapter IV should be such that the data are as accessible, high-quality, ready and suitable for the purpose of creating scientific, innovative and societal value and quality as*

> *possible. Work on the implementation of the EHDS and further dataset
> improvements should be conducted in a manner that **prioritises the datasets that
> are the most suitable for creating such value and quality.***

EHDS Regulation Art. 80) Minimum specifications for datasets of high impact

> *The Commission may, by means of implementing acts, determine the minimum specifications
> for datasets of high impact for secondary use, taking into account existing Union
> infrastructures, standards, guidelines and recommendations. Those implementing acts shall be
> adopted in accordance with the examination procedure referred to in Article 98(2).*
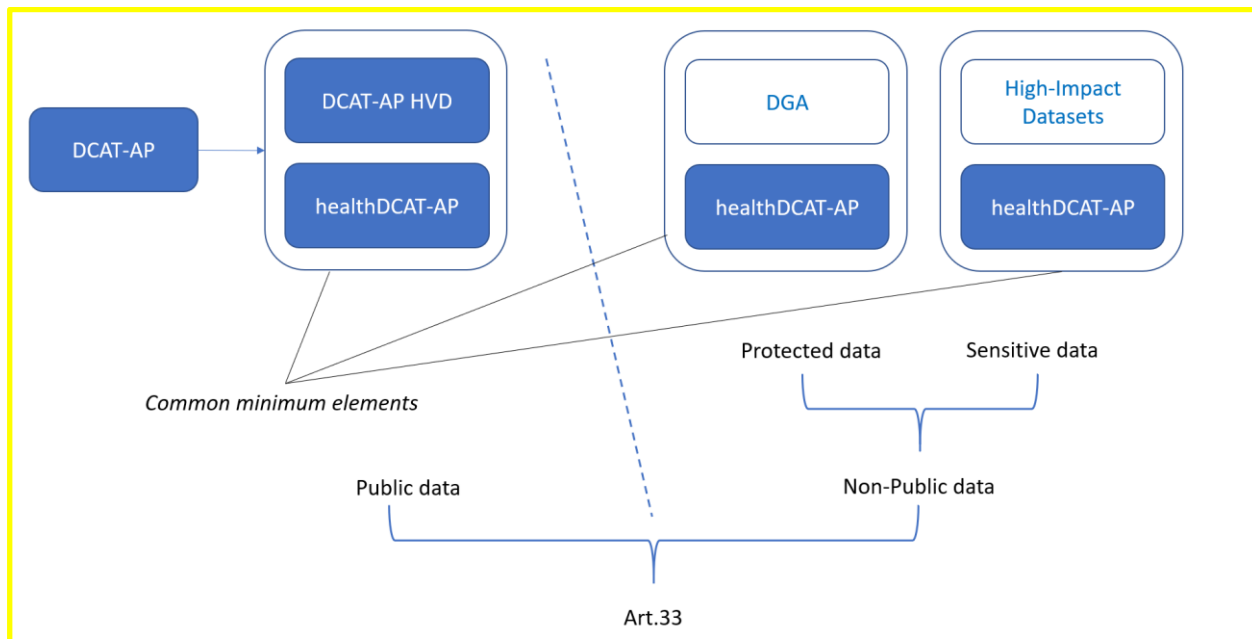
Figure 1: HealthDCAT-AP vs Art.33 defining datasets in scope of the EHDS

Expected Scenario: When, for instance, a statistical institute complies with the HVD IR, it will also meet the requirements of HealthDCAT-AP, as both application profiles share the same minimum metadata elements for the publication of open data (e.g., statistical data). This dual compliance ensures that Health Data Access Bodies (HDABs) can seamlessly integrate and populate health catalogues with DCAT-AP HVD records that fall under the scope of Article 51.

* <u>Attention Point</u>: Health-related datasets

<u>Challenge</u>: The EHDS ELI URI must be used in the "Applicable Legislation" property, which is mandatory in DCAT-AP HVD. Since the "Theme" property is not mandatory in either DCAT-AP or DCAT-AP HVD, it is important to ensure that the "Applicable Legislation" field is accurately and comprehensively filled for High-Value Datasets (HVDs) to facilitate the identification and harvesting of datasets under the scope of EHDS Article 33.

<u>Solution</u>: To assist data holders in complying with the EHDS Regulation and other legal frameworks, the European Commission could implement an AI-based metadata

enhancement service (e.g., API service). This service, similar to the automated 'Eurovoc' tagging in data.europa.eu, would help identify applicable legal frameworks, such as the EHDS, and suggest enhancements to dataset descriptions. Such a service would ensure that metadata aligns with the relevant DCAT-Application Profiles and supports legal compliance across sectors.

A notable precedent is the automatic "Eurovoc" tagging introduced by data.europa.eu, which offers a valuable use case for SEMIC experts to explore further. This tagging system provides a novel approach for implementing properties that support faceted search, a concept that should be considered in the future development of DCAT Application Profiles to improve the efficiency of dataset catalogues.

https://data.europa.eu/en/apps https://data.europa.eu/annif/

Annif is an open-source toolkit for automated subject indexing. It integrates several machine learning and AI based algorithms for text classification. This implementation helps to tag EU Vocabulary properties for datasets.

Comment: HealthDCAT-AP extends DCAT-AP by introducing two additional properties, hasCodingSystem and hasCodeValues, which enable the tagging of datasets using any controlled vocabulary, such as Eurovoc.

5. HealthDCAT-AP model

5.1. Informal description

The HealthDCAT-AP is meant to facilitate metadata management and data interoperability under the EHDS framework, which regulates the secondary use of health data for research, policy-making, and other non-primary purposes. While it works in conjunction with the healthData@EU infrastructure, its design is primarily driven by the legal requirements of the EHDS Regulation concerning data cataloguing and interoperability.

It retains the foundational structure and concepts of DCAT-AP, while incorporating additional classes and metadata elements to suit the unique needs of the health domain.

The purpose of HealthDCAT-AP is to streamline data exchange within the healthData@EU infrastructure and ensure interoperability with other EU data spaces.

5.2. Extensions and specific usage for description of health datasets

During the Technical Working Group (TWG) sessions on the design of the HealthDCAT Application Profile, a number of requirements were identified and analysed[15] for managing and sharing health datasets. New metadata properties were introduced to extend the DCAT-AP vocabulary, addressing those requirements. Some of the properties are to be defined in a new namespace for HealthDCAT-AP and other properties are reused from existing RDF vocabularies (e.g., DPV[16] or DQV[17]). The suggested namespace prefix for HealthDCAT-AP is:

---

[15] All minutes of the TWG sessions are compiled in the Milestone M6.2 "Technical working group on the transition from existing metadata templates to HealthDCAT-AP - Working group minutes".
[16] Data Privacy Vocabulary (DPV)

"healthdcatap".

The key elements of the vision outlined in this chapter offer compelling reasons for adopting the HealthDCAT-AP model, helping stakeholders understand the rationale behind extending DCAT-AP for health data and the resulting benefits.

5.2.1. Metadata management and building the health data space as a knowledge graph

5.2.1.1. Metadata identifiers

HealthDCAT-AP mandates the use of persistent dereferenceable URIs (i.e., HTTP URIs) for the identifier of the metadata record. (i.e.: dct:identifier –

Ex: https://hdab-catalogue.org/publisher_0/datasets/68112f77-8f2c-496a-8398-77e52b60c883).

A persistent identifier is a stable and unique reference to a metadata record or dataset. In HealthDCAT-AP, these identifiers are typically HTTP URIs - web addresses that point directly to the specific dataset's description or metadata. By using persistent URIs, the data can be reliably accessed and shared across platforms, ensuring that even if the metadata is moved or copied, the original reference remains intact. This allows for consistent data management and discoverability over time.

The use of persistent dereferenceable URIs also applies to other DCAT properties linking a dataset to other datasets (e.g.: IsVersionOf, …) or external resources (e.g.: IsReferencedBy,…) where persistent dereferenceable URIs are expected. The use of persistent dereferenceable URIs are essential for ensuring the stability, reliability, and accessibility of resources within the EU Health Data Space and beyond.

We know the benefit of Digital Object Identifiers (DOI)[18] for providing stable, reliable, and accessible identifiers for digital objects. A DOI is a unique alphanumeric string assigned to a digital object that provides a permanent link to its location on the Internet. It enhances the organisation, sharing, and citation of digital content in a wide range of fields like scientific publications. The same logic applies to metadata records that will be shared and exchanged between dataset catalogues. Similarly, in HealthDCAT-AP, each metadata record is assigned a persistent dereferenceable URI (HTTP URIs), ensuring that the original metadata can always be reliably accessed. The primary dataset catalogue serves as the single source of truth, and any updates must occur at the source. Copies made for discoverability must always reference the original metadata.

---

[17] [Data on the Web Best Practices: Data Quality Vocabulary](#)
[18] [Using DCAT-AP for research data](#)
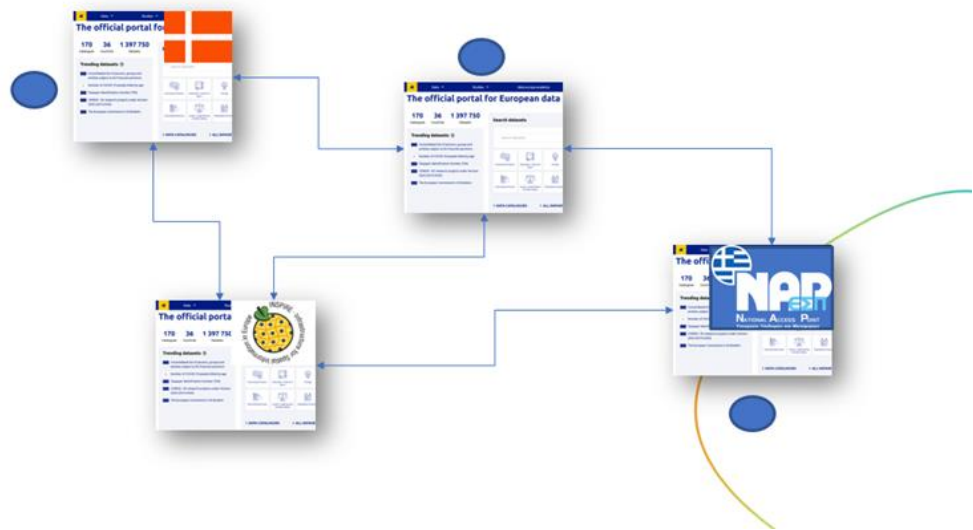
Webinar DCAT-AP Health - Session 1 (June 2023) by DIGIT.D2 – Interoperability – Use Case 1: Harvesting "PURIs allow to explore the network of catalogues beyond a single portal."

While copies of the master metadata can be made for enhancing discoverability purposes on the Web, only the master metadata must be updated at the primary source by the data holder. Maintaining persistent dereferenceable URIs for metadata identifiers allows efficient metadata management. "*It is the user that has to ensure that its use of a dataset is legally acceptable, not the data catalogue. And they can do this because users can retrieve the original metadata*"[19]. Thus at any time, a copy of a metadata can be retrieved by returning at its source thanks to its main identifier.

If a metadata record does not meet the HealthDCAT-AP requirements for metadata identifiers, Health Data Access Bodies (HDABs) are required to assign an additional persistent, dereferenceable URI (HTTP URI) to the metadata identifier while retaining the original identifier provided by the data holder. When aggregating metadata, HDABs must ensure the integrity of metadata. DCAT-AP 3.0's Catalogue Record class provides HDABs with the necessary metadata elements for managing aggregated metadata:

- Listing date: The date on which the description of the Resource was included in the Catalogue.
- Modification date: The most recent date on which the Catalogue entry was changed or modified.
- Source metadata: The original metadata that was used in creating metadata for the Dataset, Data Service or Dataset Series.
- Etc.

5.2.1.2. EU Health Data Space as a knowledge graph

DCAT is a RDF vocabulary representing a dataset catalogue. It is a knowledge graph (Linked

---

[19] Guidelines on the management of identifiers

Data) which relies on:

- [Dublin Core](#) to represent metadata elements
- [SKOS](#) to represent glossaries and thesauri
- [PROV](#) to represent provenance and data lineage
- [CSVW](#) to represent tabular data
- [Data Quality vocabulary](#) (DQV-AP) to describe quality
- [Data Protection vocabulary](#) (DPV) to describe data privacy

All these Web ontologies participate to create a knowledge graph: Fine grained structure of information unambiguously interpretable by machines. By using a knowledge graph, stakeholders can more easily find, understand, and use datasets, as the relationships between datasets and resources (like publications or codes) are explicitly defined.
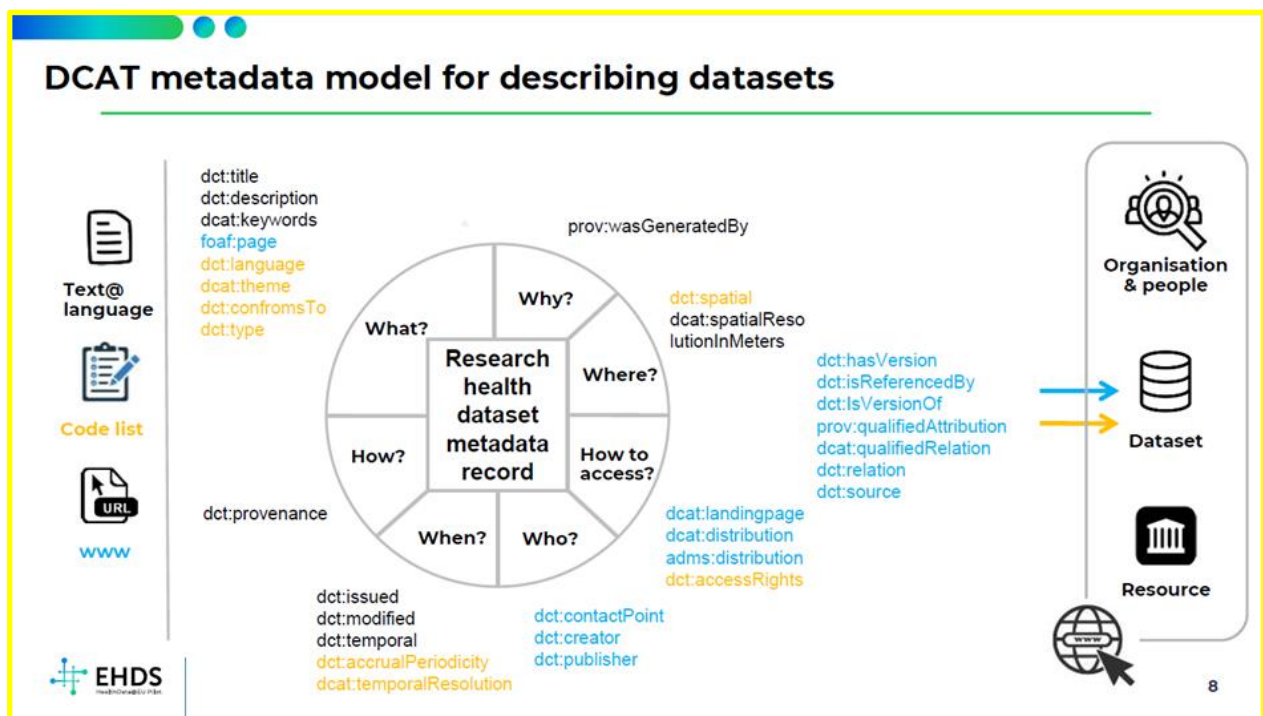
Annex 2: In describing a dataset, the DCAT-AP metadata model includes more relationships, such as SKOS code lists (shown in orange) and Web resources (shown in blue), expressed as HTTP URIs, than textual information (shown in black).

Figure 2: DCAT-AP includes properties that link a dataset to other datasets or resources using HTTP URIs.

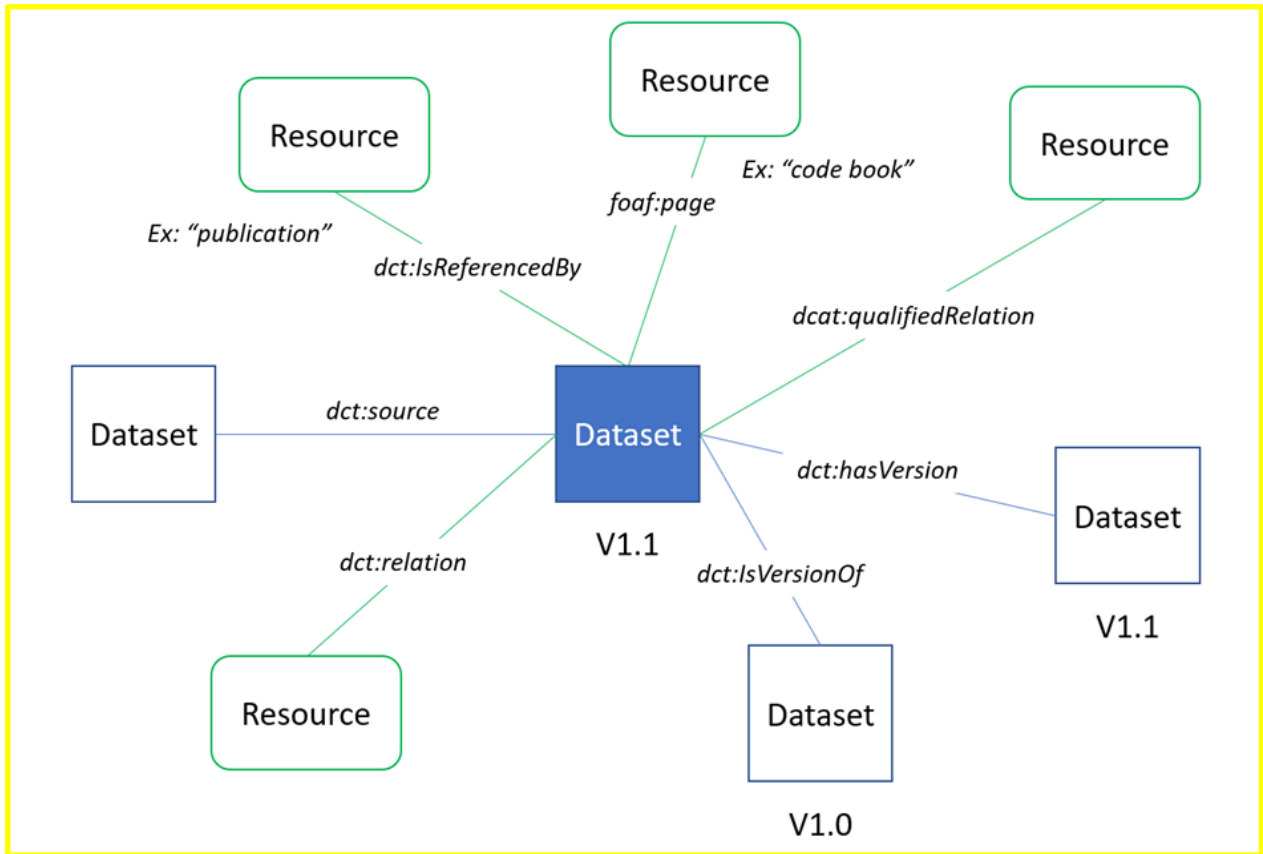To help data users understand how a dataset can be used, investigating its previous applications can be highly beneficial. As a knowledge graph, DCAT helps bridge input data with research data, offering insights into how datasets have been utilised (Annex 2). The requirement of using persistent URIs for metadata identifiers is not only of interest for the management of a system of distributed catalogues, it is also beneficial for building the European Health Data Space as an extended data graph where all resources are interconnected over the Web via communication standards and where machines can create knowledge (See: 2.1 Background - DCAT-AP in the EU Data Spaces).

Linked data principles:

- Use URIs as names for things
- Use HTTP URIs so that people and machines (i.e., applications) can look up those names.
- When someone looks up a URI, provide useful information.
- Include links to other URIs. so that they can discover more things.

One requirement for operating the EU Health Data Space is to care for the HTTP URIs and to be able to identify broken URIs (i.e., relationships). Another requirement for governing the EHDS will be to measure its level of maturity: percentage of URIs, level of achieved interoperability, for instance, metadata being a dataset proxy it provides information on the data conformity to standards (ref: dct:conformsTo), etc.

The principles of linked data (i.e., using URIs, providing useful information, and linking to other URIs) are foundational for building this interconnected space. Maintaining functional URIs and measuring the level of interoperability will be key to ensuring the maturity and effectiveness of the European Health Data Space.

5.2.1.3. Metadata harvesting or publishing – Metadata review process

The EHDS Regulation introduces the obligation for data holders to review their metadata records (aka "dataset description") once a year.

EHDS Regulation Art. 60 Duties of health data holders

> *3. ... The health data holder shall, at a minimum on an annual basis, check that its dataset description in the national dataset catalogue is accurate and up to date.*

This newly introduced rule for data holders requires a metadata control management process, adding a new feature to DCAT-AP. In the CatalogRecord class of DCAT-AP 3.0, the mandatory property dct:modified shows when the metadata was last updated, but it doesn't indicate when it was reviewed. In the same Class, the "listing date" informs on which date, the description of the dataset was included in the catalogue. If the EU central health dataset catalogue or national catalogues will manage the review process, HealthDCAT-AP would need a new property, like healthdcatap:reviewDate, to show when the metadata was last reviewed or updated by the data holder. Alternatively, an internal technical metadata property, "reviewDate," could be managed within the data holder's internal system without being shared across catalogues to address this requirement.

According to the EHDS Regulation, data holders are responsible for reviewing their dataset descriptions, with no mandate given to dataset catalogues for this process.

In both harvesting and publishing scenarios, it is best practice to include the DCAT CatalogRecord alongside the Dataset, Distribution, and DataService classes. The CatalogRecord is essential because it provides metadata specifically about the metadata entry itself. Notably, it indicates the Application Profile used for the metadata record (through dct:conformsTo), specifying which profile the catalogued resource's metadata adheres to. This enables aggregator agents to validate the metadata against specific constraints during integration. This information is particularly valuable as HealthDCAT-AP introduces three distinct sets of cardinalities, making it crucial for ensuring accurate validation and interoperability across catalogues.

5.2.2. Metadata fit for the purpose of Generative AI

Semantic search is an advanced search technique that enhances search accuracy by understanding the meaning behind queries and the content of indexed data, rather than only matching keywords. It leverages natural language processing (NLP), machine learning (such as Large Language Models), and ontologies to interpret the context, intent, and relationships between words in a query. Unlike traditional keyword-based search, semantic search algorithms consider the broader context to better interpret user intent.

For example, Elasticsearch or OpenSearch are advanced search tools that go beyond basic keyword searches. They power semantic search, which understands the intent behind a query instead of just matching keywords. By leveraging Natural Language Processing (NLP)

and machine learning techniques, like Large Language Models (LLMs), ElasticSearch enhances how datasets are searched. For example, it recognises that 'myocardial infarction' and 'heart attack' are related terms, providing users with more accurate and relevant search results.

For non-technical users, this means they can use more natural language queries, such as 'What are the main causes of heart disease in Europe?' and receive contextually appropriate results. Technical users, on the other hand, will benefit from vector-based search, where terms are represented as numerical vectors, making it easier to capture relationships between different datasets.

Recognising the significance of semantic search as a major evolution in search technology, HealthDCAT-AP has introduced new free-text properties and RDF concepts to further extend the DCAT knowledge graph:

- Free-text Properties in HealthDCAT-AP: Properties such as dct:description, healthdcatap:populationcoverage (what), dpv:hasPurpose (why), dct:provenance (how), and healthdcatap:publishernote (who) are designed to enhance NLP/LLM capabilities.
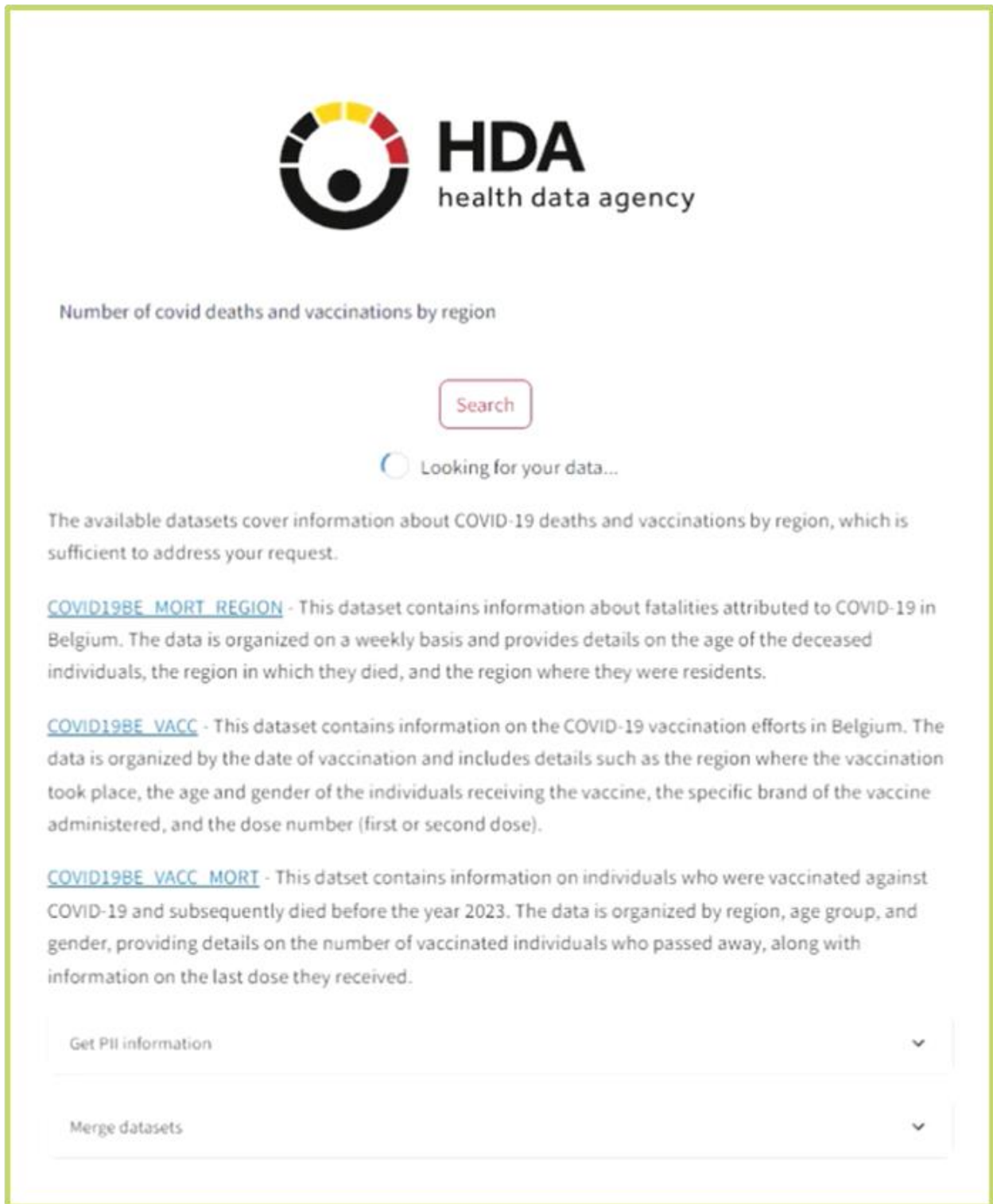
Figure 3: Screenshot of a Proof Of Concept genAI application integrated in a dataset catalogue developed by the Belgian Health Data Agency.

- **Enhancing semantic search through ontology integration** with the extended HealthDCAT-AP knowledge graph, enriched with additional ontologies like the Data Privacy Vocabulary, healthdcatap:healthTheme and healthdcatap:hasCodeValues (based on SKOS), leverages Linked Data principles to create a more interconnected and meaningful search experience. By utilising the SPARQL Protocol and RDF Query Language (SPARQL), users can

perform precise and targeted queries that explore the rich relationships and contextual relevance embedded within the graph. This integration enables users to discover information that is both semantically enriched and contextually relevant, maximising the power of Linked Data to connect and interpret disparate data sources.

This approach is exemplified in Rajaram Kaliyaperumal's presentation, "Metadata and FAIR Data Point," particularly in the slides titled "Why storing metadata in RDF is better?" In this presentation, two use cases are demonstrated, both relying on Linked Data principles to showcase the advantages of RDF-based metadata storage. These use cases highlight the potential of linked data infrastructures, illustrating how semantic web technologies can significantly enhance data discoverability and interoperability.

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX ejp: <http://purl.org/ejp-rd/vocabulary/>
PREFIX dcterm: <http://purl.org/dc/terms/>
PREFIX fdo: <http://rdf.biosemantics.org/ontologies/fdp-o#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>

SELECT DISTINCT ?resource ?id ?title ?description ?homepage (STR(?disease_name) AS ?disease)
?country_name ?country_code WHERE {

  VALUES ?disease_iri {<http://www.orpha.net/ORDO/Orphanet_98056>}

  # GET all the sub diseases of our input disease class
  ?ordo_class_iri rdfs:subClassOf* ?disease_iri;
  rdfs:label ?disease_name.

  SERVICE <http://178.63.49.197:7300/repositories/ordo-catalog-fdp> {
  ?resource a ?type ;
  dcat:theme ?ordo_class_iri;
  dcterm:description ?description;
  dcterm:title ?title;
  dcterm:publisher [dcterm:spatial ?publisher_location];
  dcat:landingPage ?homepage;
  fdo:metadataIdentifier [dcterm:identifier ?id].

  ?publisher_location skos:relatedMatch ?wiki_data_uri. ?wiki_data_uri rdfs:label ?country_name;
  wdt:P297 ?country_code.
  }
}
```

Example of the SPARQL query 1 (Source and copyright 2022: Rajaram Kaliyaperumal - Semantic Web expert with software developer background). Based on the Linked Data principles, the query associates remote resources such as a DCAT catalogue (FAIR DATA POINT), the Orphanet ORDO ontology, wikidata.org to retrieve information on available resources.
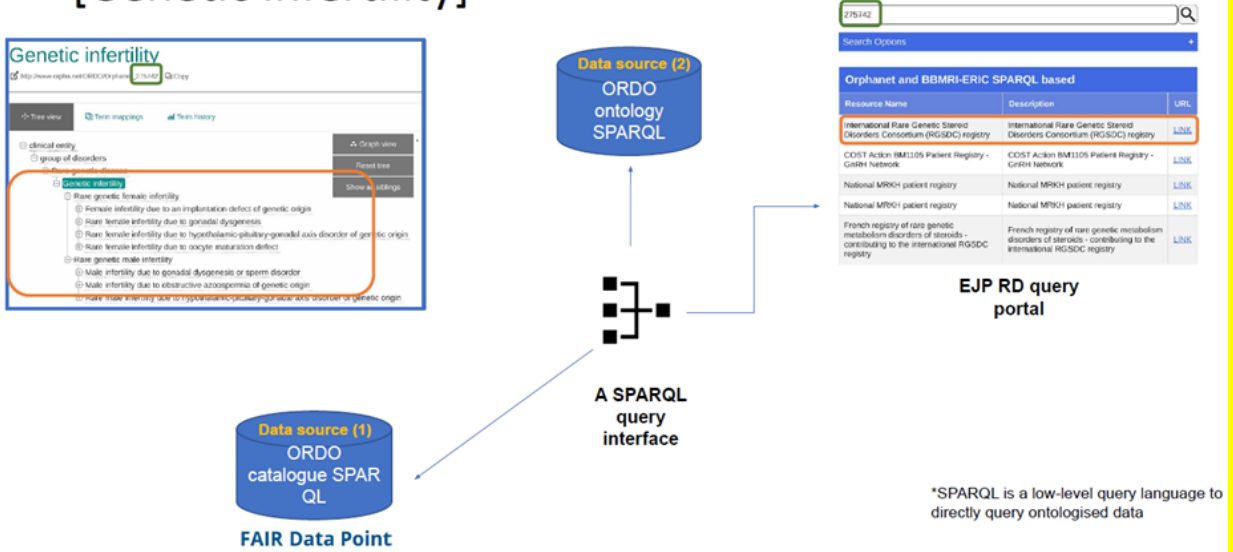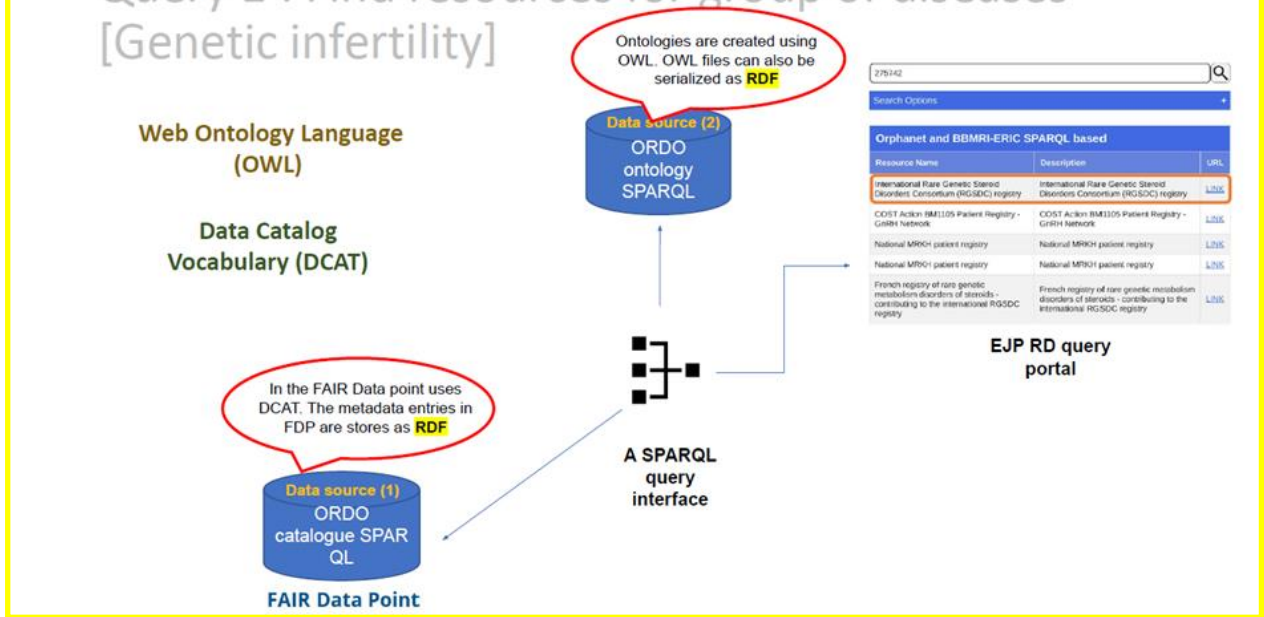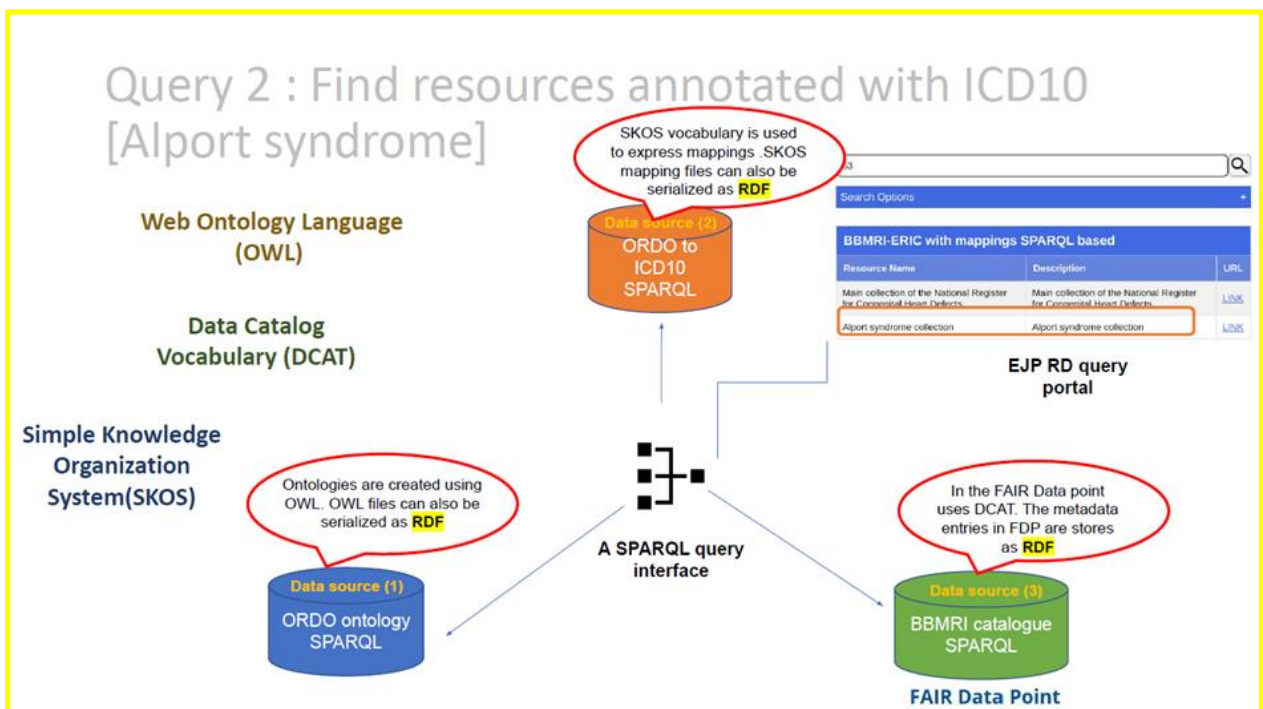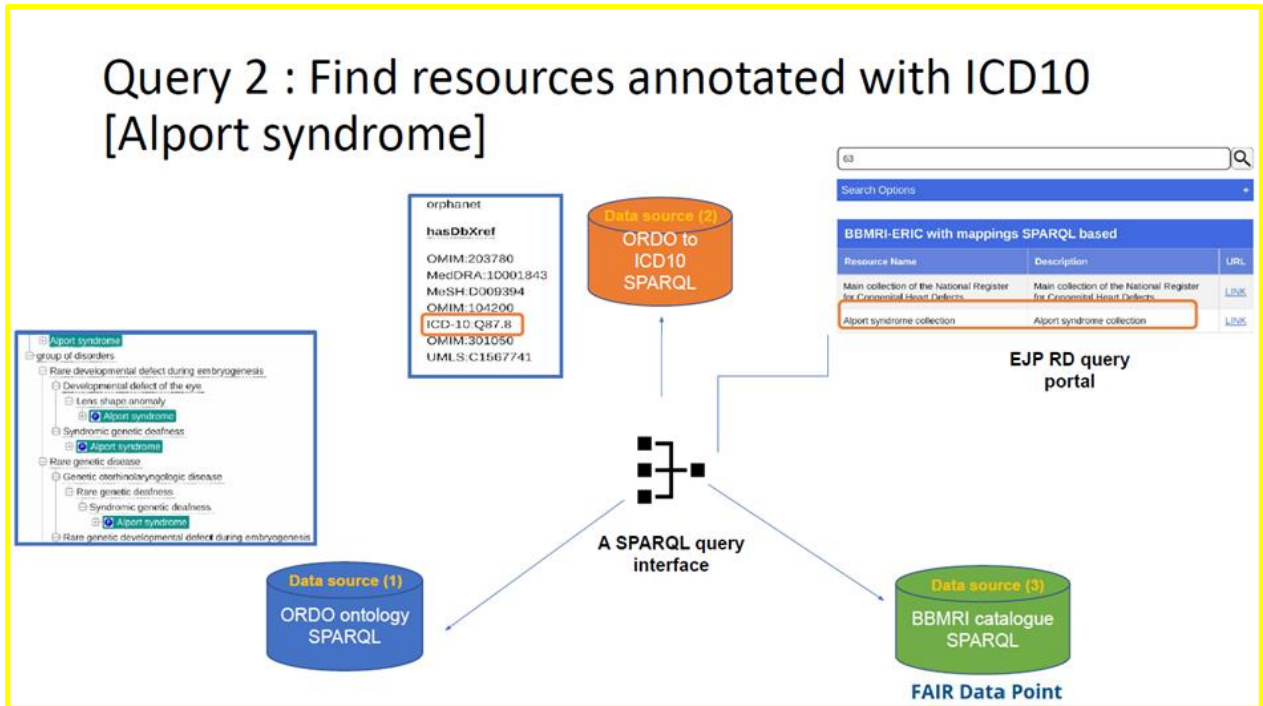
Source: [Metadata and FAIR DATA POINT](#) (Rajaram Kaliyaperumal (Biosemantics group, Leiden))

5.2.3. Faceted search

Faceted search is a technique used in search interfaces, particularly within dataset catalogues like those using the Dataset Catalogue Vocabulary (DCAT) standard. It allows users to filter and refine search results using multiple attributes or categories, known as facets. These facets are usually derived from metadata properties associated with the datasets. Faceted search allows users to efficiently narrow down large volumes of datasets

based on specific criteria. Each facet represents a metadata element that describes the datasets, and users can select one or more facets to refine their search results. For example facets might include: keywords, publisher, themes or categories, temporal coverage, geographic coverage, data formats, licences, … When a user selects a facet, the catalogue dynamically filters the list of available datasets to only those that match the selected criteria.

To make faceted search operational, DCAT-AP provides already a bulk of key metadata elements: Keywords (dct:keyword), Publisher (dct:publisher), Theme (dcat:theme) linked to a controlled vocabulary helping users to find datasets based on subject areas like health, Temporal Coverage (dct:temporal), Geographic Coverage (dct:spatial), Format (dcat:mediaType or dct:format),... Other elements can be used to enhance the faceted search capabilities depending on the specific use case. HealthDCAT-AP has introduced a number of new facets to improve the search user experience.
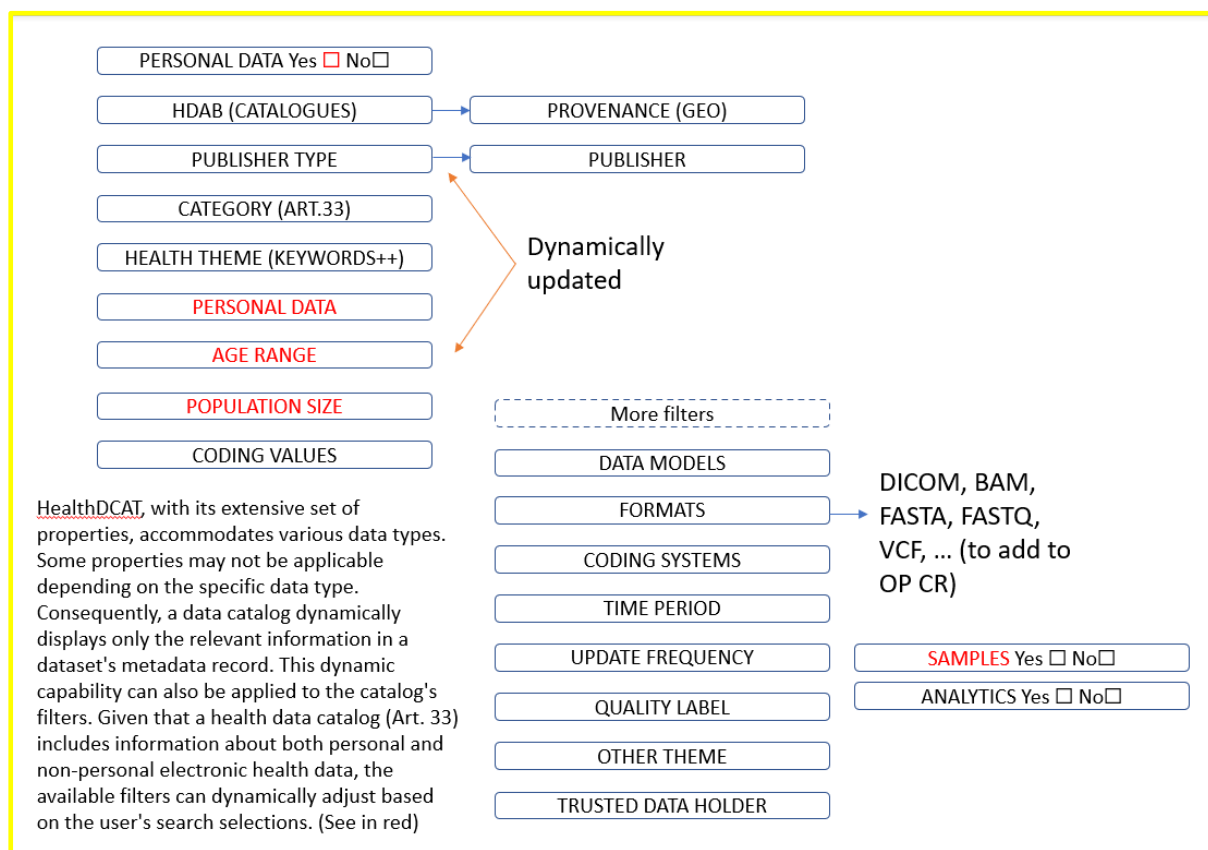
Figure 4: Optimised Filters for HealthDCAT-AP Dataset Catalogues: Dynamic Display Based on Personal Data Inclusion

For instance, a mandatory controlled vocabulary denoting health data within the scope of the Commission Regulation (ref: Art.51 - 17 top health categories) MUST be used in healthdcatap:healthCategory to enable the 'Category' facet. Additional facets specific to HealthDCAT-AP include:

- Health Data Access Bodies,
- Publisher Type,
- Trusted Data Holder,
- Personal Data Classes,
- Age Ranges,

- Population Size,
- Quality Label
- and Coding Values.

These new properties introduced in HealthDCAT-AP significantly enhance the description of health datasets and are designed to improve the data search experience.

Implementation consideration: To implement effective faceted search, metadata must be sufficiently detailed. In this context, the implementing act that will define the minimum metadata elements is essential for the future performance of health dataset catalogues. (i.e., minimum metadata elements for Art.51 and Art.80)

[EHDS Regulation Article 77) Dataset description and dataset catalogue](#)

4. … The Commission shall, by means of implementing acts, set out the minimum ▌ elements health data holders are to provide for datasets and the characteristics of those elements. Those implementing acts shall be adopted in accordance with the examination procedure referred to in Article 98(2).

In addition, the search interface must be designed to present these facets clearly and intuitively to users. It is important to consider UX improvement in two ways: prioritising the facets and dynamically adapting the list of filters.

- **Prioritising the facets:** Not all facets have the same level of relevance, and the pilot use cases have shown that data queries can have both commonalities and specificities (e.g., searching for a theme or a specific dataset format). The filter interface should be designed based on these insights.

- **Dynamically adapting the list of filters:** A health dataset catalogue governed by Article 33 encompasses both personal and non-personal electronic health data. Several metadata elements introduced by Health DCAT-AP, which function as facets, are applicable only to specific types of data (e.g., either personal or non-personal). As a result, the filters can be dynamically adjusted based on the user's search criteria, ensuring that users are presented with the most relevant options. (See Fig. 4: "Optimised Filters for HealthDCAT-AP Dataset Catalogues: Dynamic Display Based on Personal Data Inclusion")

5.2.4. Wikidata as global knowledge hub for the European Health Data Space

The EU Data spaces need their own semantic framework (semantic data model) to achieve consistency across their entire domain and beyond. A semantic framework defines meanings and relationships between the domain concepts. Each concept is identified with a persistent unique identifier. Two datasets sharing the same concept are easily identified because their semantic annotation is consistent across the entire data space.

The common abstractions of a semantic framework are:

Entities represent objects, concepts, or things within a domain. Each entity is a distinct object that can be uniquely identified.
Attributes are properties or characteristics of entities. They provide more information about

an entity.

Relationships describe how entities are connected or related to one another. They help to define the interactions between different entities.

Relation abstractions used in a semantic data model:

- Classification – "instance_of" relations
- Aggregation – "has_a" relations
- Generalisation – "is_a" relations

Classes (or types) are used to categorise entities into groups based on shared characteristics. Each instance of a class is an individual entity.

Hierarchies and taxonomies are used to organise entities into parent-child relationships, creating a structured classification system.

Ontologies are comprehensive frameworks that define the entities, attributes, relationships, and rules within a specific domain. They provide a formal representation of knowledge.

Identifiers are unique values assigned to entities to distinguish them from one another.

There are 2 options to implement an effective harmonised semantic framework for the EU Health Data Space:

Option 1: Implement the semantic framework from scratch using Linked Data technologies

For example, Wikidata and Wikibase are the tools SEMIC has chosen to enable collaboration in creating semantic data models.

Tools: https://joinup.ec.europa.eu/collection/semic-support-centre/wikidata-and-wikibase

Example of domain semantic framework:
https://linkedopendata.eu/wiki/The_EU_Knowledge_Graph

(Resources provided by **Andrea Perego**)

Option 2: Use an existing operational semantic framework to speed up the implementation phase of the data space

The proposed solution for the European Health Data Space and in use in HealthDCAT-AP is to rely on [Wikidata.org] as large-scale, human-readable, machine-readable, multilingual, multidisciplinary, collaborative, centralised, editable, structured, and linked knowledge-base.

The Scientific Publication "Wikidata: A large-scale collaborative ontological medical database"[20] provides rationales about the advantages offered by Wikidata.org as "large-scale semantic framework" and "valuable medical resource" for the EHDS. Additionally, the paper "Wikidata as a knowledge graph for the life sciences"[21] provides further insights into Wikidata's suitability for the EHDS.

---

[20] Houcemeddine Turki, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, Helmi Hamdi, "Wikidata: A large-scale collaborative ontological medical database," Journal of Biomedical Informatics, Volume 99, 2019, 103292, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2019.103292.

[21] Waagmeester A, Stupp G, Burgstaller-Muehlbacher S, Good BM, Griffith M, Griffith OL, Hanspers K, Hermjakob H, Hudson TS, Hybiske K, Keating SM, Manske M, Mayers M, Mietchen D, Mitraka E, Pico AR, Putman T, Riutta A, Queralt-Rosinach N, Schriml LM, Shafee T, Slenter D, Stephan R, Thornton K, Tsueng G, Tu R, Ul-Hasan S, Willighagen E, Wu C, Su AI. Wikidata as a knowledge graph for the life sciences. Elife. 2020 Mar 17;9:e52614. doi: 10.7554/eLife.52614. PMID: 32180547; PMCID: PMC7077981.

"The lack of an integrated and structured version of biomedical knowledge hinders efficient querying or mining of that information, thus preventing the full utilisation of our accumulated scientific knowledge."

"Interoperable: Wikidata items are extensively cross-linked to other biomedical resources using Universal Resource Identifiers (URIs), which unambiguously anchor these concepts in the Linked Open Data cloud (Jacobsen et al., 2018). Wikidata is also available in many standard formats in computer programming and knowledge management, including JSON, XML, and RDF."

GEONAMES: an external semantic resource used in DCAT-AP: Geonames.org is a geographical database that integrates various geographical data sources accessible through various web services, under a Creative Commons attribution 4.0 licence and has a community-driven model. The database contains over 25 million geographical names and consists of 11 million unique features. These features include information such as location names, coordinates, and additional metadata. GeoNames identifiers can be used as persistent URIs in DCAT-AP dct:spatial property to interlink geographical data with other datasets, enhancing the semantic richness of data integration.

Comment: HealthDCAT-AP has introduced several new properties, and DCAT-AP also includes properties where it is beneficial to use Wikidata as a large authority vocabulary. This is because standardising the values of these properties as HTTP URIs is essential, and ensuring that these values come from an authoritative vocabulary enhances consistency:

• healthdcatap:healthTheme (Wikidata URIs)

• healthdcatap:hasCodingSystem (Wikidata URIs)

• healthdcatap:hasCodeValues (Wikidata associated with external standards like ICD-10)

• dct:conformsTo (Wikidata URIs)

5.2.5. Metadata fit for the purpose of semantic annotation and semantic search

The HealthDCAT-AP introduces new properties that rely on ontologies or controlled vocabularies, reinforcing the use of URIs and RDF to ensure consistent description of health datasets, which is crucial for interoperability. These new properties enhance the semantic annotation, ensuring that metadata can accurately describe health datasets. The enhanced semantic annotations in HealthDCAT-AP also support advanced search capabilities, making it easier to retrieve relevant data through more precise SPARQL queries.

New HealthDCAT-AP properties enhancing semantic annotation include:

• healthdcatap:healthCategory (Controlled vocabulary)

• healthdcatap:publisherType (Controlled vocabulary)

• healthdcatap:healthTheme (Wikidata)

• healthdcatap:hasCodingSystem (Wikidata)

• healthdcatap:hasCodeValues (Wikidata associated with external standards like ICD-

10)

- dpv:hasPersonalData (Personal Data Categories (PD) of the DPV ontology)

For the DCAT-AP property dct:conformsTo, the Wikidata URIs MUST also be used.

5.2.6 Sample Distribution

Property	Sample
URI	adms:sample
Range	dct:Distribution
Definition	A sample distribution of the dataset.
Comment	The sample can be anonymised or synthetic subsets that retain the original dataset's essential characteristics without revealing any personal information, or it might solely exhibit the dataset's structure
Usage note	When a health Dataset is categorised as non-public data, implementers MUST provide descriptions for, at least, one sample Distribution of the dataset.

When a health dataset is categorised as non-public data (i.e., protected or sensitive data), one requirement raised during the Technical Working Group sessions was the need to provide descriptions for at least one sample distribution of the dataset. This retained approach involves making a subset of the dataset available using mock-up data in replacement of the sensitive data. This can be an anonymised or synthetic subset. The Population Health Information Research Infrastructure | phiri.eu . This will allow data users to experiment with the data by coding, preparing scripts and analysis, or training ML/AI models. Additionally, the sample distribution must be accompanied by a data dictionary, providing definitions and descriptions of all data elements.

The Population Health Information Research Infrastructure | phiri.eu PHIRI Tacks 7.5 "upgrading options" (Project Horizon 2020 - grant agreement No 101018317)

> **Populating the digital twin with mock-up data**
> Mock-up data is a type of synthetic data that is designed to mimic the characteristics of real-world data without containing any sensitive information.
>
> By using techniques such as data sampling, transformation, generation, and augmentation, it is possible to create mock-up data that closely resembles the original dataset while protecting any sensitive information that may be present. Depending of the goals of the digital twin model and legal constraints to process sensitive data, one or another techniques can be considered:

Data Sampling: One approach to creating mock-up data is to randomly sample data from the original dataset, using techniques such as random sampling or stratified sampling . This can help to preserve the statistical properties of the original dataset while eliminating any sensitive information (e.g.: the "avatar" method).

Data Transformation: Another approach to creating mock-up data is to transform the original data in some way, such as by scaling or re-scaling the values, adding noise, or applying mathematical functions. This can help to preserve the structure of the original data while obscuring any sensitive information.

Data Generation: In some cases, it may be necessary to generate entirely new data in order to create mock-up data. This can be done using techniques such as Data profiling which can capture statistical properties of a dataset and aim to create high-quality, realistic mock-up data. Data profiling refers to the analysis of information for use in a data warehouse in order to clarify the structure, content, relationships, and derivation rules of the data . Data profiles are obtained from an attribute–value system such as a flat file or a spreadsheet (rectangular data) . For instance, in case of a relational database, a SQL query extracts information and exports it as a csv file. A data profile can be produced from the csv file and serves as a reference document to generate mock-up data.
More advanced techniques like generative adversarial networks (GANs) can generate new data that very closely resembles it.

Data Augmentation: Data augmentation involves adding new data points to the existing dataset, either by duplicating existing data points or by creating new data points through some other means. This can help to increase the size of the dataset and improve the accuracy of the digital twin model.

As already mentioned, all these techniques for creating mock-up data requires careful consideration of the legal framework in force in silos for the sharing and processing of sensitive data. There are various legal constraints (e.g.: GDPR) from one silo to another and data profiling as it guarantees privacy in an unequivocal way is the most appropriate technique to implement the population's digital twin. However, any other technique may be used. It may have the advantage of including patterns in the mock data that enrich the analysis. It will then be important to specify in the digital twin the quality of the mock-up data.

Simplifying data access negotiations
As sensitive personal health data is never directly accessed, the Population Health Digital Twin offers the possibility to overpass the pitfall of long and complex administrative data access negotiations with the data permit authorities as no individual personal data is directly accessed.

In analysing the required capabilities of the future EU central health dataset catalogues, two key requirements have been identified:

1- Variable listing in Data Permit/Access Applications: Data users submitting a data permit/access application form must have the option to list the specific variables they wish to access for reuse. They are not intended to access all variables according to the Data minimisation principle (ref: EHDS Regulation - Article 66 "Data minimisation and purpose

limitation").

EHDS Regulation Article 68) Data permit

1. For the purposes of granting access to electronic health data, the health data access bodies shall assess whether all the following criteria are fulfilled:
...
(c) the processing complies with Article 6(1) of Regulation (EU) 2016/679 and, in the case of pseudonymised data, there is sufficient justification that the purpose cannot be achieved with anonymised data;

2- Analysis script submission with Data Request Applications: When requesting a data request, users must be able to attach an analysis script which is meant to produce the anonymised statistical data.

EHDS Regulation Article 69) Health data request

1. The health data applicant may submit a health data request for the purposes referred to in Article 53 with the aim of obtaining a response only in an anonymised statistical format. A health data access body shall not provide a response to a health data request in any other format and the health data user shall have no access to the electronic health data used to provide that response.

5.2.6.1. The challenge of producing "harmonised" data dictionaries

Article 33 of the Data Act mandates that the healthData@EU infrastructure must unequivocally comply with this regulation to function effectively as a Data Space. This includes the requirement to "describe in a publicly available and consistent manner" the data structures.

Publishing data structures, formats, and vocabularies consistently within the European Health Data Space (EHDS) is crucial for handling sensitive health data. Such standardisation facilitates interoperability and ensures uniform data interpretation across different systems. For a clinician or researcher accessing health data across different Member States, harmonised data dictionaries ensure that common terms like 'sex,' 'weight,' or 'blood pressure' are consistently defined, improving their ability to work with this data seamlessly, no matter where it was generated

Data Act Article 33

Essential requirements regarding interoperability of data, of data sharing mechanisms and services, as well as of common European data spaces
1. Participants in data spaces that offer data or data services to other participants shall comply with the following essential requirements to facilitate the interoperability of data, of data sharing mechanisms and services, ...
(b) the data structures, data formats, vocabularies, classification schemes, taxonomies and code lists, where available, shall be described in a publicly available and consistent manner;

Iate: Data Structure definition

> *A set of structural metadata associated with a data set, which includes information about how concepts are associated with the measures, dimensions, and attributes of a hypercube, along with information about the representation of data and related descriptive metadata.*

In HealthDCAT-AP, this can be achieved by publishing at least one sample Distribution, accompanied by a data dictionary.

The proposed solution made to the Technical Working Group (TWG) to publish and harmonise the production of data dictionaries involves utilising the "Tabular Data on the Web" (CSVW) vocabulary. CSVW allows the annotation of tabular data (e.g., CSV files) with metadata that describes the data structure, types, formats, and relationships. CSVW provides namespace vocabulary terms and definitions specifically for tabular data, making it simple to implement. This metadata effectively serves as a data dictionary, providing comprehensive information about the dataset. See example:

EXAMPLE 1

Recommended use of CSVW terms for RDF-izing variable descriptions

```
<LINKVACC> a csvw:TableGroup ;
csvw:table <STATBEL>, <VACCINET+>, <COBRHA> .

<STATBEL> a csvw:Table ;
dcterms:title "Statbel"@en ;
dcat:keyword "Statbel"@en ;
csvw:url <http://example.org/tree-ops-ext.csv> ;
csvw:column [
a csvw:Column ;
csvw:name "CD_RN_STATUS" ;
csvw:datatype xsd:float ;
dcterms:description "Patient status: Whether patient is deceased, has migrated or is de-registred
], [
a csvw:Column ;
csvw:name "CD_COD_COVID" ;
csvw:datatype xsd:string ;
dcterms:description "COVID-19 specific death Patients deceased from COVID-19"@en ;
], [
a csvw:Column ;
csvw:name "CD_EDU" ;
csvw:datatype xsd:string ;
dcterms:description "Patient educational level using ISCED classification"@en ;
], [
a csvw:Column ;
csvw:name "HH_TYPE_LIPRO" ;
csvw:datatype xsd:string ;
dcterms:description "Patient household status"@en ;
] .
```

Figure 5: Draft version of the HealthDCAT-AP (RDF examples)

W3C Tabular Data on the Web: https://www.w3.org/TR/tabular-data-primer/

W3C CSV on the Web: https://www.w3.org/TR/tabular-data-model/

By effectively RDF-izing variable descriptions using the CSVW vocabulary, the data dictionary can be presented as RDF in a sample Distribution. This ensures alignment with the RDF framework of DCAT facilitating better data interoperability and usability which is expected for supporting the data access application service of the EU health dataset catalogue.

To summarise, HealthDCAT-AP requires data holders to provide a sample distribution of the dataset (e.g., mock-up data, anonymised data, synthetic data, etc.) in any computer-readable format (e.g., CSV, JSON). If applicable, a data dictionary should also be published. The data dictionary must be published using CSVW, resulting in an RDF format for the sample distribution.

A more complex use case involves merging both requirements by simultaneously producing the dataset sample as tabular data along with its data dictionary using CSVW:

Dataset example: Health Phenotype Data

Patient_ID	Time_Point (days)	Geographic_Location (Latitude, Longitude)	Height (cm)	Weight (kg)	Blood_Pressure (mmHg)
P001	0	(40.7128, -74.0060)	170	70	120/80
P002	7	(34.0522, -118.2437)	165	65	115/75
P003	14	(51.5074, -0.1278)	180	80	130/85

Dimensions of the dataset:

- **Time_Point**: Measured in days from the first recorded observation.
- **Geographic_Location**: Latitude and Longitude coordinates (WGS84 (World Geodetic System 1984)).
- **Height**: Recorded in centimeters (cm).
- **Weight**: Recorded in kilograms (kg).
- **Blood Pressure**: Measured in millimeters of mercury (mmHg).

To fully understand a dataset, both the dimensions and their semantics are essential (e.g.: data types and taxonomies). The CSVW vocabulary addresses these requirements which are not covered by DCAT-AP. Alternative solutions like DataCube are also available.

Data structure and semantic - RDF representation compliant to CSVW	
JSON-LD serialisation	TURTLE serialisation
{	@prefix csvw: <http://www.w3.org/ns/csvw#> .

```
"@context": "http://www.w3.org/ns/csvw",
"url": "patients_data.csv",
"tableSchema": {
  "columns": [
    {
      "name": "Patient_ID",
      "titles": "Patient_ID",
      "datatype": "string",
      "propertyUrl": "http://example.org/vocab#PatientID"
    },
    {
      "name": "Time_Point",
      "titles": "Time_Point (days)",
      "datatype": "integer",
      "propertyUrl": "http://example.org/vocab#TimePoint"
    },
    {
      "name": "Geographic_Location",
      "titles": "Geographic_Location (Latitude, Longitude)",
      "datatype": "string",
      "propertyUrl":
"http://www.w3.org/2003/01/geo/wgs84_pos#location",
      "separator": ","
    },
    {
      "name": "Height",
      "titles": "Height (cm)",
      "datatype": "integer",
      "unit": "http://qudt.org/vocab/unit#Centimeter",
      "propertyUrl": "http://example.org/vocab#Height"
    },
    {
      "name": "Weight",
      "titles": "Weight (kg)",
      "datatype": "integer",
      "unit": "http://qudt.org/vocab/unit#Kilogram",
      "propertyUrl": "http://example.org/vocab#Weight"
    },
    {
      "name": "Blood_Pressure",
      "titles": "Blood_Pressure (mmHg)",
      "datatype": "string",
      "propertyUrl": "http://example.org/vocab#BloodPressure"
    }
  ],
  "primaryKey": "Patient_ID"
}
}
```

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix ex: <http://example.org/vocab#> .
@prefix qudt: <http://qudt.org/vocab/unit#> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .

<#Table> a csvw:Table ;
   csvw:url "patients_data.csv" ;
   csvw:tableSchema [
      a csvw:Schema ;
      csvw:columns (
         [
            csvw:name "Patient_ID" ;
            csvw:titles "Patient_ID" ;
            csvw:datatype "string" ;
            csvw:propertyUrl ex:PatientID
         ]
         [
            csvw:name "Time_Point" ;
            csvw:titles "Time_Point (days)" ;
            csvw:datatype "integer" ;
            csvw:propertyUrl ex:TimePoint
         ]
         [
            csvw:name "Geographic_Location" ;
            csvw:titles "Geographic_Location (Latitude, Longitude)" ;
            csvw:datatype "string" ;
            csvw:propertyUrl geo:location ;
            csvw:separator ","
         ]
         [
            csvw:name "Height" ;
            csvw:titles "Height (cm)" ;
            csvw:datatype "integer" ;
            csvw:propertyUrl ex:Height ;
            csvw:unit qudt:Centimeter
         ]
         [
            csvw:name "Weight" ;
            csvw:titles "Weight (kg)" ;
            csvw:datatype "integer" ;
            csvw:propertyUrl ex:Weight ;
            csvw:unit qudt:Kilogram
         ]
         [
            csvw:name "Blood_Pressure" ;
            csvw:titles "Blood_Pressure (mmHg)" ;
            csvw:datatype "string" ;
            csvw:propertyUrl ex:BloodPressure
         ]
      ) ;
      csvw:primaryKey "Patient_ID"
   ] .
```

Dataset sample as RDF Representation compliant to CSVW

JSON-LD serialisation	TURTLE serialisation

```
{
  "@context": {
    "ex": "http://example.org/vocab#",
    "geo": "http://www.w3.org/2003/01/geo/wgs84_pos#",
    "qudt": "http://qudt.org/vocab/unit#"
  },
  "@graph": [
    {
      "@id": "ex:P001",
      "@type": "ex:Patient",
      "ex:PatientID": "P001",
      "ex:TimePoint": 0,
      "geo:location": {
        "geo:lat": "40.7128",
        "geo:long": "-74.0060"
      },
      "ex:Height": {
        "qudt:unit": "qudt:Centimeter",
```

```
@prefix ex: <http://example.org/vocab#> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix qudt: <http://qudt.org/vocab/unit#> .

# Patient P001
ex:P001 a ex:Patient ;
   ex:PatientID "P001" ;
   ex:TimePoint 0 ;
   geo:location [
      geo:lat "40.7128" ;
      geo:long "-74.0060"
   ] ;
   ex:Height [
      qudt:unit qudt:Centimeter ;
      qudt:value 170
   ] ;
   ex:Weight [
      qudt:unit qudt:Kilogram ;
```

```
    "qudt:value": 170
  },
  "ex:Weight": {
    "qudt:unit": "qudt:Kilogram",
    "qudt:value": 70
  },
  "ex:BloodPressure": "120/80"
},
{
  "@id": "ex:P002",
  "@type": "ex:Patient",
  "ex:PatientID": "P002",
  "ex:TimePoint": 7,
  "geo:location": {
    "geo:lat": "34.0522",
    "geo:long": "-118.2437"
  },
  "ex:Height": {
    "qudt:unit": "qudt:Centimeter",
    "qudt:value": 165
  },
  "ex:Weight": {
    "qudt:unit": "qudt:Kilogram",
    "qudt:value": 65
  },
  "ex:BloodPressure": "115/75"
},
{
  "@id": "ex:P003",
  "@type": "ex:Patient",
  "ex:PatientID": "P003",
  "ex:TimePoint": 14,
  "geo:location": {
    "geo:lat": "51.5074",
    "geo:long": "-0.1278"
  },
  "ex:Height": {
    "qudt:unit": "qudt:Centimeter",
    "qudt:value": 180
  },
  "ex:Weight": {
    "qudt:unit": "qudt:Kilogram",
    "qudt:value": 80
  },
  "ex:BloodPressure": "130/85"
}
]
}
```

```
    qudt:value 70
  ] ;
  ex:BloodPressure "120/80" .

# Patient P002
ex:P002 a ex:Patient ;
  ex:PatientID "P002" ;
  ex:TimePoint 7 ;
  geo:location [
    geo:lat "34.0522" ;
    geo:long "-118.2437"
  ] ;
  ex:Height [
    qudt:unit qudt:Centimeter ;
    qudt:value 165
  ] ;
  ex:Weight [
    qudt:unit qudt:Kilogram ;
    qudt:value 65
  ] ;
  ex:BloodPressure "115/75" .

# Patient P003
ex:P003 a ex:Patient ;
  ex:PatientID "P003" ;
  ex:TimePoint 14 ;
  geo:location [
    geo:lat "51.5074" ;
    geo:long "-0.1278"
  ] ;
  ex:Height [
    qudt:unit qudt:Centimeter ;
    qudt:value 180
  ] ;
  ex:Weight [
    qudt:unit qudt:Kilogram ;
    qudt:value 80
  ] ;
  ex:BloodPressure "130/85" .
```

Real-world example: The dataset "Legati: Misiones diplomáticas hispanomusulmanas (1492-1708)" on the EU Data Portal serves as an illustrative case to demonstrate the advantages of integrating the CSVW vocabulary into HealthDCAT-AP.

The dataset description includes three distributions in various formats, providing a useful example for discussion:

- 1/ LEGATI.v2.xlsx
- 2/ Legati_Readme.txt
- 3/ Normas.pdf

The first distribution, in XLSX format, contains the actual data. This data can be automatically displayed on the EU Data Portal using a CSV reader and the dataset's download link:

Preview of the xlsx distribution provided for the dataset "Legati: misiones diplomáticas hispanomusulmanas (1492-1708)" LEGATI.v2.xlsx

A similar tool could be developed to display an adms:sample distribution, in CSVW format, in any health dataset catalogues.

The second distribution, in TXT format, is a README file that contains no data and should not be listed as a dataset distribution. However, it is notable that the file includes structural metadata that could be provided through a CSVW distribution:

Legati_Readme.txt

…
DATA-SPECIFIC INFORMATION:
1. Number of variables: 23
2. Number of cases/rows: 61
3. Variable List:
 - Número / number (Código interno identificando cada misión / Internal code identifying each mission)
 - Nombre del enviado / Agent's name
 - Otros nombres / Other names (alias y nombres en otros idiomas de los agentes / agents' nicknames and alternative names in other languages)
 - Biografía del enviado / agent's biography
…
4. Missing data codes:

5. Specialized formats or other abbreviations used:
- AGS Archivo General de Simancas
  o CC: Cámara de Castilla
  o E: Estado
…

6. Dictionaries/codebooks used:

7. Controlled vocabularies/ontologies used:

README txt file provided as a distribution for the dataset "Legati: misiones diplomáticas hispanomusulmanas (1492-1708)"

The third distribution, a PDF containing data entry guidelines, represents the dataset's data model. Although this PDF is not a true data distribution, it could be formalized as an entity-relationship diagram or table structure, which CSVW could standardise as a more structured solution.

In summary, we aim to standardise sample distributions for health datasets and incorporate them into the HealthDCAT Application Profile to ensure interoperability in publishing sample datasets and data dictionaries. This principle of harmonised data specifications for health data samples and data dictionaries can be effectively realised using the CSVW vocabulary.

5.2.6.2. The challenge of defining datasets: "How to break down large data warehouses into logical datasets?"

Many discussions during the design of the HealthDCAT-AP focused on the "dataset" definition. For health data sources that comprise a large data warehouse, this proves to be a particularly challenging exercise. How to decide which datasets are suitable for publishing? How to break down large health registries into logical or "virtual" datasets that are relevant for the secondary use? The European Health Data Space (EHDS) Recital paragraph Alinea (56) provides a definition of a dataset that guides this process:

EHDS Regulation Recital 56)

> *… Electronic health data for secondary use* ▌ *should* ▌ *be made*
> *available preferably in a structured electronic format that facilitates their*
> *processing by computer systems. Examples of structured electronic formats include*
> *records in a relational database, XML documents or CSV files and free text,*
> *audios, videos and images provided as computer-readable files.*

A dataset is, in essence, a structured collection of data ready for analysis, while a data source is the provider of that data (e.g.: A data warehouse or health registry). The Oxford Dictionary defines dataset as "a collection of data, typically in tabular form, especially one that can be accessed and manipulated by computer software". In summary, the data source supplies the raw data, which is then processed and organised into datasets for various analytical purposes.

During the exercise of mapping the national metadata model of the French Health Data Hub to HealthDCAT-AP (Task 9.4), we extensively discussed the associated challenges. In order to generate HealthDCAT-AP metadata records, the agreed-upon approach was to define coherent data collections from the data warehouses. This strategy ensures that datasets from the data warehouse are consistently prepared and structured, making them ready for analysis.

In the health data landscape, it is a common situation to find data providers such as registries, biobanks curating data and storing data in a data warehouse. These organisations will face the challenge to define coherent data collections for the EHDS:

The challenge of defining datasets from an IT technical perspective: Many health data providers curate their data within the context of data warehousing. Data marts, as subsets of these data warehouses, are specifically designed to address the unique needs of

individual departments within an organisation.

Here's a more detailed explanation:

State of the art of data management and business intelligence:

A **Data Warehouse** is a centralised repository that stores integrated data from multiple sources. It is designed to support business analysis and decision-making processes. The data warehouse contains historical data and is optimised for query and analysis. It typically covers a wide range of subject areas.

The **Data Mart** is a smaller, more focused subset of a data warehouse designed to meet the specific needs of a particular department, team, or business function, such as a research project. It contains data relevant to a particular subject area or line of business.

How Data Warehouse and data mart work together

1.   Data Collection:

        o   Data from various operational systems and external sources are collected and integrated into the data warehouse.

2.   Data Storage:

        o   The data warehouse stores large volumes of data, often spanning multiple subject areas and time periods.

3.   Data Access via Data Marts:

        o   Data marts are created to provide more accessible, department-specific views of the data stored in the data warehouse.

        o   These data marts can be optimised for performance and tailored to the specific analytical needs of the business unit they serve.

4.   Usage:

        o   Users within specific departments can access data marts to perform analysis, generate reports, and make data-driven decisions.

        o   This reduces the complexity for end users, as they interact with a more focused subset of the overall data.

By making data accessible in data marts, organisations can ensure that relevant and actionable insights are more readily available to specific departments, improving the efficiency and effectiveness of their business processes.
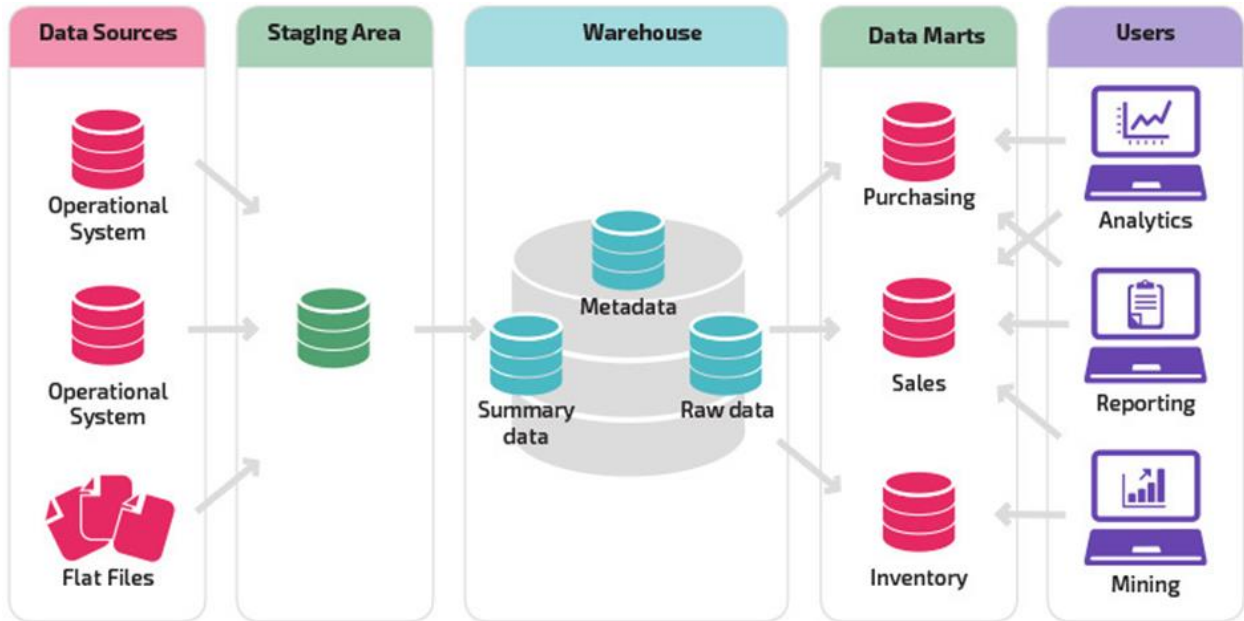
Figure 6: Data mart vs data warehouse

Within the framework of the HealthData@EU infrastructure, a recommended "user-centric" approach for data warehouse holders is to create "virtual" data marts. Each "virtual" data mart would define a coherent dataset. A HealthDCAT-AP record can then be generated and maintained for each data mart. Additionally, a data dictionary containing definitions and descriptions of all data elements for the "virtual" data mart is produced. This approach enhances data discovery and aligns with a more service-oriented strategy. It supports the EHDS user journey by allowing users to request access for data in a Secure Processing Environment (SPE), which can be considered a "virtual" secured data mart.

In Finland, this approach has already been adopted. The catalogue lists a collection of data resources, with each resource offering a set of accessible datasets:

Datasets

Variables are divided in 3 datasets

HUSLAB Laboratory, main results, Multilab	HUSLAB Laboratory additional research data, Multilab	HUSLAB Laboratory, microbial results, Multilab
33 Variable ▸ GO TO DATASET	13 Variable ▸ GO TO DATASET	17 Variable ▸ GO TO DATASET

Figure 7: https://aineistokatalogi.fi/catalog/studies/7167ba11-87f3-4ef4-9142-c11583be5c74

Alternatively, the feasibility of creating a single healthDCAT metadata record for an entire data warehouse can be questioned. A data warehouse is not a dataset; it lacks the

granularity necessary to effectively describe the diverse datasets it contains. A single metadata record would likely be too broad and general, making it difficult for users to discover the specific datasets they need. Instead, approaching datasets as products - each with its own healthDCAT metadata record - would offer a more granular and user-friendly solution. This would allow data users to better understand and access specific datasets tailored to their needs, ensuring they can effectively use the data. Treating datasets as individual products also enables more precise metadata, enhancing data discoverability and usability.

Few resources and studies are covering this topic: "How can Data Catalog Vocabulary (DCAT) be used to address the needs of databases?"

Other resource by the same Author: https://www.w3.org/2016/11/sdsvoc/beata

Below is a figure that shows the hierarchical structure of the Norwegian Cause Of Death Registry:

- The registry is classified as a National Health Registry, a "data source"
- It's broken down to a "collection" named "DÅR Analyse". Most of the data sources consists of several collections (su registries)
- The data sources and collections have classified the variables into themes. Unfortunately, this classification is not standardised across data sources and remains proprietary to each one.
- Each collection contains multiple variables, with some variables often being the same across different collections. In some cases, key variables are separated into a master collection.
- For each variable, you can find information about the associated code list, such as its classification, terminology, or value set, including the standard it is based on.
- Additionally, you will find information about the codes, including both the code itself and its corresponding text. This information is essential, as it is used not only to specify the data a user applies for but also to ensure semantic interoperability when requesting data from multiple sources.
- Additionally, some data sources provide statistics at the variable level, offering users valuable insights into both data quality and coding practices. Moreover, most data sources make these statistics available as open data (Ref: 5.2.7 Analytics - HealthDCAT-AP model includes a property of class Distribution to associate a dataset to its statistics).

In this context, it's not always intuitive to determine what qualifies as a dataset and what does not. In future processes, it will be important to establish a common understanding and develop a method to embed these breakdown structures within dataset descriptions.
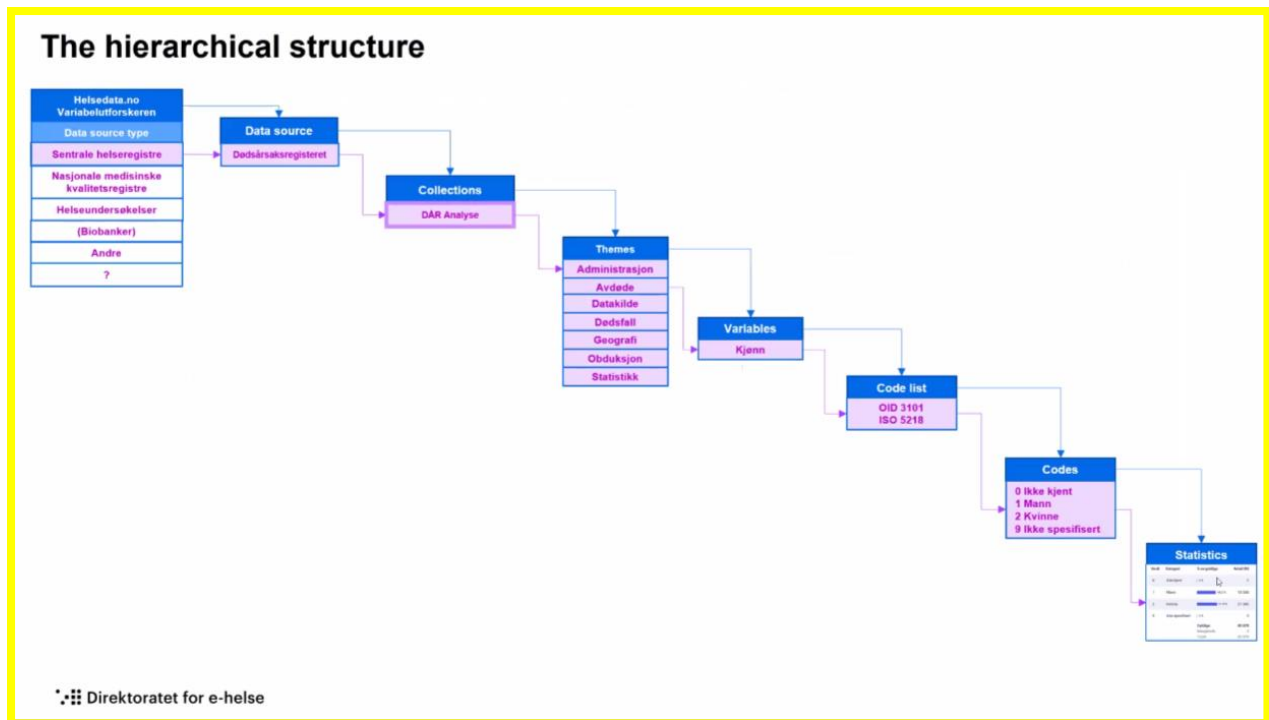
Figure 8: Example of a data warehouse's structure: the Norwegian Cause of Death Registry

5.2.6.2.1 Common Harmonised datasets for federated analysis and learning

Given the federated nature of the healthDCAT@EU infrastructure, the concept of "virtual" datasets, as discussed above, can be considered across the EHDS. Through HealthDCAT-AP, common harmonised data models and datasets could be promoted and accurately described. Figure 9 presents a graphical representation of the EHDS components supporting secondary data use, divided into two main categories: on the left, the provision of relevant health data for health data holders and Health Data Access Bodies, and on the right, the consumption of relevant data for health data users.

The concept of this figure comes from the JRC Science for Policy report INSPIRE – A public sector contribution to the European Green Deal data space (Fig. 11, p. 36). Additional components have been added, illustrating the data flow to PROCESSING SERVICES (Secure Processing Environments), common DATA ACCESS APPLICATION forms and clarifying that the EHDS does not required raw data to be harmonised before publication. While the EHDS does not required and generalised data harmonisation, this does not preclude organisations, such as research infrastructures, from coordinating and agreeing on common data models and virtual datasets for the EHDS. This approach supports federated analysis and learning by making the data ready for processing. (Note: Interoperability requirements for Processing Services essential for federated analysis and learning fall outside the scope of this document.)

How can such a virtual, federated 'dataset' be promoted? When a dataset conforms to a specific data model or standard, the dct:conformsTo property in HealthDCAT-AP should ideally reference a URI. For example, a dataset using the OMOP data model could use the URI https://www.wikidata.org/wiki/Q47219554 for OMOP. Additionally, a common HARMONISED DATA VALIDATOR could be implemented as a new component of the HealthDATA@EU infrastructure to ensure the quality and consistency of datasets.
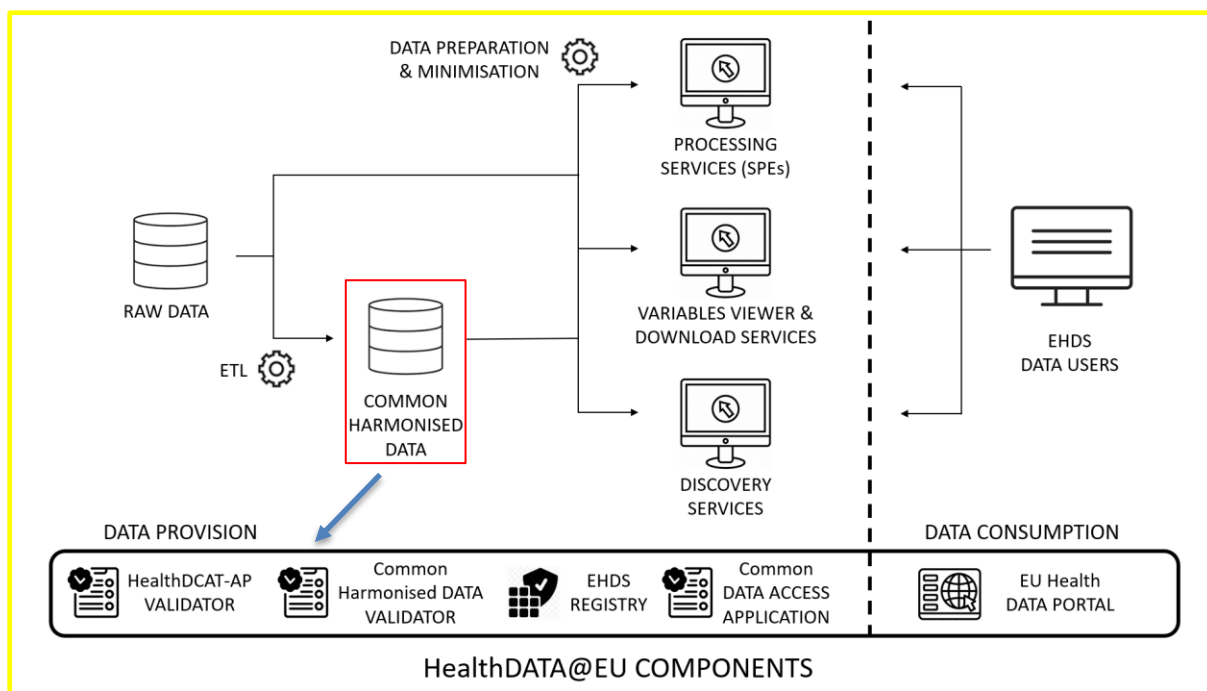
Figure 9: Overview of the components of the HealthDCAT@EU infrastructure.

Real-world example: DARWIN EU (Data Analysis and Real World Interrogation Network)[22] is a European initiative established by the European Medicines Agency (EMA) to harness real-world healthcare data for regulatory decision-making. Its primary objective is to provide timely and reliable evidence on the use, safety, and effectiveness of medicines across the European Union (EU).

The network comprises a coordination centre and a growing consortium of data partners, including hospitals, primary care providers, health insurance organisations, patient registries, and biobanks. These partners contribute anonymised healthcare data, which is standardized into a **common data model** to facilitate efficient analysis.

5.2.6.3. Intellectual Property Rights (IPR)

EHDS Regulation Article 52) Intellectual property rights and trade secrets

> *2. Health data holders shall inform the health data access body of any electronic*
>
> *health data containing content or information protected by intellectual property*
>
> *rights, trade secrets or covered by the regulatory data protection right laid down in*
>
> *Article 10(1) of Directive 2001/83/EC or Article 14(11) of Regulation (EC)*
>
> *726/2004.*
>
> *Health data holders shall identify which parts of the datasets are*
>
> *concerned and justify the need for the specific protection of the data. Health data*
>
> *holders shall provide that information when communicating to the health data*
>
> *access body the description of the dataset they hold pursuant to Article 60(3) of this*

---

[22] https://darwin-eu.org/

> *Regulation or, at the latest, following a request received from the health data*
>
> *access body.*

To comply with the requirements of the Data Governance Act, HealthDCAT-AP has made the 'Conditions for re-use (Rights)' a mandatory property using dct:rights, a dct:RightsStatement. This property refers to 'a statement that specifies rights associated with the Dataset Distribution,' allowing data holders to provide the necessary information to meet the Intellectual Property Rights assertion requirements under the EHDS Regulation. This requirement also applies to sample Distributions. However, due to IPR constraints, if any electronic health data contains content protected by Intellectual Property Rights, it may prevent the provision of a dataset subset. In such cases, it may still be possible to produce a redacted data dictionary, with health data holders indicating which parts of the datasets are affected by IPRs and justifying why specific protection is required. A sample distribution providing a redacted data dictionary and conditions for re-use would be beneficial to ensure dataset discoverability. Therefore, even when IPRs apply, a data dictionary should remain mandatory for HealthDCAT-AP to enhance data discoverability.

Real-world example: Publishing redacted data - where sensitive or confidential information is edited or censored - is a common practice before sharing or publishing. A notable example comes from biodiversity data management, especially within platforms like the Global Biodiversity Information Facility (GBIF) and similar environmental data repositories. In these cases, datasets containing sensitive information, such as precise locations of endangered species, are redacted to address conservation and ethical concerns. Only non-sensitive metadata and summary statistics, like species counts and general habitat types, are shared, while specific geographic coordinates and other details that could risk exposing protected locations are excluded.

Publishing redacted data is also common in marine data management, as seen with some of the European Marine Observation and Data Network (EMODnet) data products. For example, datasets containing sensitive information about submarine power and telecommunication cables are often redacted to protect infrastructure security and integrity. While general information, such as the presence of cables and broad geographical areas, may be shared, the exact coordinates and detailed pathways are withheld. This raw, precise information remains under the copyright of the companies that own or operate the cables, ensuring both data security and compliance with proprietary rights.

5.2.7. Analytics (healthdcatap:analytics)

HealthDCAT-AP introduces a new property modelled as Distribution to link to visualisations, analytics services (e.g., dashboards), technical reports (e.g., dataset metrics), quality and usability indicators, querying services (e.g., Beacon API), and more. This property allows access to or requests for associated resources, which facilitate data discoverability and understanding without directly accessing the underlying data. This approach is beneficial for protecting sensitive information and managing data that is too large or complex for direct access.

Data holders can include HTTP URIs to publicly available dashboards, summary reports, or links to APIs that allow users to query high-level statistics on the dataset. This not only improves transparency but also facilitates decision-making without requiring access to the

raw data.

Property	Analytics
URI	healthdcatap:analytics
Range	dct:Distribution
Definition	An analytics distribution of the dataset.
Usage note	Publishers are encouraged to provide URLs pointing to API endpoints or document repositories where users can access or request associated resources such as technical reports of the dataset, quality measurements, usability indicators,... or analytics services.

HealthDCAT-AP Analytics Distribution - three examples of Dashboard:

Example 1: The Surveillance Atlas of Infectious Diseases published by ECDC is a healthdcatap:analytics Distribution of the sensitive data that the EU agency is collecting from the EU Member States.
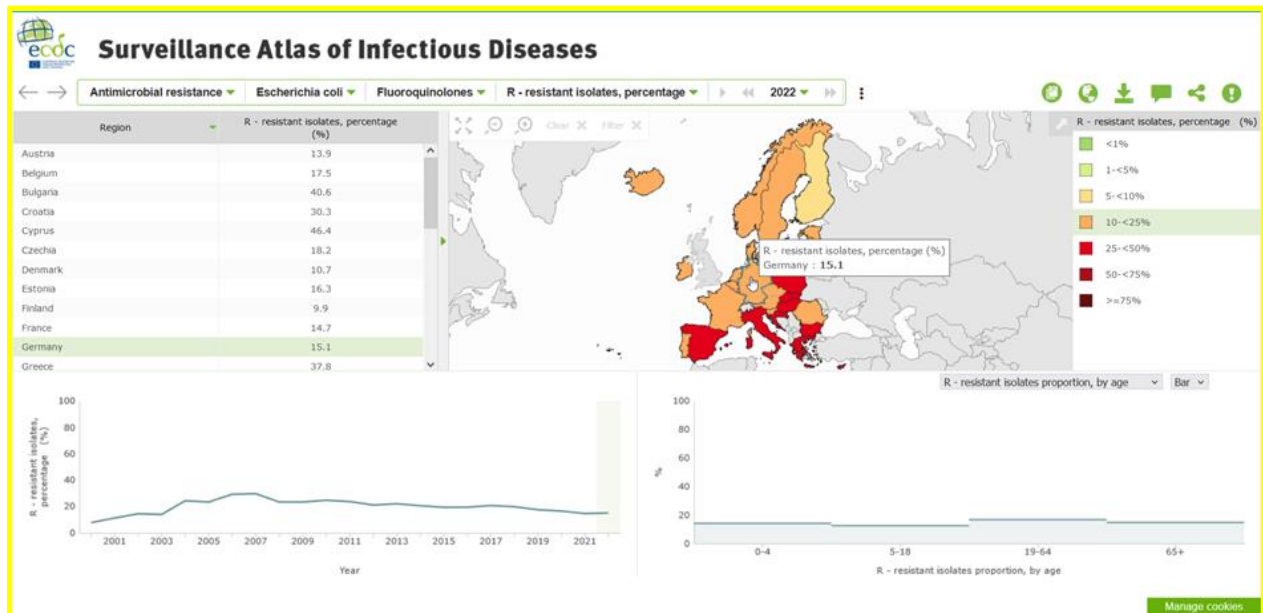
Example 2: the Austrian National Cancer Registry dashboard provides information on the different types of cancer, focusing on the organ and tissue type affected by cancer. The basis is the Austrian National Cancer Registry maintained by Statistics Austria, in which cases diagnosed or treated in hospitals are registered.
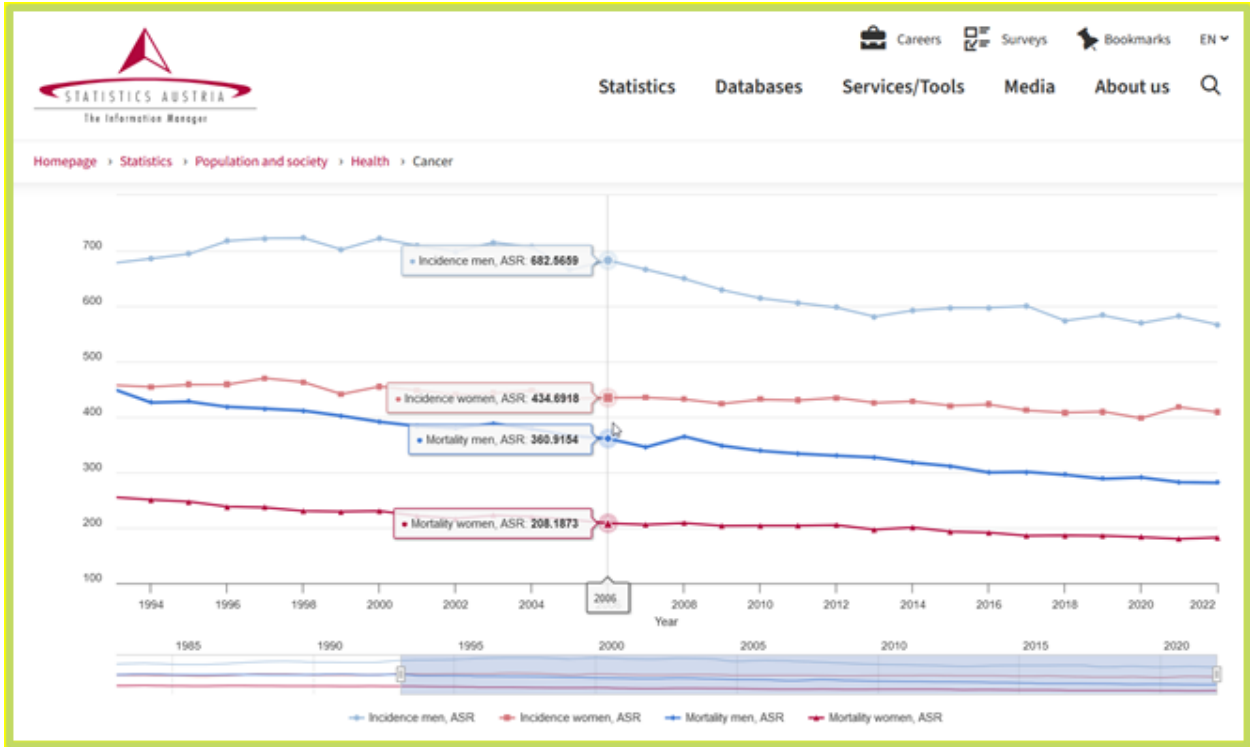
Example 3: the **Spanish Ministry of Health** publishes a statistical portal allowing interactive consultation of the National Health System.
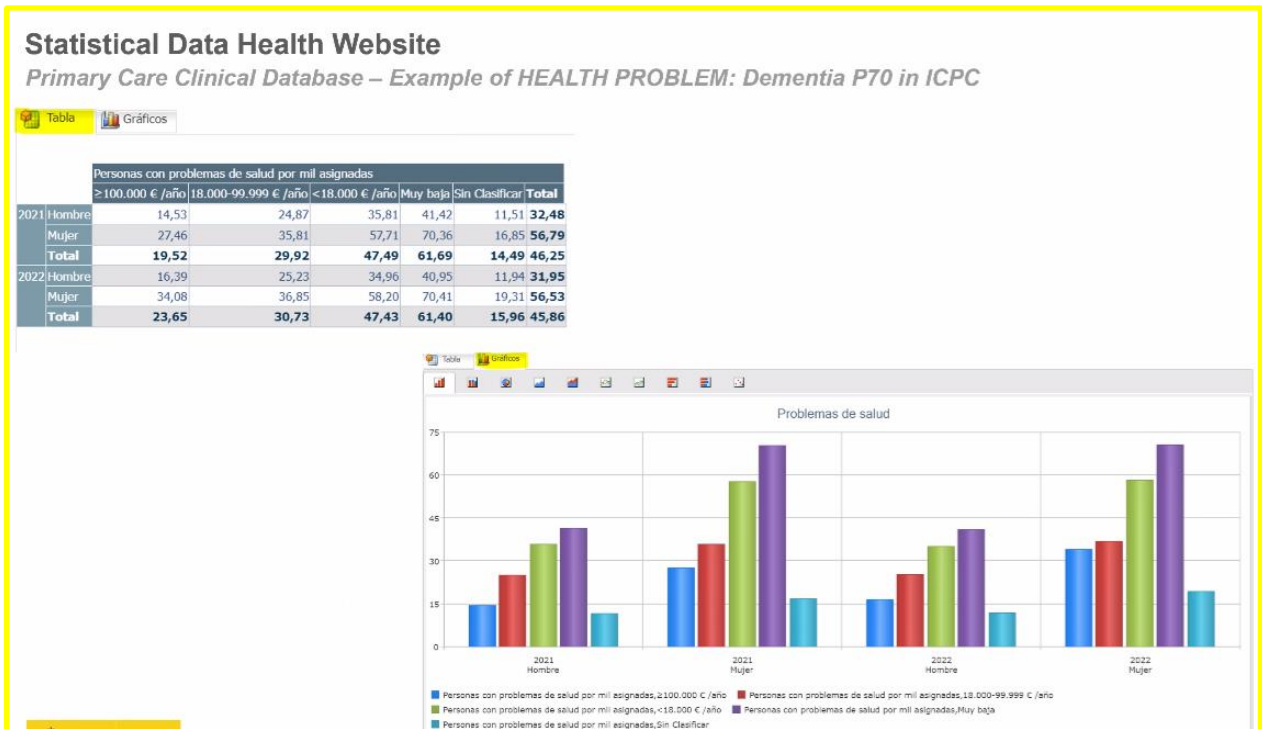
Lastly, the EHDS Regulation mandates that Health Data Access Bodies (HDABs) provide responses to health data requests in anonymized statistical formats where this is sufficient. Depending on the capabilities of the analytical systems implemented by data publishers, the HealthDCAT-AP analytics Distribution can help address this requirement through a data

service.

5.2.8. Quality annotation

According to Article 2 (2 aa) of the EHDS Regulation, a « *'data quality and utility label' means a graphic diagram, including a scale, describing the data quality and conditions of use of a dataset;*».

In this context, the EU Data Portal has provided a general approach for the HealthDCAT-AP draft specification. The EU Data Portal defines a quality framework for checking metadata against various indicators, termed "Metadata Quality Assurance". This quality assurance is based on five dimensions: Findability, Accessibility, Interoperability, Reusability, and Contextuality. These dimensions are practical implementations of the FAIR data principles, and derived from the computation of selected DCAT-AP properties.

For instance, the Reusability dimension is evaluated by analysing information provided for "Licence information," "Licence vocabulary," "Access restrictions," "Access restrictions vocabulary," "Contact point," and "Publisher". An average score, termed "Rating evolution," is generated, and a visualisation output is available for each dimension:

Figure 10: Example of the **Reusability** graphic diagram of the data.europa.eu portal

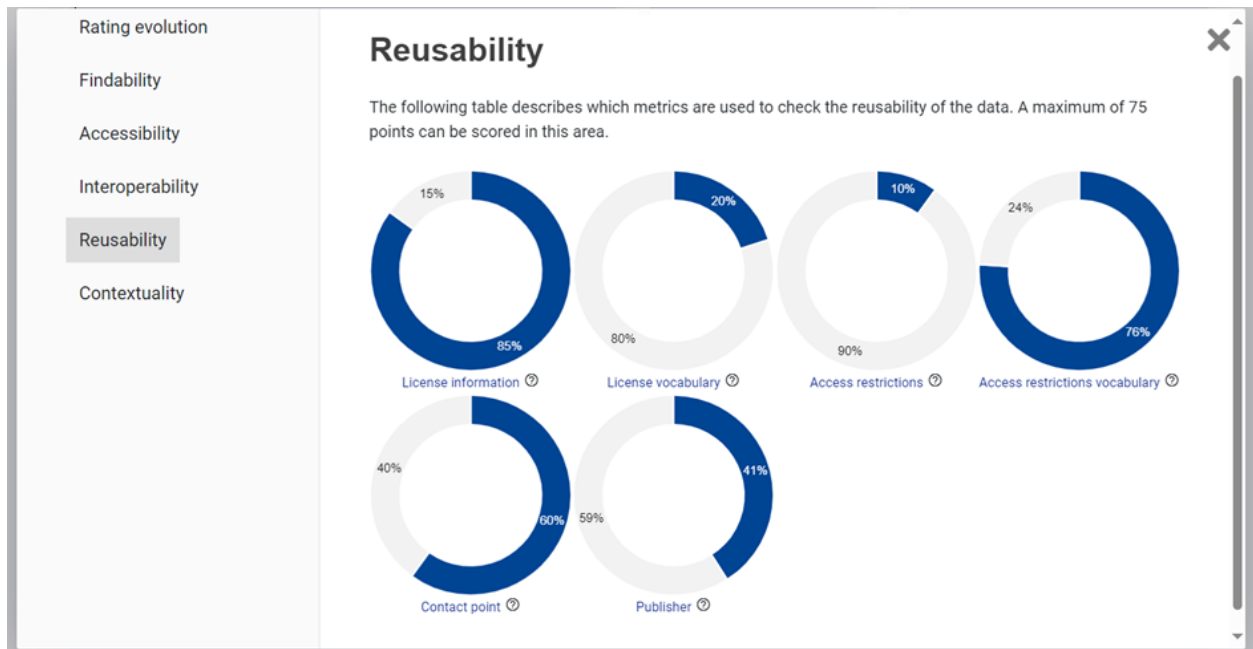The metadata quality evaluation in the EU Data Portal aids data providers and portals in improving their metadata. It also serves as a crucial filter for data users, enhancing their search experience.

Article 59 of the EHDS Regulation mandates that Health Data Access Bodies (HDABs) produce a biennial activity report. This report must include the "*(i) number of data quality labels issued by data holders*" for their data portal, "*disaggregated by quality category*".

The expected quality categories for the reporting are listed in the Article 78 "*Data quality and utility label*". For paragraphs (a), (c), (d), (e), and (f), DCAT-AP and the new properties of HealthDCAT-AP provide the necessary metrics for HDABs to generate data quality labels from metadata (Metadata Quality Dimensions according to the EU Data Portal). This alignment is because part of the requirements for the HealthDCAT design are directly derived from the FAIR data principles and the EHDS Regulation.

EHDS Regulation Article 78) Data quality and utility label

The data quality and utility label shall cover the following elements, where applicable:
(a) for data documentation: metadata, support documentation, the data dictionary, the format and standards used, the source of the data and, where applicable, the data model;
(b) for assessment of technical quality: completeness, uniqueness, accuracy, validity, timeliness and consistency of the data;
(c) for data quality management processes: the level of maturity of the data quality management processes, including review and audit processes, and bias examination;
(d) for assessment of coverage: the period, population coverage and, where applicable, representativity of the population sampled, and the average timeframe in which a natural person appears in a dataset;
(e) for information on access and provision: the time between the collection of the electronic health data and their addition to the dataset and the time needed to provide electronic health data following the issuing of a data permit or a health data request approval;
(f) for information on data modifications: merging and adding data to an existing dataset, including links with other datasets.

EHDS Regulation Article 60) Duties of health data holders

4. Where a data quality and utility label accompanies the dataset pursuant to Article 78, the health data holder shall provide sufficient documentation to the health data access body for that body to verify the accuracy of the label.

The missing category, where no metrics are available from metadata, is paragraph (b) "*for assessment of technical quality: completeness, uniqueness, accuracy, validity, timeliness and consistency of the data;*". Moreover, user feedback, which provides valuable insights for assessing the quality and utility ranking of the dataset, is not covered by Article 78.

To include information on any applicable quality and utility labels for a dataset, HealthDCAT-AP utilises the dqv:hasQualityAnnotation property. This property, sourced from the Data Quality Vocabulary (DQV), is an RDF vocabulary developed by the W3C's Data on the Web Working Group.

Property	Quality annotation

URI	dqv:hasQualityAnnotation
Range	oa:Annotation
Definition	A statement related to quality of the Dataset, including rating, quality certificate, feedback that can be associated to datasets or distributions.
Comment	The information may include quality aspects such as accuracy, reliability, comparability, coherence, relevance, timeliness etc.
Usage note	The annotation requires the provision of information about the motivation of the annotation (oa:motivation), and an explicit link to the resource being annotated (oa:hasTarget) together with either a link to a resource that contains the annotation (oa:hasBody) or text filed (oa:bodyText).

The approach undertaken by HealthDCAT-AP is also aligned with the StatDCAT Application Profile and the MobilityDCAT-AP published in April 2024.

Usage note for quality description in MobilityDCAT-AP

> *This property MAY describe any quality aspects regarding the delivered content, in particular methods, metrics/indicators and results of a quality assessment in the responsibility of the Rights Holder (see property dct:rightsHolder). This information SHOULD assist data consumers in determining the value of data, before actually accessing and processing it. Thus, the information SHOULD be publicly available. Furthermore, it can be helpful for validation processes by 3rd parties, e.g., a National Body in context of EU Delegated Regulations.*

To summarise, HealthDCAT-AP enables health data holders to provide the necessary information to generate the quality labels defined by the EHDS Regulation. However, how to assess technical quality still needs to be determined. And methods for collecting user feedback also need to be streamlined. Health Data Access Bodies are responsible for producing and reporting these labels by category, as outlined in the EHDS Regulation. The process for generating overall quality and utility labels according to a common framework also remains to be defined. This represents a broader challenge for the EU Data Spaces as they work to develop a comprehensive quality framework. A standardised approach to describe the quality of datasets across EU data spaces should be adopted. HealthDCAT-AP recommends using the Data Quality Vocabulary annotation property (dqv:hasQualityAnnotation) to provide information on quality, aligning with the approach taken by StatDCAT-AP and MobilityDCAT-AP.

Here's an example of how to include a quality certificate in a HealthDCAT-AP metadata record using the property dqv:hasQualityAnnotation. The oa:hasBody property points to an HTTP URI that serves as a direct link to the quality certificate, enabling users to access the certificate's full details and verify the dataset's quality standards.

```
dqv:hasQualityAnnotation [
a dqv:QualityCertificate ;
oa:hasTarget <https://…mydataset> ;
oa:hasBody <https://acertificateserver.eu/mycertificate> ;
oa:motivatedBy dqv:qualityAssessment];
```

For example, as illustrated in the vocabulary DQV examples ([vocab-dqv-examples](#)):

```
<https://acertificateserver.eu/mycertificate>
dqv:hasQualityMeasurement[a dqv:QualityMeasurement ;
dqv:isMeasurementOf :
[a dqv:Metric ; skos:definition "Ratio between the number of objects represented in
the csv and the number of  objects expected to be represented according to the
declared dataset scope."@en ;
dqv:expectedDataType xsd:double ;
dqv:inDimension [ a dqv:Dimension ;
skos:prefLabel "Completeness"@en ;
skos:definition "Completeness refers to the degree to which all required information
is present in a particular dataset."@en ;
dqv:inCategory :intrinsicDimensions
];
];
dqv:value "0.5"^^xsd:double];
…
```

A quality certificate relying on the Data Quality Vocabulary (DQV) can encompass a range of detailed information, creating a comprehensive assessment of data quality. This may include:

- a quality label,
- a utility label,
- specific quality dimensions (such as accuracy and completeness),
- associated metrics,
- and user feedback.

Real-world example: The dataset "[European Quality of Life Time Series, 2007 and 2011](#)" has achieved [Platinum (Expert level) certification](#) from the [Open Data Institute](#) (ODI).

The Open Data Institute facilitates an [auto-certification](#) process by retrieving DCAT metadata records. Although the certificate is not available in RDF format (and thus lacks the RDF-based DQV vocabulary structure), it is still accessible in a machine-readable [JSON format](#) (see example below). The quality certificate generated by the ODI service can be incorporated into HealthDCAT-AP using the dqv:hasQualityAnnotation property as a dqv:QualityCertificate.

The service implemented by ODI offers significant advantages, enabling any data holder to automatically generate a data quality certificate based on their dataset description. Data holders can enrich the certificate by adding additional information, such as technical quality metrics and user feedback reported, for instance, by the Health Data Access Body (HDAB). The HDAB can then use the generated certificate to create a graphical representation, as outlined in Article 2 (ae) of the EHDS Regulation, by accessing its machine-readable format. This flexibility allows the HDAB to customise the appearance of the graphic in its data catalogue. Furthermore, HealthDCAT-AP's approach supports data holders in subscribing to various quality and utility frameworks, enhancing the adaptability and interoperability of their datasets. These frameworks could be designed by the HDAB, a Member State, the EHDS, the common EU Data Spaces, or any relevant organisation (e.g., EOSC).

JSON   Raw Data   Headers
Save Copy Collapse All Expand All ▽ Filter JSON

version: 0.1
license: "http://opendatacommons.org/licenses/odbl/"
▼ certificate:
  ▼ title: "Open Data Certificate for European Quality of Life Time Series, 2007-2011: Open Access"
  ▼ uri: "https://certificates.theodi.org/en/datasets/1621/certificate"
    jurisdiction: "GB"
    status: "final"
    certification_type: "self certified"
  ▼ badges:
    ▼ application/javascript: "https://certificates.theodi.org/en/datasets/1621/certificate/badge.js"
    ▼ text/html: "https://certificates.theodi.org/en/datasets/1621/certificate/badge.html"
    ▼ image/png: "https://certificates.theodi.org/en/datasets/1621/certificate/badge.png"
  ▼ dataset:
    ▼ title: "European Quality of Life Time Series, 2007-2011: Open Access"
      publisher: "UK Data Service"
      uri: "https://certificates.theodi.org/en/datasets/1621"
      dataLicense: null
      contentLicense: null
      documentationUrl: "http://dx.doi.org/10.5255/UKDA-SN-7724-1"
      publisherUrl: "http://discover.ukdataservice.ac.uk/"
      publisherOrigin: false
    ▼ sourceDocumentationUrl: "http://discover.ukdataservice.ac.uk/catalogue/?sn=7724"
      sourceDocumentationMetadata: true
    ▼ copyrightURL: "http://discover.ukdataservice.ac.uk/catalogue/?sn=7724"
      contentRights: "samerights"
    ▼ copyrightStatementMetadata:
        0:                      "dataLicense"
        1:                      "contentLicense"
        2:                      "attribution"
        3:                      "attributionURL"
        4:                      "copyrightNotice"
        5:                      "copyrightYear"
        6:                      "copyrightHolder"
...
    ▼ distributionMetadata:
        0:                      "title"
        1:                      "description"
        2:                      "issued"
        3:                      "modified"
        4:                      "rights"
        5:                      "accessURL"
        6:                      "downloadURL"
        7:                      "mediaType"
      codelistDocumentationUrl: "http://esds.ac.uk/DDI25/7724.xml"
    ▼ engagementTeamUrl: "http://www.eurofound.europa.eu/surveys/about-eurofound-surveys/international-collaboration"
    ▼ libraries: "http://discover.ukdataservice.ac.uk/?sf=User%20guides&df=GuideTypeTextFacet_Methods%20and%20statistics"
      schemaDocumentationUrl: "http://esds.ac.uk/DDI25/7724.xml"
    level: "platinum"
    created_at: "2015-08-03T10:48:59Z"
                                                        Mozilla Firefox

Example of an ODI dataset quality certificate at the 'Platinum' level.

5.2.9. Purpose for collecting data

HealthDCAT-AP enhances the core DCAT-AP vocabulary by incorporating new properties from the Data Privacy Vocabulary, specifically designed to provide detailed information about the purpose for which data is collected. This is a critical addition, as understanding the purpose behind data collection is relevant to any dataset, whether personal or non-personal. While this type of metadata is beneficial for all datasets, it becomes especially relevant in the case of personal data, as it aligns with the requirements of the General Data

Protection Regulation (GDPR). Under GDPR, organisations must clearly articulate the legal basis and purpose for processing personal data, making this information essential for datasets involving personal information. The inclusion of such properties within HealthDCAT-AP ensures a more complete understanding of the context and rationale for data collection, improving dataset transparency and aiding users in making informed decisions when accessing and utilising health data.

Property	Purpose
URI	dpv:hasPurpose
Range	Literal
Definition	A free text statement of the purpose of the processing of data or personal data.
Usage note	The purpose or goal here is intended to sufficiently describe the intention or objective of why the data or technology is being used, and should be broader than mere technical descriptions of achieving a capability. For example, "Analyse Data" is an abstract purpose with no indication of what the analyses is for as compared to a purpose such as "Marketing" or "Service Provision" which provide clarity and comprehension of the 'purpose' and can be enhanced with additional descriptions.

Property	Legal basis
URI	dpv:hasLegalBasis
Range	rdfs:Resource, expressed as a URI.
Definition	The legal basis used to justify processing of personal data.
Usage note	Legal basis (plural: legal bases) are defined by legislations and regulations, whose applicability is usually restricted to specific jurisdictions which can be represented using dpv:hasJurisdiction or dpv:hasLaw. Legal basis can be used without such declarations, e.g. 'Consent', however their interpretation will require association with a law, e.g. 'EU GDPR'.

HealthDCAT-AP also incorporates the [Personal Data extension](#) of the Data Privacy Vocabulary (DPV) specification providing additional concepts to represent different types and categories of personal data. The Personal Data extension (DPV-PD) offers as such a harmonised, standardised RDF framework of describing sensitive information in the context of HealthDCAT-AP. By extending DCAT-AP with DPV-PD, it enables the sensitive nature of health datasets containing personal information to be uniformly described and understood

across various dataset catalogues, promoting interoperability and consistency.

Property	Personal Data
URI	dpv:hasPersonalData
Range	rdfs:Resource, expressed as a URI.
Definition	Key elements that represent an individual in the dataset.
Usage note	This definition of personal data encompasses the concepts used in GDPR Art.4-1 for 'personal data' and ISO/IEC 2700 for 'personally identifiable information (PII)'.

Real-world example: The dataset "Linking of registers for COVID-19 vaccine surveillance" links selected variables from existing registries for COVID-19 vaccine surveillance, in order to ensure the monitoring of COVID- 19 vaccines in the phase following their marketing authorization. Given that this dataset contains personal-level information on Belgian citizens, its creation required approval from the Information Security Committee. The dataset's legal basis, purpose, and personal data categories are transparently communicated to data users through the use of the [Data Privacy Vocabulary](#) properties: dpv:hasLegalBasis, dpv:hasPurpose, and dpv:hasPersonalData.

dpv:hasPurpose dpv:Purpose

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

dpv:hasLegalBasis dpv:LegalBasis

```
dpv:hasLegalBasis [
a dpv:LegalBasis ;
dct:description "CSI Deliberation no. 21/028 of february 18, 2021, last
amended on june 18, 2021, relating to the communication of data to
pseudonymized personal character relating to the health of vaccinnet+,
healthdata covid-19 database i and ii, healthdata covid-19 clinical
database, cobrha, statbel and the agency intermutualist in sciensano, as
part of the link-vacc project and the subsequent processing of personal data
pseudonymised by the federal drug agency in view monitoring the safety of
covid-19 vaccines"@en;
```

```
dct:source
<https://www.ehealth.fgov.be/ehealthplatform/file/view/AXkNfdPml9vUUfvGGfJr?
filename=21-028-f212-AFMPS-vaccinnet-
modifi%C3%A9e%20le%2018%20juin%202021.pdf>,
<https://www.ehealth.fgov.be/ehealthplatform/file/view/AX_-
9sZSuwVJMAnC0ENo?filename=21-028-f166-LINK-VACC-
modifi%C3%A9e%20le%205%20avril%202022.pdf> ;
];
```

dpv:hasPersonalData dpv:PersonalData

```
dpv:hasPersonalData dpv-pd:Gender, dpv-pd:Age, dpv-pd:Location, dpv-
pd:Nationality, dpv-pd:Education, dpv-pd:HealthRecord;
```

5.3. Minimum HealthDCAT-AP elements

EHDS Regulation Article 77) Dataset description and dataset catalogue

*4. … The Commission shall, by means of implementing acts, set out the minimum ▌ elements
health data holders are to provide for datasets and the characteristics of those
elements.*

We propose categorising health datasets (i.e.: datasets in scope of Art.51) into three distinct
categories:

1. Non-personal electronic health data available as open data [Open data]
2. Non-personal electronic health data available as non-open data [Protected data or
   data with restricted access]
3. Personal electronic health data [Sensitive data or non-public data]

Based on this categorisation, we can further discuss the minimum metadata elements
required for a HealthDCAT-AP record by taking into account existing applicable EU
legislations.

5.3.1.1 Non-personal electronic health data available as [open data]

The original scope of the DCAT-AP specification was to enable the description and exchange
of publicly accessible Open Data across Open Data Portals in Europe. The dct:accessRights
property, combined with the controlled vocabulary "Access Rights" maintained by the
Publications Office, allows data users to identify datasets as Open Data.

Publications Office: Access right authority list:

The access-right authority table is a controlled vocabulary listing the access rights or
restrictions to resources. It is designed for but not limited to DCAT descriptions of datasets.
This authority table is maintained by the Publications Office of the European Union and
disseminated on the EU Vocabularies website.

*PUBLIC: Publicly accessible by everyone. Usage note: Permissible obstacles include registration and request for
API keys, as long as anyone can request such registration and/or API keys.*

DCAT-AP 3.0 minimum elements

The minimum elements required in a DCAT-AP 3.0 metadata record are the 'title' and 'description.' If a dataset distribution is included, the 'access URL' also becomes mandatory.

DCAT-AP 3.0 Dataset

title	Literal	1..*	A name given to the Dataset.	This property can be repeated for parallel language versions of the name.

description	Literal	1..*	A free-text account of the Dataset.	This property can be repeated for parallel language versions of the description.

DCAT-AP 3.0 Distribution

access URL	Resource	1..*	A URL that gives access to a Distribution of the Dataset.	The resource at the access URL may contain information about how to get the Dataset.

For 'non-personal electronic health data,' the EHDS Regulation mandates that the data must be open and accessible, which necessitates the inclusion of at least one distribution. However, this requirement is not addressed by the minimum metadata elements defined in DCAT-AP 3.0.

EHDS Regulation Article 60) Duties of the health data holders (open data)

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

This requirement is addressed by the minimum metadata elements defined in DCAT-AP High-Value Datasets:

DCAT-AP HVD minimum elements

The High-Value Datasets Implementing Regulation has introduced minimum elements to enhance metadata quality, thereby increasing the level of reuse for a selected group of core datasets within the public sector—namely Geospatial, Earth Observation and Environment, Meteorological, Statistics, Companies and Company Ownership, and Mobility. For instance, the 'Applicable Legislation' property, introduced in DCAT-AP version 3.0, is defined as a mandatory metadata element. Furthermore, any HVD metadata record must include at least one dataset distribution, a requirement also expressed in the European Health Data Space (EHDS) Regulation, as specified in Whereas 6 and expected in HealthDCAT-AP.

DCAT-AP HVD Dataset

applicable legislation	Legal Resource	1..*	The legislation that mandates the creation or management of the Dataset.	For HVD the value must include the ELI http://data.europa.eu/eli/reg_impl/2023/138/oj. As multiple legislations may apply to the resource the maximum cardinality is not limited.

dataset distribution	Distribution	1..*	An available Distribution for the Dataset.	The HVD IR is a quality improvement of existing datasets. The intention is that HVD datasets are publicly and open accessible. Therefore a Distribution is expected to be present. (Article 3.1)

The High-Value Datasets Implementing Regulation has also introduced a mandatory controlled vocabulary to classify datasets according to defined core datasets by the IR. The "Health" doesn't belong to the core datasets in scope of the HVD IR.

HVD Category	Concept	1..*	The HVD category to which this Dataset belongs.

However, a dataset under the scope of the High-Value Datasets Implementing Regulation (HVD IR) can also fall within the scope of the EHDS Regulation. To effectively populate health dataset catalogues with High-Value Datasets, we propose aligning the HealthDCAT-AP minimum elements with those of DCAT-AP HVD. Metadata conforming to DCAT-AP 3.0 can only be considered valid if at least one distribution is available. The 'applicable legislation' property should enable filtering for High-Value Datasets that fall under the EHDS.

Expecting HVD data holders to consistently tag the 'applicable legislation' with the EHDS Regulation may be unrealistic. Therefore, Health Data Access Bodies need to consider an automatic metadata enrichment process based on sufficient indicators that a dataset is within the scope of the EHDS Regulation. Implementing a centralised open AI service to automate this filtering task could be an efficient and effective solution.

Reminder: Nearly 27,000 records are labelled as health-related datasets using the DCAT 'theme' property, which requires the use of the Dataset Theme Vocabulary maintained by the EC Publications Office.

DCAT-AP HVD Distribution

applicable legislation	Legal Resource	1..*	The legislation that mandates the creation or management of the Distribution	For HVD the value must include the ELI http://data.europa.eu/eli/reg_impl/2023/138/oj. As multiple legislations may apply to the resource the maximum cardinality is not limited.

5.3.1.2 Non health personal electronic data available as non-public data [Protected data]

Publications Office: Access right authority list:

RESTRICTED: Only available under certain conditions. Usage note: This category may include resources that require payment, resources shared under non-disclosure agreements, resources for which the publisher or owner has not yet decided if they can be publicly released.

The Data Governance Act introduced the concepts of NSIP (National Single Information Point) metadata and non-open NSIP data that can be searched in the European Data portal as European Register for Protected Data. These data according to the DGA can be health datasets and would fail in the scope of the EHDS Regulation. The exercise of defining HealthDCAT-AP minimum metadata elements implies to consider the requirements introduced by the DGA for non-open data. The minimum NSIP metadata elements include information on publishers and conditions for the re-use of data (both relating to the dataset metadata level; dct:publisher and dct:rights) as well as information on the format and size of individual distributions (dct:format and dcat:byteSize).

Harvesting guidelines for the European Register for Protected Data (ERPD)

The Data Governance Act does not provide concrete technical instructions for the implementation of the NSIPs and the ERPD, neither on the technology or data level. The harvesting guidelines presented here are therefore based on an interpretation of the relevant articles of the Data Governance Act. This concerns in particular Articles 5-8 of the Data Governance Act.
…
Metadata that can be identified as being mandatory from the Data Governance Act are therefore mapped into DCAT-AP properties and must be structured correspondingly by NSIPs to enable harvesting. This relates to the requested information on the titles of datasets, their descriptions, their publisher, conditions for reuse and access procedure, format, and size.

The following metadata is mandatory for **NSIP datasets**:

Property	URI	Range	Usage note	Cardinal

				ity
Title (M)	dct:title *	rdfs:Literal	This property contains a name given to the Dataset. This property can be repeated for parallel language versions of the name.	1..n
Descripti on (M)	dct:descr iption *	rdfs:Literal	This property contains a free-text account of the Dataset. This property can be repeated for parallel language versions of the description.	1..n
Publisher (M)	dct:publis her	foaf:Agent	This property refers to an entity (organisation) responsible for making the Dataset available.	1..1
Access rights (M)	dct:acces sRights	dct:Rights Statement	This property refers to information that indicates whether the dataset is open data, has access restrictions, or is not public. From the controlled vocabulary of the Publications Office of the EU 6, the following codes should be used for NSIP data: "non-public" or "restricted". "open" is prohibited for NSIP data.	1..1
Distributi on (M)	dct:distri bution	dcat:Distrib ution	Distribution(s) available for a dataset.	1..n

The following metadata is mandatory for **NSIP Distributions**:

Property	URI	Range	Usage note	Cardinality
Format (M)	dct:format	dct:MediaTypeOrExtent	This property refers to the file format of the Distribution. You can only specify one format per Distribution. If an NSIP offers the same data in different formats, each format must be specified as a separate distribution.	1..1
Size (M)	dcat:byteSize	rdfs:Literal	The size in bytes can be approximated (as a decimal) if the precise size is not known. If data is offered via an API or other endpoint, size should refer to the overall size of the underlying dataset.	1..1
Access procedure (M)	dcat:accessURL *	rdfs:Resource	A URL of a Website that enables either access to the described data or that contains information on how to request the data.	1..n
Conditions for re-use (Rights) (M)	dct:rights	dct:RightsStatement	This property refers to a statement that specifies rights associated with the Distribution.	1..1

Considering that a non-open NSIP dataset may fall within the scope of the EHDS Regulation, we propose aligning the HealthDCAT-AP minimum elements for non-personal electronic health data, available as non-open data, with those of NSIP metadata. The mandatory NSIP metadata for distributions is well-suited for accessing datasets without requiring a data request or permit application. However, it may be insufficient for personal electronic health

data, where a formal application process must be managed by a Health Data Access Body.

5.3.1.3 Health personal electronic data [Sensitive data]

When a data request or permit application is required, the following minimum metadata elements have been discussed and proposed, along with the rationale behind the decisions on their cardinality

Publications Office: Access right authority list:

NON PUBLIC: Not publicly accessible for privacy, security or other reasons. Usage note: This category may include resources that contain sensitive or personal information.

Mandatory properties of HealthDCAT-AP for **personal electronic health data**

Property	Reasoning about cardinality's decision
Access rights	If information on data access is not provided, we cannot determine if the dataset needs to be linked to an application form. Therefore, health data holders MUST classify their datasets as either PUBLIC, RESTRICTED, or NON-PUBLIC. For HealthDCAT-AP, we will use only PUBLIC and NON-PUBLIC categories:
	PUBLIC: Publicly accessible by everyone. Note: "Permissible obstacles include registration and request for API keys, as long as anyone can request such registration and/or API keys."
	NON-PUBLIC: Not publicly accessible for privacy, security, or other reasons. Note: This category may include resources with sensitive or personal information.
	If a dataset is PUBLIC, the metadata MUST include a distribution with an access URL or download link, as per the EHDS Regulation. If it is NON-PUBLIC, the metadata MUST include a distribution with a HDAB landing page or equivalent authority, according to the EHDS Regulation.
Applicable legislations	This property is essential for efficiently managing metadata records in a catalogue with multiple DCAT application profiles. Without a proper filter, customising an API to filter datasets is challenging. This information helps knowledgeable data users understand the reuse conditions of datasets.
	A HDAB could automatically add the EHDS Regulation to this property for all metadata records in its dataset catalogue. A dataset can be in scope of multiple Directives and Regulations. Ex: INSPIRE health Theme Another example: a statistical population dataset (Census) is not dct:theme= HEAL but applicable legislation is HVD and EHDS.
	(Comment: The conformsTo property in the catalogRecord class offers another filtering option if the DCAT Application Profile is provided, though this is more technical.) Upload mandatory condition to the EU health Data catalogue: applicableLegislation= EHDS (ELI link)
Contact point	A contact point should always be available for data users seeking more information about a dataset. This way, data users don't need to determine whether to contact the creators, publishers, or HDABs; the contact point serves as the dataset's "helpdesk."
Dataset distribution	In HealthDCAT-AP, there is always at least one distribution available. For sensitive data, this is typically an HDAB landing page. For data with public access rights, it may be a landing page, a direct download link, or a data service. Access to the data may require users to log in.

Description	Mandatory in DCAT-AP
Geographical coverage	Do all datasets listed in Art. 33 have a geographical signature? It appears they do, but this needs confirmation. This property is an important filter when considering populations. Providing geographical coverage can be technically challenging due to various possibilities. We recommend using controlled vocabularies for this property (see the usage note of DCAT-AP 3). Associating these vocabularies with spatial geometries would be beneficial.
Health category	Within a domain-specific catalogue, efficiently categorising datasets is essential for helping data users navigate and find relevant datasets more easily. Can TEHDAS2 assist in defining short labels for all categories in Art. 33, which is challenging for data holders? Additionally, could it be possible to develop a tool that automatically selects the correct category for data holders?
Health data access body	For all datasets where dct:type=personal data and/or dct:accessRights=non public, a HDAB or equivalent authority acts as a data access gateway, making this property mandatory. (This use case should be included in the Implementing Act for high impact datasets.) If dct:accessRights=public, there's no need for the HDAB contact point, but a HDAB may choose to add it to metadata records in its catalogue. HDABs will need to update many metadata records from various sources. To include its catalogue in the central health EU catalogue, this property is mandatory. This means HealthDCAT-AP is an application profile supporting the healthDATA@EU infrastructure.
Health theme	Similar to keywords, but designed for machines, the health theme property relies on Wikidata—a large-scale, human-readable, machine-readable, multilingual, multidisciplinary, centralised, editable, structured, and linked knowledge base. This significantly enhances the quality and usability of dataset descriptions for machines. Evaluating the use of Wikidata and confirming its suitability could be an activity for TEHDAS2.
Identifier	DCAT-AP HVD rational on identifiers To ensure the operational management of metadata records, Identifiers are defined as mandatory PURI for HealthDCAT-AP for high impact datasets
Keyword	It is important to include relevant keywords that characterise the dataset. While always applicable, this should be restricted to the main keywords. Data holders should avoid copy-pasting extensive lists and instead limit the keywords to 10-15 per language. For instance, if it consists of code values, it is preferable to use the property "code values" or to provide access to the codebook in adms:sample (consider to use CSV on the Web to describe the variables)
Provenance	It (dataset lineage) provides transparency about the origins and history of a dataset, ensuring data reliability and trustworthiness. It helps users understand how the dataset was created, modified, and by whom, which is crucial for assessing its quality and relevance. TEHDAS2 could work on improving the usage note
Publisher	The publisher acts as the data holder and is accountable for making the dataset accessible.
Publisher note	Understanding the nature of the publisher provides valuable context for the dataset and can be indicative of its overall quality. For example, a dataset published by a research institute may have higher credibility and reliability compared to others.

Publisher type	The same rationales apply as for Publisher Type, but they should be provided as full text.
Purpose	It explicitly explains why the data was collected and outlines its primary use. Do we need a definition of primary users and secondary users? "Peer reviewed journals would like to have this information listed, whether you reuse the data or are the original creator of the data set or not."
Sample	Mandatory (experimental) for personal electronic health data, this DCAT-AP 3.0 property is particularly valuable for sensitive data. It allows the exposure of a dataset representation artefact, such as mock-up data, synthetic data, anonymous data, or a codebook (data dictionary) that reveals the dataset's structure.
Theme	All records in the HDABs' catalogues must tag datasets as related to health using the OP CR entry "HEAL."
Title	Mandatory in DCAT-AP 3.0
Data type	Art. 33 encompasses both personal and non-personal electronic health data. It is crucial to distinguish between these data types, as defined by the Controlled Vocabulary (CV) that can be applied. The CV includes entries such as geospatial data, statistical data, ontologies, etc. We propose adding personal data to this CV. For more information, you can refer to the EU Vocabularies concept scheme: https://op.europa.eu/en/web/eu-vocabularies/concept-scheme/-/resource?uri=http://publications.europa.eu/resource/authority/dataset-type Additionally, does Art. 33 include ontologies, such as the Rare Diseases Ontology? This exercise is part of TEHDAS2's comparison of Art. 33 and the Dataset-type authority table.

Recommended properties of HealthDCAT-AP for **personal electronic health data**

Analytics	This property has been introduced in HealthDCAT-AP to expose metrics and insights about the dataset. It is of the type Distribution, similar to dct:distribution and adms:sample, as it also represents data. However, it should not be considered a standard distribution of the dataset itself. Instead, it may take the form of an analytics dashboard (advanced use case), a technical report (statistical data), or an API for querying the dataset. It does not provide any direct download or subset possibilities of the dataset. The ECDC Atlas, used in testing HealthDCAT-AP, perfectly illustrates the use of this property.
Code values	This property can enhance discoverability. For instance, a data user might search for a specific disease using a coding system like ICD-10. Moreover, this property is designed to be machine-actionable, facilitating automated processes and searches.
Coding systems	This property indicates the readiness of the dataset for reuse and can enhance its discoverability. For instance, a data user might search for datasets that utilise ICD-10 for coding diseases. Additionally, this property is designed to be machine-actionable, facilitating automated processes and searches.
Conforms to	This property provides information on the readiness of the dataset for reuse. It can improve the discoverability of a dataset. A data user might search for a dataset compliant to the OMOP data model.
Documentation	Publishing documentation about datasets is a valuable and common practice that enhances understanding of the dataset. However, maintaining publicly available documentation can be challenging for data holders. Documentation serves as a quality element in the Regulation.

frequency	
is referenced by	FAIR I3: (Meta)data include qualified references to other (meta)data.
Landing page	A landing page is defined as "a web page that provides access to the dataset, its distributions, and/or additional information. It is intended to point to a landing page at the original data provider" (DCAT-AP 3). Support for implementation: We need to explicitly define two concurrent properties: 1. dcat:landingPage in the dataset class (Optional): This property refers to a web page that provides comprehensive access to the dataset and related information. 2. dcat:accessURL in the Distribution class (Mandatory): This is a URL that gives access to a specific distribution of the dataset. The resource at the access URL may include information on how to obtain the dataset. Considering that in HealthDCAT-AP a distribution always exists and that the access URL is a mandatory property within a distribution, we need to clarify whether the landing page should also be a mandatory property and what specific information (URI) it should contain. In any case, it is a recommended property.
Language	A data user expects to have this information when it pertains to health data if applicable.
Legal basis	The collection of personal data must always have a legal basis under the GDPR. If such a legal basis exists, it should be provided to help data users better understand the premise of the dataset's (secondary) use.
Maximum typical age	Applicable only for population datasets. Useful filter to explore a catalogue. This concept of typical age exists in several health dataset catalogues.
Minimum typical age	Applicable only for population datasets. Useful filter to explore a catalogue. This concept of typical age exists in several health dataset catalogues.
Number of Records	Referring to Art. 8 of the Data Governance Act: "... with relevant information describing the available data, including at least the data format and size and the conditions for their re-use," one can observe that specifying the size is mandatory. Beyond offering quantitative information, it allows the data user to estimate the 'value' of the dataset.
Number of unique individual	Beyond providing quantitative information, it enables the data user to estimate the 'value' of the dataset.
Personal data	Understanding the nature of the publisher provides valuable context for the dataset and can be indicative of its overall quality. For example, a dataset published by a research institute may have higher credibility and reliability compared to others.
Population coverage	Definition: 'A description of the population within the dataset.' This is important for understanding the dataset, but it is only applicable if the dataset contains a population.
quality annotation	EHDS reg. Whereas (59) The data quality and utility label should not prevent datasets from being made available through the EHDS

Related resource (dct:relation)	FAIR I3: (Meta)data include qualified references to other (meta)data. All relationships to other datasets and resources help to understand how a dataset has been used and, ultimately, how it can be used. This practice extends the dataset's DCAT knowledge graph.
Source (A related dataset dct:source)	FAIR I3: (Meta)data include qualified references to other (meta)data. Relationships to other datasets and resources help understand how a dataset has been used and how it can be used in the future. This process extends the dataset's DCAT knowledge graph.
Temporal coverage	The Temporal Coverage property, defined by two timestamps (start date and end date), specifies the period that the dataset covers. If data collection is ongoing, no end date is provided. Is this approach sufficient? (It could be discussed in TEHDAS2. Keep it simple.)
Temporal resolution	It is an asset to have this information as it allows one to determine the reuse applicability.

Optional properties of HealthDCAT-AP for **personal electronic health data**

Retention period	If the dataset must be deleted after a specific date, this property is mandatory, as the dataset cannot be included in the EU central dataset catalogue "basket." However, the metadata record must remain available because some publications may reference it.
Spatial resolution (meters)	Rarely applicable, except for geospatial data.
Version	A good practice in data management is to maintain a versioning strategy for datasets. While not essential for the discoverability and understanding of the dataset, it is up to the data holder to decide whether to provide this information.
...	

5.3.1.4 Conclusion

In summary, DCAT-AP metadata can populate a health dataset catalogue if it meets specific criteria, depending on the type of health data being described:

- [**Open data**] HealthDCAT-AP for non-personal electronic health data is categorised by applying the filter condition dcat:theme=HEAL. For open data under the High-Value Datasets Implementing Regulation (HVD IR), essential metadata fields such as dct:title, dct:description, dcatap:applicableLegislation=ELI HVD, and distribution elements like Access URL must be provided.

- [**Protected data**] For non-open data under the Data Governance Act (DGA), additional fields, including dct:accessRights=restricted, dct:publisher, and distribution details such as format, size, and re-use conditions, are required.

- [**Sensitive Data**] HealthDCAT-AP for personal electronic health data requires a more specific filter condition: dcat:theme=HEAL + dct:type=personal_data. This activates additional cardinalities in HealthDCAT-AP to ensure metadata

completeness, including fields such as Access Rights, Applicable Legislation, Contact Point, Dataset Distribution, Description, Geographical Coverage, Health Category, Health Data Access Body, Identifier, Publisher, Purpose, and more.

HealthDCAT-AP defines three conditional sets of minimum metadata elements depending on the sensitivity of the data, reflecting the variety of datasets in scope under the EHDS Regulation (Article 33). This flexibility ensures that both personal and non-personal health data are appropriately managed and described, supporting the diverse needs of health data governance.

HealthDCAT-AP cardinalities depending of the access right on the dataset:

PUBLIC	RESTRICTED	NON_PUBLIC
Mandatory properties		
dct:description: rdfs:Literal [1..n] dct:title: rdfs:Literal [1..n] dct:identifier: rdfs:Literal: xsd:anyURI [1..n] dcatap:applicableLegislation rdfs:Resource [1..n] dcat:theme (dct:subject): skos:Concept [1..n] dct:accessRights: dct:RightsStatement [1..1] dcat:distribution: dcat:Distribution [1..n] healthdcatap:hdab foaf:Agent [1..1] heathdcatap:healthCategory: (dct:subject) skos:Concept [1..n]	dct:description: rdfs:Literal [1..n] dct:title: rdfs:Literal [1..n] dct:identifier: rdfs:Literal: xsd:anyURI [1..n] dcatap:applicableLegislation rdfs:Resource [1..n] dcat:theme (dct:subject): skos:Concept [1..n] dct:accessRights: dct:RightsStatement [1..1] dct:publisher: foaf:Agent [0..1] dcat:distribution: dcat:Distribution [1..n] healthdcatap:hdab foaf:Agent [1..1] heathdcatap:healthCategory: (dct:subject) skos:Concept [1..n]	adms:sample: dcat:Distribution [1..n] dcat:contactPoint: vcard:Kind [1..n] dcat:distribution: dcat:Distribution [1..n] dcat:keyword: rdfs:Literal [1..n] dcat:theme (dct:subject): skos:Concept [1..n] dcatap:applicableLegislation rdfs:Resource [1..n] dct:accessRights: dct:RightsStatement [1..1] dct:description: rdfs:Literal [1..n] dct:identifier: rdfs:Literal: xsd:anyURI [1..n] dct:provenance: dct:ProvenanceStatement [1..n] dct:publisher: foaf:Agent [1..1] dct:spatial: dct:Location [1..n] dct:title: rdfs:Literal [1..n] dct:type: skos:Concept [1..1] dpv:hasPurpose dpv:Purpose [1..n] healthdcatap:hdab foaf:Agent [1..1] heathdcatap:healthCategory: (dct:subject) skos:Concept [1..n] healthdcatap:healthTheme: (dct:subject) skos:Concept [1..n]
Recommended properties		
dcat:contactPoint: vcard:Kind [0..n] dcat:keyword: rdfs:Literal [0..n]	dcat:contactPoint: vcard:Kind [0..n] dcat:keyword: rdfs:Literal [0..n]	dcat:landingPage: foaf:Document [0..n] dcat:temporalResolution rdfs:Literal: xsd:duration [0..1] dct:accrualPeriodicity: dct:Frequency [0..1] dct:conformsTo: dct:Standard [0..n] dct:isReferencedBy: rdfs:Resource [0..n] dct:language: dct:LinguisticSystem [0..n] dct:relation: rdfs:Resource [0..n] dct:source: dcat:Dataset [0..n] dct:temporal: dct:PeriodOfTime [0..n] dpv:hasLegalBasis dpv:LegalBasis [1..n] dpv:hasPersonalData dpv:PersonalData [0..n] dqv:hasQualityAnnotation dqv:QualityCertificate [1..n]

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

rdfs:Literal [0..n] healthdcatap:retentionPeriod dct:PeriodOfTime [0..1] owl:versionInfo: rdfs:Literal [0..1] prov:qualifiedAttribution prov:attribution [0..n] prov:wasGeneratedBy: prov:Activity [0..n]	rdfs:Literal [0..n] healthdcatap:retentionPeriod dct:PeriodOfTime [0..1] owl:versionInfo: rdfs:Literal [0..1] prov:qualifiedAttribution prov:attribution [0..n] prov:wasGeneratedBy: prov:Activity [0..n]	

The most recent version of HealthDCAT-AP cardinalities is available at https://HealthDCAT-AP.github.io

6. Uptake strategies

The operationalisation of the healthDCAT@EU infrastructure will require an implementation strategy that includes developing technical specifications, guidelines, and proofs of concept. This strategy must independently address the needs of various stakeholders: data holders, catalogue managers (metadata managers), and data users. HealthDCAT-AP, an RDF linked data technology, provides machine-readable and actionable information, enhancing the discovery and reuse of health datasets. As a technology of the Semantic Web, it supports the interoperability of a new generation of dataset catalogues that may be unknown to the health domain experts. The implementation strategy must therefore be tailored to the specific perspectives of these stakeholders, avoiding unnecessary technical complexity for non-technical experts. It should provide hands-on support and tools, use cases, and training accordingly.

- Data holders need to establish solutions for maintaining their data and metadata according to the FAIR principles
- Data holders need to set up a way of transforming their metadata to HealthDCAT-AP if the format of their metadata is not originally compliant to the HealthDCAT-AP specification.
- Data publishers (aka HDABs) need to understand HealthDCAT-AP in order to implement and manage health dataset catalogues
- Both data holders and publishers may need to understand the possibilities and limitations of RDF linked data technology.
- Data users need to understand how to use health dataset catalogues to find and access health datasets. It includes advanced search queries (e.g., SPARQL as a query language).
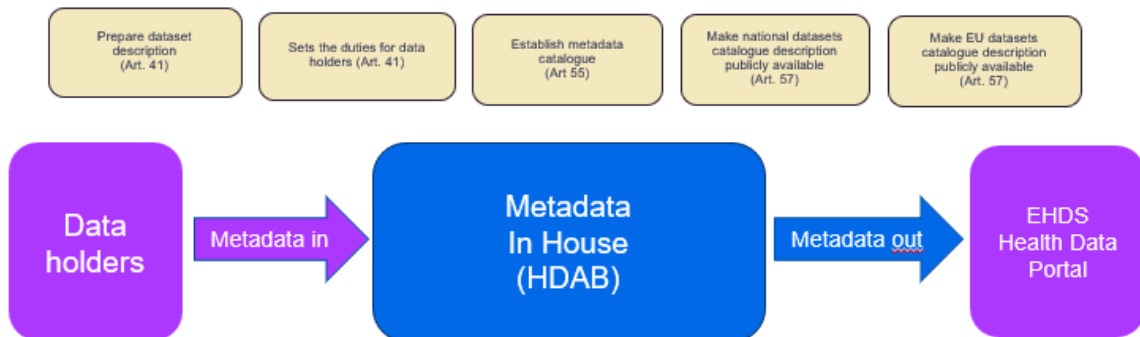
Figure 11: High level metadata value chain from data holders, via the HDAB and to the EU Dataset catalogue (aka, EHDS Health Data Portal) (Scope of the EHDS regulation)

The data and metadata value chain

To establish an efficient metadata value chain, it is essential to understand the flow of metadata from its origin, through the data holders, to the Health Data Access Bodies (HDABs), and ultimately to external dataset catalogues like the EU Dataset catalogue. Only with this clear understanding would it be possible to accurately define the scope and deliverables in line with the EHDS regulation, while also addressing the additional capabilities required for active metadata and data management.

Questions to be answered during the Second Joint Action Towards the European Health Data Space (TEHDAS2) which were not in scope of the Work Package 6 of the EHDS2 Pilot project:

- What are the key requirements for data holders in managing and maintaining metadata?
- How will data holders share their metadata with the National Health Data Access Body (HDAB)?
- How will data holders establish an efficient service for producing standardised data products (datasets) and data on demand, including sensitive (anonymized or pseudonymized), aggregated statistical data, sample datasets, and synthetic datasets?
- What are the key requirements for the Health Data Access Body (HDAB) in managing and maintaining metadata?
- How will HDABs share their metadata with other dataset catalogues? How will they transform metadata from other standards into HealthDCAT-AP?

In the figures below, we illustrate the metadata value chain in greater detail, highlighting the key capabilities and components necessary for active metadata management at both the data holder level and within the HDAB. While this could have been expanded to include solutions for broader data management, analysis systems, and the production of open datasets and statistics, these areas fall outside the scope of WP6. However, we strongly recommend focusing on the interface between metadata and data management as the EHDS continues to be developed.
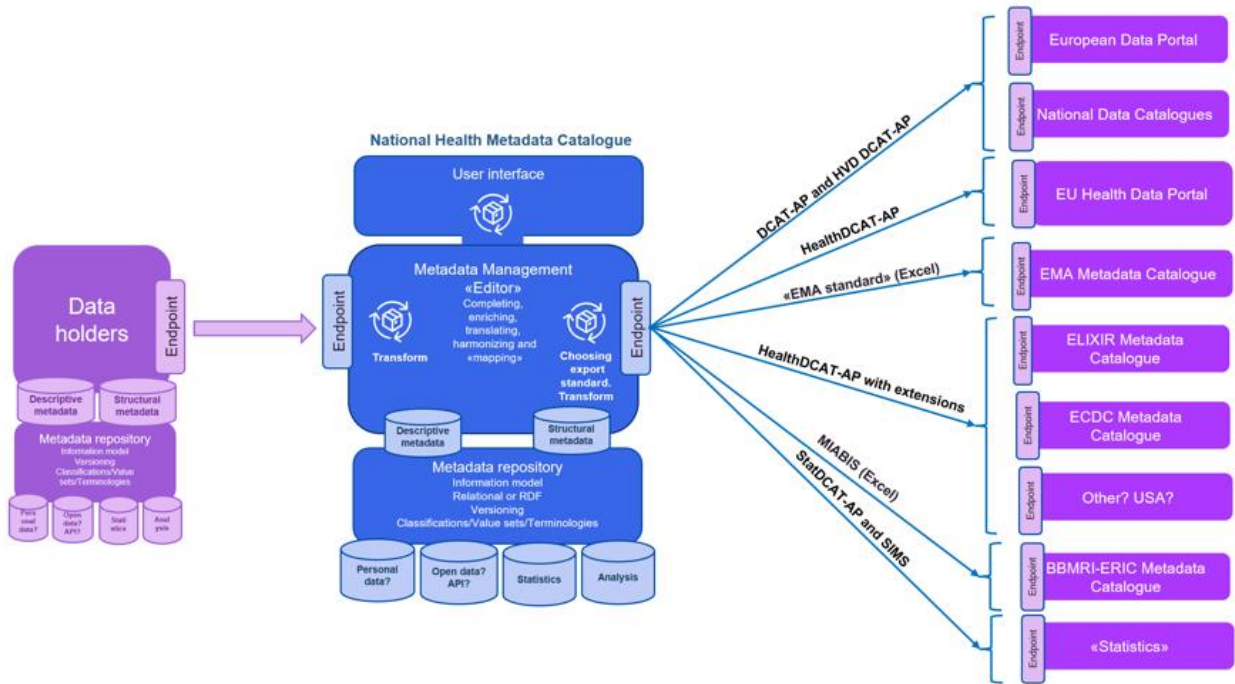
Figure 12: The metadata value chain from data holders, via the HDAB and to different dataset catalogues
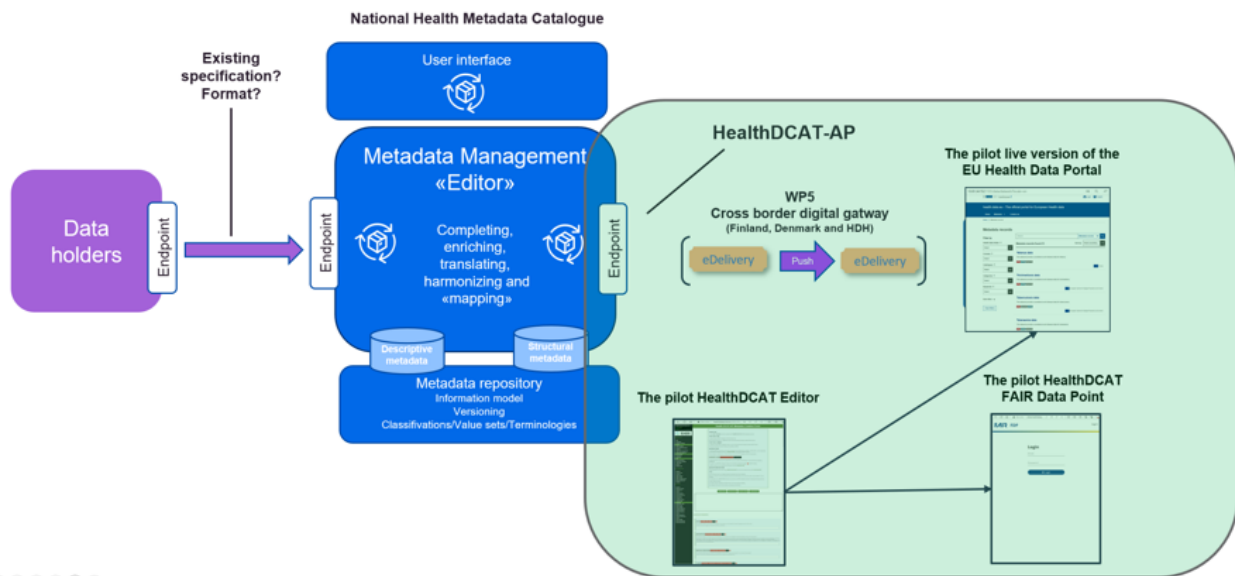

Figure 13: Scope of WP6 – Enabling Participants to produce HealthDCAT-AP compliant metadata and share it with the EU Health Data Portal

The Enabling Environments

To fulfil the responsibilities of both data holders and HDABs, it is essential to consider the necessary enabling environments. The WHO has developed a toolkit for National eHealth strategy development that effectively illustrates this concept, highlighting that for sustainable ICT environments - such as "Services and Applications" and "Infrastructure" - to thrive, certain enabling environments must be in place:

- Leadership, governance and engagement

- Strategy and investments
- Legislation, policies and compliance
- Strategy and investments
- Standard and interoperability
- Workforce

Figure 14: WHO National eHealth Strategy toolkit

The European Interoperability Framework addresses similar concepts, though with a more technical focus, emphasising interoperability as a key priority:

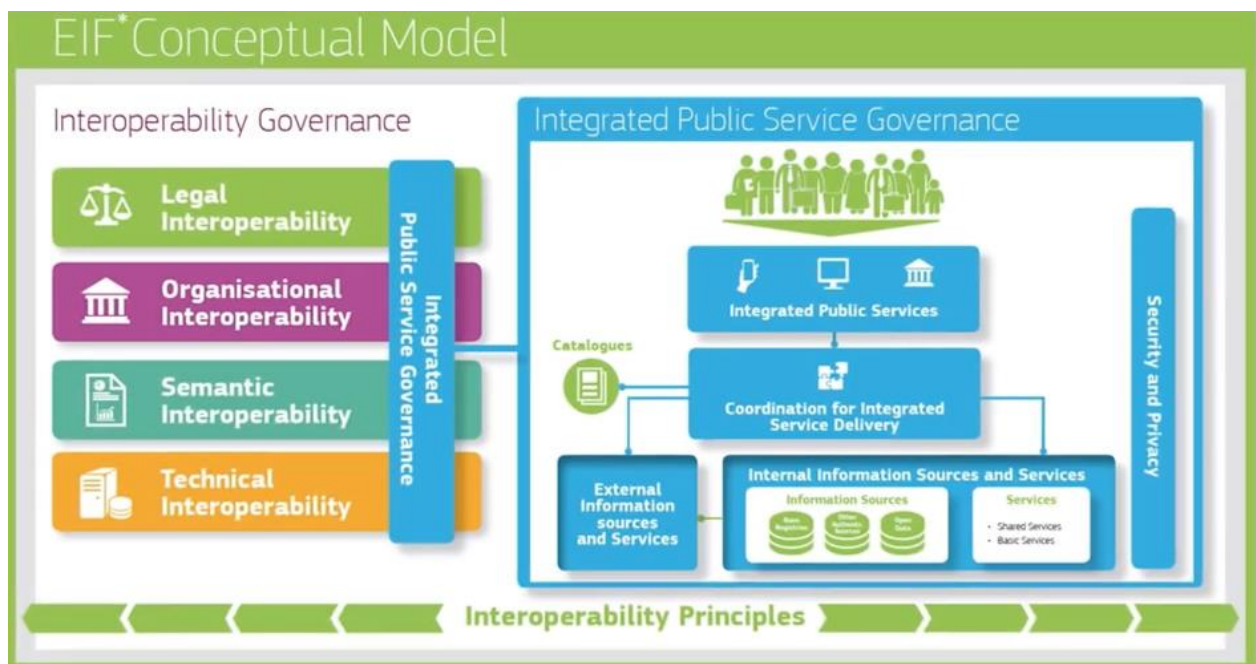Figure 15: European Interoperability Framework - EIF Conceptual Model

6.1. Data holders

The responsibilities and requirements at the data holder level, including establishing an infrastructure for sharing metadata with the National HDAB, were not within the scope of WP6. However, we did address this topic to some extent by posing questions to participants and engaging in related discussions:

DATA HOLDERS	METADATA IN
Who are your dataholders?	How do you harvest your metadata from your data holders? By Excel-sheets?
Do your data holders prepare their dataset descriptions themselves?	Are the metadata according to a specification?
Do they use their metadata to manage their raw data?	Is the metadata harvested as part of a project or as an established routine?
	Do you have routines for updating the metadata?

All these questions have to be taken into further account when making guidelines for both the HDABs especially if the HDABs will be responsible for following up the duties of data holders.

The answers of these questions from the participants in WP6-Task6.3 is summarised in the report on the implementation of the HealthDCAT-AP by the nodes.

The next steps must clarify the responsibilities of the HDABs in implementing, managing, and supporting data holders to ensure they fulfil their obligations:

- Identify responsible metadata managers and anchor this in their organisations
- Set up interdisciplinary teams
- Training and guidance
- Tools
- Setting up endpoints (Infrastructure)
- Capacity building
- Financing

We have found that it is important to designate a "responsible" metadata manager at both the data holder and HDAB levels. Equally important is the formation of interdisciplinary teams, as highlighted in the "Workforce" section of the WHO model. These teams require strong support from their leaders, access to the right tools, and proper training and guidance. A key tool would be a user-friendly, multilingual platform (requiring no technical expertise) for creating and updating HealthDCAT-AP metadata records. One potential solution for this is the HealthDCAT-AP editor, developed by Sciensano and used in the EHDS2 pilot project to support the Work Package 6 activities.

It's important to underline that the data holder is the owner/editor of the

metadata - not the HDAB.

Other important issues that needs further attention are:

- How can feedback from data users on data quality and utility be effectively captured? What should the feedback flow look like, and what role would Health Data Access Bodies (HDABs) play as intermediaries in this process?
- How to ensure feedback from data users if they have enriched the datasets, for example when it comes to samples and new analysis results?
- How to enrich the data and metadata at data holder level with this feedback and routines for updating the shared metadata?

Persistent HTTP URIs for metadata identifiers, as required by HealthDCAT-AP and maintained at the data holder level, are essential for establishing effective feedback mechanisms.

Comment on the importance of user feedback: User feedback plays a critical role in enriching health metadata records. Results and outputs provide valuable insights into how an electronic health data source has been used, which, in turn, informs potential future uses. Therefore, it is essential that data holders are informed of these results or outputs. For instance, health data users should provide data holders with the DOI of any scientific publications derived from the dataset, allowing the data holder to reference the publication in the metadata record of the electronic health data source.

Additionally, health data users must cite the electronic health data source using its metadata identifier in their publications. If they publish their research datasets as open data, they should also include the metadata identifier of the original electronic health data source in the DCAT metadata record of their research dataset. This ensures proper attribution and enhances the traceability and discoverability of both the original data and the research derived from it.

6.1.2. Guidelines for data holders

Guidelines are essential for the successful implementation of the EHDS. In this regard, the Data Provider Manual from data.europa.eu offers comprehensive documentation and guidance for data holders:

[Documentation of data.europa.eu (DEU)](#)

- Who we are
- Our metadata model (DCAT-AP)
- How to search for datasets
- How to publish on data.europa.eu
- High-value datasets
- API Documentation
- Publications and education
- Metadata quality
- Data quality
- Data citation
- Data visualisation
- Accessibility Statement
- Legal notice
- Glossary of terms

6.1.2. Tools for data holders

Guidelines, while essential, are not sufficient on their own to ensure effective implementation and management of metadata records within the EHDS. Data holders, especially in the health sector, need practical tools that go beyond guidelines to support the creation, management, and quality assurance of their metadata. Such tools would enable data holders to apply best practices consistently and efficiently, reducing the complexity, administrative burden and time involved in adhering to standards such as HealthDCAT-AP.

For instance, automated tools for metadata validation can help data holders quickly identify and correct errors, ensuring their records meet required quality standards before publication.
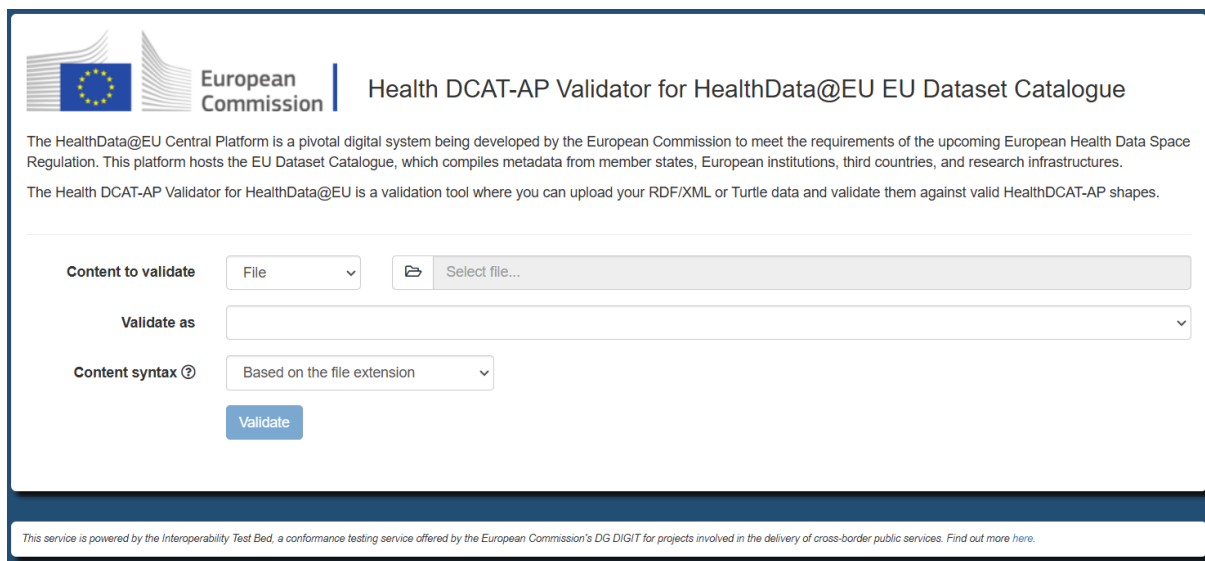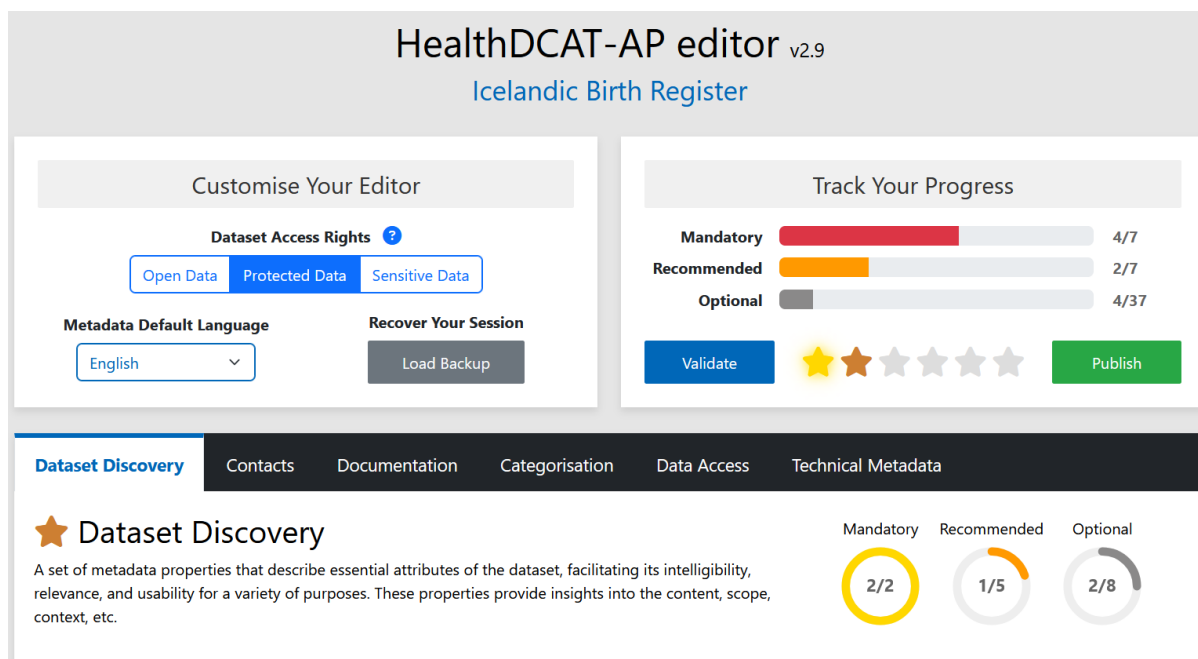
Figure 16: HealthDCAT-AP validator

Additional Information on the Tool: Which processes and tools could be used to manage the quality of metadata?

Additionally, tools for metadata generation and cataloguing can simplify the process of creating new records, helping data holders maintain accuracy and completeness while managing large datasets. By providing actionable support, these tools can significantly enhance the quality, interoperability, and reliability of health metadata records across the EHDS, ultimately contributing to a more integrated and accessible data ecosystem.

Example of a HealthDCAT-AP editor developed by the Innovation in Health Information Systems Unit at Sciensano, the national Public Health Institute of Belgium.

6.1.3. Ensuring multilingualism in metadata generation: Challenges and best practices

Metadata records created in the native language of data holders are often of higher quality than those produced through translation. When metadata is created in the original language, it reflects a more accurate and nuanced understanding of the dataset's context, structure, and content. Therefore, encouraging data holders to create metadata in their native language ensures higher fidelity and a more accurate representation of the datasets.

EHDS Regulation Article 77) Dataset description and dataset catalogue

> *2. The dataset descriptions in the national dataset catalogue shall be available in at least one official language of the Union. The dataset catalogue for Union institutions, bodies, offices and agencies provided by the Union health data access service shall be available in all official languages of the Union.*

DCAT, as an RDF-based standard, inherently supports multilingualism by allowing metadata fields to include language tags. This feature enables data holders to provide metadata values in multiple languages, ensuring accessibility across linguistic groups. By attaching language tags (e.g., `@en` for English, `@fr` for French), DCAT facilitates the seamless inclusion of multilingual content, making dataset descriptions universally accessible in multilingual data catalogues, in line with the requirements outlined in Article 55 of the EHDS Regulation.

6.2. Metadata "In house" (HDAB)

In Task 6.3, the objective was to focus on implementing HealthDCAT-AP within the EHDS nodes of the consortium, as these nodes were responsible for testing and evaluating the

proof of concepts developed by the EHDS2 pilot project. The task also aimed to support the transition from existing metadata templates to HealthDCAT-AP wherever possible.

The objectives of Task 6.3 were clearly defined as follows:

- **Implementation**
  - Produce metadata
    - Either by transition and completing of your existing metadata
    - And/or by using the pilot editor

- **Push the metadata according to HealthDCAT-AP to**
  - First to the FAIR Data Point (and the Sandbox)
  - Then to the pilot of EU Health Dataset Catalogue

- **If you already have a metadata catalogue**
  - Make your existing metadata catalogues HealthDCAT-AP compliant and able to feed into the EU Health Dataset Catalogue

- **If you don't have a metadata catalogue**
  - Set up a basis HealthDCAT-AP compliant metadata catalogue that are able to feed into the EU Health Dataset Catalogue

- **Scope:**
  - All the datasets needed/used in the use cases
    - Both for the central research databases
    - And the datasets initially needed for setting up this databases

- **Assessment**
  - Feedback on HealthDCAT-AP using the feedback form
  - The pilot of EU Health Dataset Catalogue (Together with Task 6.4)

- **It does not include:**
  - Setting up duties for data holders
  - Setting up a national health metadata catalogue for national needs that could be wider than our «mission»
  - Setting up eDelivery as a digital cross border gateway (WP5: Finland, Denmark, France)

Before addressing the deliverables, we asked participants a series of questions to gauge their starting point for the upcoming work. The responses from WP6-Task 6.3 participants are summarised in the report on the implementation of HealthDCAT-AP by the nodes. These insights provide a strong foundation for the further development and deployment of the participating nodes and will guide future efforts in TEHDAS2.

METADATA IN HOUSE	METADATA OUT
Are you a data holder yourself?	How do you share your metadata? Through an API?
Do you have established a metadata catalogue?	Are they in DCAT-format?
How detailed is the metadata? Variables with codes?	Do you transform your existing metadata to DCAT manually or by an micro-service?
Do you use your metadata to manage your raw data?	
Do the reserachers use your metadata to define and specify the data they are applying for?	

To distinguish this step from the data-related responsibilities, we referred to it as "**Metadata In-House**" and "**Metadata Out**" within the metadata value chain. This was done primarily to broaden the scope of what metadata is and how it could be utilised in the future.

We believed that adopting this broader perspective was crucial for setting up flexible **metadata capabilities** within the HDAB. This approach ensures that the system is not only capable of sharing metadata according to HealthDCAT-AP but also adaptable to meet future needs.

By **capabilities**, we refer to the functions and services that a HDAB must manage in relation to metadata. This encompasses both the necessary tools and applications, as well as the human resources required to effectively handle metadata management:

- **Metadata in (Infrastructure):** How to harvest the metadata from the data holders? Setting up endpoints and secure gateways.

- **Metadata management (Applications):** A tool where metadata (and data) managers may improve and transform metadata they receive from data holders. The editor developed and used in the pilot is a very good starting point, but will not cover all the future need for active metadata management in a node.

- **Metadata repository (In house infrastructure):** A place where you store metadata and also code lists, classifications, value sets, terminologies, information models and other concepts that you would like to add to the foundational resources that HealthDCAT-AP relies on (e.g. required controlled vocabularies)

- **The National Health metadata catalogue:** A portal where users can find and explore information (metadata) about data (re)sources and belonging datasets according to the FAIR principles. A portal that is so easy to use that the users can find, define and specify the real data they will add to their application.

- **The metadata GAP:** Experienced metadata managers use metadata for efficient data management, i.e. running scripts for producing standardised data products (datasets) and ad-hoc datasets according to applications. This has to be taken into account when establishing a metadata capability in a node that also are data holder, if not there will be a gap between the metadata and data.

- **Data and solutions for doing analysis** This is quite connected to the "metadata gap". By using the metadata for data management it's a lot easier to set up services for analysis i.e. cohort explorers, analyse solutions on aggregated data, and more.

- **Metadata out (Infrastructure)** Setting up endpoints and gateways for sharing metadata and data need more technical skills than metadata managers normally possess. Experience so far indicates that this could be a "bottleneck" for efficient sharing of metadata.

We recommend that these questions and capabilities be further explored when establishing metadata and data management solutions within the HDABs. Adopting a scope that is too narrow could hinder the potential for more active metadata and data management in the future.

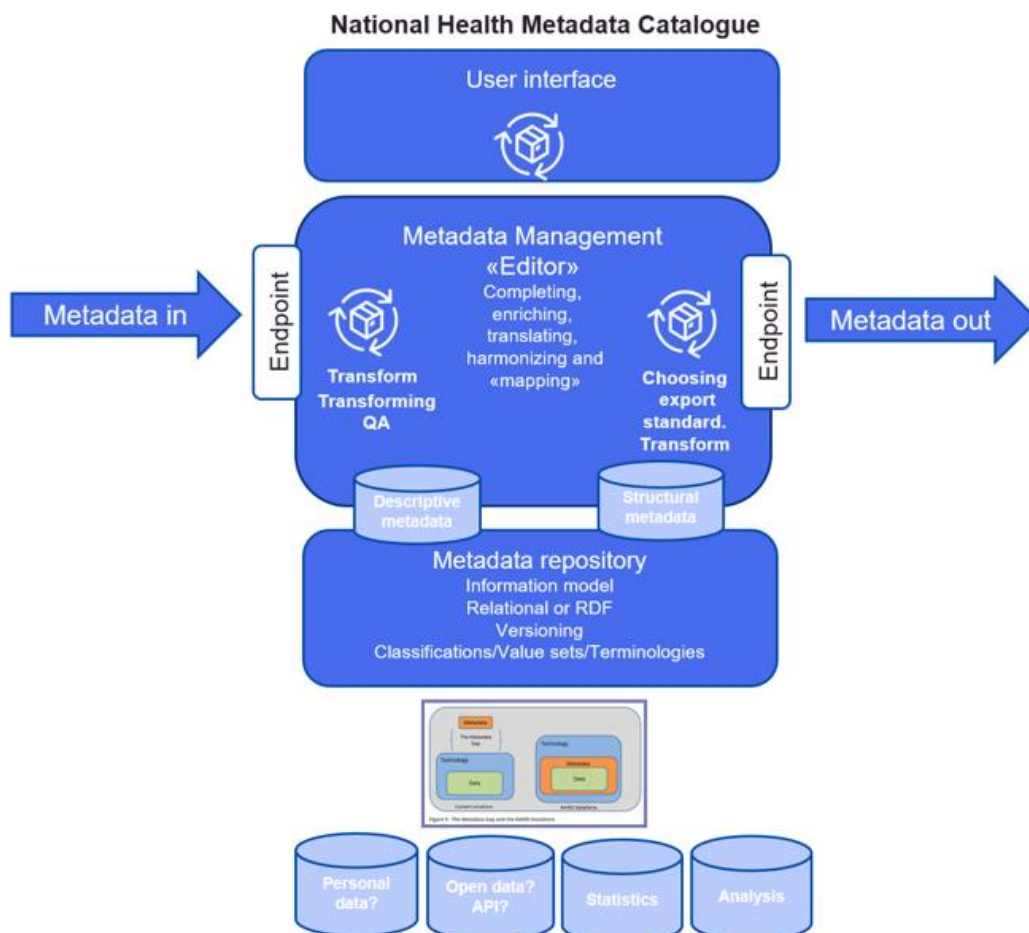We have tried to visualise the main capabilities in the figure below:

National Health Metadata Catalogue

Figure 17: Capabilities of a National health data catalogue

To align with the concrete insights gained from the pilot project, we need to:

- Develop guidelines, functional high level specifications, for setting up capabilities needed to manage metadata (and data) at both data holder and HDAB level.
- Develop technical guidelines for implementers that cover:
    - Transforming existing metadata (if existing) to HealthDCAT-AP
    - A toolkit to edit and maintain DCAT-based metadata including metadata quality validation
    - An interface for sharing DCAT-based metadata from an endpoint and through a secure gateway (eDelivery). First of all with the EU Health Dataset catalogue but also other catalogues like the EU Data Portal.
    - An in-house DCAT-based metadata catalogue with essential functionalities to meet users' needs for discovering, defining, and specifying the data they wish to apply for. The catalogue's core capability requirements include the production of quality labels, management of review periods, and other key features to ensure comprehensive metadata management.
    - An interface between the applications and the data specified in bullet point over.
- Further develop the hHalthDCAT-AP specification:
    - Testing findability and discoverability by using the EU Health Dataset catalogue. Identify candidate properties from users feedback?
    - Need for updating mandatory properties?
    - Need for more precise controlled vocabularies (existing) and/or need controlled vocabularies for special health issues?
    - Documentation of data quality and utility (Specification carried out through the QUANTUM project)
- Ensure interoperability between national health dataset catalogues managed by the HDABs, the EU Health Data Portal, the EU Data Portal, and other European health dataset catalogues, as well as relevant data-sharing infrastructures.
- Revision - maintenance procedures (and responsibility) of HealthDCAT-AP
- Investigate the practical consequences for the data holders and the HDABs

EHDS Regulation Recital 86)

> *The EU dataset catalogue should minimise the administrative burden for the health*
> *data holders and other database users, be user-friendly, accessible and cost-effective,*
> *connect national dataset catalogues and avoid redundant registration of datasets.*
> *Without prejudice to the requirements set out in Regulation (EU) 2022/868, the EU*
> *dataset catalogue could be aligned with the data.europa.eu initiative. Interoperability*
> *should be ensured between the EU dataset catalogue, the national dataset*
> *catalogues and the dataset catalogues from European research infrastructures and*
> *other relevant data sharing infrastructures.*

EHDS Regulation Art. 57) Tasks of health data access bodies

> *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;*
> *The national dataset catalogue referred to in point (j)(i) of this paragraph shall also*

EHDS Regulation Article 77) Dataset description and dataset catalogue

3. The dataset catalogue shall be made available to single information points established or designated under Article 8 of Regulation (EU) 2022/868.

Regulation (EU) 2022/868 (DGA - Data Governance Act): The DGA will apply from 24 September 2023, establishing important guidelines for the re-use within the EU of certain categories of data held by public bodies, as well as a framework for the collection and provision of data brokerage services.

Data Governance Act Article 8 2)

« The single information point shall make available by electronic means a searchable asset list containing an overview of all available data resources including, where relevant, those data resources that are available at sectoral, regional or local information points, with relevant information describing the available data, including at least the data format and size and the conditions for their re-use. »

6.2.1. Validating HealthDCAT-AP metadata and ensuring meta(data) quality

Metadata aggregators, such as HDABs, can incorporate the HealthDCAT-AP validator into their metadata management workflow (see Figure 16: HealthDCAT-AP Validator). This integration enables metadata to be checked for conformance with HealthDCAT-AP, ensuring that mandatory elements are included and that metadata values adhere to the specified data types or reference the controlled vocabularies defined in the HealthDCAT-AP profile.

HDABs play also an active and significant role in ensuring the quality of dataset descriptions provided by data holders. According to Article 37, HDABs are required to "cooperate" with data holders, acting as intermediaries between data users and data holders. They also gather and report user feedback (Recital 39aa), which can contribute to improve metadata quality. This quality aspects is related to the whether the actual values in the metadata are accurate and reflect appropriately what the dataset is and what it is about. This is a much more challenging aspect than checking the conformity of metadata against the HealthDCAT-AP profile and depends for a large part on the understanding of the creator of the metadata and the interactions with data holders.

EHDS Regulation Recital 57)

Health data users who benefit from access to datasets provided for under this Regulation could enrich the data in those datasets with various corrections, annotations and other improvements, for instance by supplementing missing or incomplete data, thus improving the accuracy, completeness or quality of the data in the datasets. Health data users should be encouraged to report critical errors in datasets to health data access bodies

EHDS Regulation Article 57) Tasks of health data access bodies

(d) cooperating with and supervising health data holders to ensure the consistent and accurate implementation of the provisions on data quality and utility label in Article 78;

(i) the number of data quality labels issued by health data holders, disaggregated per quality category;

Health Data Access Bodies (HDABs) play a crucial role in managing metadata for health datasets. While they are not the original owners of this information, they are responsible for maintaining its integrity. To achieve this, HDABs should establish effective communication channels with data holders, ensuring that any updates or corrections to metadata are promptly integrated into their systems.

In the context of their catalogues, HDABs might implement certain technical adjustment. For example, they can automate the completion of the dcatap:applicableLegislation property, which links datasets to the EHDS legal framework. Additionally, the healthdcatap:hdab property, designed to identify the HDAB linked to a health dataset (and vice versa), should be consistently applied across all dataset descriptions in the national health dataset catalogue. In the HealthDCAT-AP specification, these "technical" properties are designated as mandatory for both PUBLIC and RESTRICTED datasets alongside other mandatory properties that have to be necessarily provided by data holders.

6.3. Data users

The scope of WP6 has been to envision the type of information required to make health datasets findable and discoverable for data users, and to "import" this information into an extended version of DCAT-AP, namely HealthDCAT-AP.

The scope of WP6 was not to develop or test a metadata catalogue against the needs of data users, but rather to provide initial dataset descriptions complying to HealthDCAT-AP, enabling users to test and validate the pilot version of the EU health Dataset catalogue.

However, throughout the design process of HealthDCAT-AP, user needs were frequently discussed, and the specification reflects the types of information crucial for making datasets findable and discoverable for users in a health dataset catalogue.

Here are some of the most discussed subjects:

- Which properties should be made mandatory to enable users to filter and query data in a precise and specific manner?
- Do users expect to find detailed information about variables and their corresponding values directly in the EU health dataset catalogue, or through a "drill-down" into national health data portals (e.g., Federated metadata analysis)?
- Will the EU health dataset catalogue be intuitive enough for users to navigate without the need for additional training?

To answer these questions, testing based on realistic use cases is essential. The 5 use cases from WP9 could be leveraged for this purpose, but specific queries and search criteria must be carefully prepared in advance.

User training may also be necessary before testing begins. For example, users should be familiar with effective filtering and search techniques. If the goal involves more advanced searches using SPARQL queries, it will be essential to set up SPARQL endpoints and provide tutorials on how to use them effectively.

Comment: All information in Wikidata is accessible through a SPARQL query interface (query.wikidata.org/), which also supports distributed queries across other Linked Data resources. The potential of leveraging the Resource Description Framework (RDF) and utilising SPARQL as a standard query language will be a crucial area for further testing and exploration in the future.

Health data users should also be informed of their reporting obligations to data holders and instructed on how to properly cite a dataset using HealthDCAT-AP identifiers.

EHDS Regulation Article 61) Duties of health data users

4. Health data users shall inform the health data access bodies from which a data permit was obtained about the results or output of secondary use and assist them to make that information public on health data access bodies' websites. Such publication shall be without prejudice to publication rights in scientific journals or other scientific publications.
When health data users use electronic health data in accordance with this Chapter, they shall acknowledge the sources of the electronic health data and the fact that the electronic health data have been obtained in the framework of the EHDS.

7. Conclusion

To conclude, Work Package 6 (WP6) focused on the development and recommendations for the adoption of HealthDCAT-AP, a DCAT application profile designed specifically for health datasets in the European Health Data Space (EHDS). It is important to clarify what was within and outside the scope of WP6:

Scope of WP6: WP6 aimed to develop HealthDCAT-AP as a standardised metadata framework that ensures interoperability, discoverability, and reuse of health datasets across the European Union. It introduced extensions to the broader DCAT-AP model to address the unique needs of the health sector, aligning with EHDS regulatory requirements.

What was not the scope of WP6: WP6 did not encompass the full technical implementation of HealthDCAT-AP across all member states or the development of operational systems for dataset catalogues. It focused on the specification and recommendations for adoption rather than addressing specific national-level system integrations.

The main elements to retain about HealthDCAT-AP include:

- Its alignment with the **FAIR data principles** (Findability, Accessibility, Interoperability, and Reusability) to enhance the exchange of health data across borders.
- The introduction of **additional metadata elements** specific to the health domain:
  - **Health-specific categories**: Metadata elements such as healthCategory and healthTheme to classify datasets according to health domain-specific standards.
  - Personal Data Categories: Classification of sensitive health data, leveraging the Data Privacy Vocabulary (DPV) to describe personal data types.
  - Quality Annotation: The inclusion of healthdcatap:qualityAnnotation to indicate data quality standards and quality labels based on EHDS

requirements.

  - …
- **Sample Distributions:** For datasets that contain restricted or non-public data, HealthDCAT-AP requires the provision of sample distributions (anonymised or synthetic subsets of the dataset) to allow potential users to explore the structure and nature of the data without exposing sensitive information.
- **Data Dictionary**: HealthDCAT-AP mandates the inclusion of a data dictionary for datasets, especially for restricted or non-public health data. The data dictionary provides a detailed description of all the variables in the dataset, including their definitions, data types, and value ranges. This helps users understand the dataset structure, making it easier to interpret and use the data correctly. It is particularly useful when providing sample distributions, as users need to know what variables are available in the dataset and how they are formatted. This approach promotes transparency and aids in data reuse for secondary purposes like research and analysis.
- **Federated Catalogue and Knowledge Graph**: HealthDCAT-AP is designed to support the development of federated health dataset catalogues within the EHDS, leveraging RDF-based linked data principles. This allows datasets to be interconnected across different catalogues while maintaining interoperability and facilitating the creation of a knowledge graph for health data.
- Ensuring compliance with **EHDS requirements** and other **EU Policies**: HealthDCAT-AP is aligned with EU regulations like the Data Governance Act (DGA), the High-Value Dataset Regulation, the Data Act and the EHDS Regulation. This ensures that health datasets across Europe are described using rich, detailed metadata in accordance with common standards for data sharing and secondary use across the common EU Data Spaces.
- **Support for Advanced Search and AI**: HealthDCAT-AP includes metadata elements that enhance the dataset's usability for semantic search and Generative AI applications. Properties like dpv:hasPurpose and healthdcatap:populationCoverage are designed to improve search accuracy using natural language processing (NLP) and machine learning algorithms.

These elements ensure that HealthDCAT-AP meets the unique needs of the health sector while supporting broader EU-wide data sharing and interoperability efforts across the EU Data Spaces.

The primary objective of the HealthDCAT-AP **uptake strategy** is to ensure widespread adoption by key stakeholders, including health data holders, Health Data Access Bodies (HDABs), and data users. It is essential to demonstrate that using HealthDCAT-AP - and, more broadly, DCAT-AP - for describing datasets in the common EU data spaces is a sound decision. In Task 6.3, the successful implementation of HealthDCAT-AP has laid the groundwork for developing health dataset catalogues that are both compliant with EU regulations and aligned with data user needs. While the pilot focused on enabling participants to produce and share HealthDCAT-AP-compliant metadata with the EU health dataset catalogue, it underscored several areas for further development. These include the roles of data holders and HDABs in managing metadata and the importance of incorporating user feedback.

Looking ahead, the adoption of HealthDCAT-AP will be critical for ensuring the findability,

accessibility, and interoperability of health datasets across the EU. Providing the community with essential tools and clear guidelines will be key to facilitate the smooth implementation of the EHDS. Any outstanding questions without clear answers could slow down this process, which is why this deliverable may appear highly technical. Its purpose, however, is to serve as a living document, enriched by community contributions.

The next steps should focus on refining the capabilities of a health dataset catalogue based on pilot feedback, ensuring it meets user needs, and fostering collaboration between national and EU-level dataset catalogues. We need to see HealthDCAT-AP in action!

By addressing these areas, the project can contribute to building a robust, future-proof infrastructure for health data sharing within the European Health Data Space (EHDS).

8. Annexes

8.1. Interview of Andrea Perego

1. **Introduction of Andrea Perego**
   Pascal asks Andrea to introduce himself and explain his work in metadata standards.

Andrea: "I'm Andrea Perego, currently working for the European Parliament. I've been involved in metadata standards for over 10 years, initially focused on geospatial metadata. I was also invited to contribute to the W3C in developing DCAT, the dataset catalogue Vocabulary. I've worked on various DCAT profiles, including those for scientific data and geospatial metadata, which are widely used across Europe, including in the European Data Portal."

* Andrea is one of the editor of DCAT 3.0 (invited expert), GeoDCAT-AP, CiteDCAT-AP (defining mappings from DataCite to DCAT-AP), DCAT-AP-JRC (an extension to DCAT-AP for multidisciplinary research data), and DCAT-EP (an extension to DCAT-AP used by the European Parliament).

2. **Why DCAT?**
   Pascal asks why DCAT is being developed. Pascal inquires about the motivations behind DCAT's development

Andrea: "DCAT is flexible and not a closed specification, allowing the addition of extra information. It was developed with government data in mind. The DCAT specification provides recommendations on how to describe elements of a dataset, such as the title and other metadata. However, there is nothing that prevents you from adding additional information to the specification. The newer versions, like DCAT 3, provide, for instance, recommendations on how to handle publications and citations, addressing requirements for both government and scientific data. This flexibility makes it possible to enhance data reuse and publication, particularly in the health sector, as you've mentioned. The only potential issue with

interoperability arises if you use a different method to express the same information.

3. **DCAT Uptake Strategies**
Pascal seeks Andrea's advice on how to ensure that the community endorses and adopts tools like DCAT for health dataset catalogues. He asks about the state of the art in metadata catalogue technologies in Europe, expressing concern about CKAN not being an RDF-based technology. He discusses the challenges of selecting compliant tools for powering metadata catalogues, sharing an example where proprietary, non-RDF solutions led to interoperability issues. Pascal emphasises the importance of using compliant, standards-based technologies and seeks Andrea's input on this issue.

Andrea: "The key is to avoid vendor lock-in and focus on open standards. Using DCAT ensures that your metadata remains interoperable, but it's important to separate catalogue management from the front-end used for searching and editing. Using standards like DCAT makes it easier to replace systems while keeping your data intact. Tools like Piveau and Entryscape are available to help create compliant metadata catalogues without deep technical knowledge of RDF. For performance reasons, catalogues often use a combination of technologies, like triple stores and Elasticsearch, for efficient data retrieval. Piveau is already being used in the European Data Portal and could be adapted to your specific needs. Aligning with these standards ensures long-term interoperability."

Andrea: "CKAN is becoming outdated and is often replaced by more efficient solutions. Many catalogue providers use open-source platforms like Entryscape and Piveau. Entryscape allows flexibility in adapting metadata schemas, while Piveau is a modular, open-source platform used by the European Data Portal. Both avoid vendor lock-in, ensuring that you can replace tools without losing data. Piveau, especially, is a good fit for health metadata catalogues since it supports metadata harvesting and quality assessment."

4. **Use of persistent dereferenceable identifiers**
Pascal asks for Andrea's opinion on the use of persistent dereferenceable URIs for metadata catalogues and whether this aspect is central.

Andrea: "For this property, there has been extensive discussion in the SEMIC community over the past years. The challenge has been how to manage metadata in federated harvesting environments. It started with the European data portal, which aggregates metadata from various sources. The question was how to trace the metadata back to its original catalogue and specific record. The solution was to establish a practice of using this property to retain the original URI of the metadata record, enabling users to trace it back to the source catalogue. This is now a common approach."

5. **Metadata management**
Pascal mentions the EHDS regulation, which requires metadata records to be revised annually. He highlights that this would involve a central European health metadata catalog, where metadata records would include a timestamp for revision, and asks if

there is a property to handle different versions of metadata records.

Andrea: "This implies having different versions for the same metadata record, not just the dataset. For example, a dataset may have different versions due to changes in the actual data. This is a key features of DCAT 3 to support the versioning of datasets.But even a small correction, like changing the title, would create a new version of the metadata record. This versioning applies to catalogue records in DCAT. The system does not typically retain versions of metadata records. While DCAT includes a modification date for a catalogue record, it doesn't provideversioning management for it. It's up to the implementation whether to keep different versions, but typically, catalogues systems should automatically maintain modification dates without manual intervention. To conclude DCAT Dataset versions can be applied to catalogue records as well (i.e., version chains and hierarchies, version replacing other versions, version information, and statuses / lifecycle."

6. **GenAI requirements**
   Pascal brings up the increasing relevance of GenAI and its potential to power next-generation metadata catalogues. He discusses the need for metadata, especially free text, to provide context like provenance, purpose, and publisher information, which are crucial for AI processing. He notes that current standards like DCAT don't fully support this, prompting the idea of adding specific properties to better serve AI.

Andrea: "I agree that full-text information will become increasingly important. For example, GeoDCAT-AP maps provenance information into free text." Andrea supports the idea of enhancing publisher information and dataset descriptions with both control vocabularies and free text, while still linking to registries for consistency.

7. **Use of Wikidata**
   Pascal explores the use of Wikidata as a common reference point for health-related metadata, especially in contexts where different coding systems are used. He discusses the challenge of using keywords versus concepts in HealthDCAT, especially considering the vast number of coding systems in health (e.g., ICD-10). He suggests using Wikidata as a way to express and organise these concepts, allowing for a more flexible and machine-readable format by creating SKOS concepts linked to Wikidata.

Andrea: "Wikidata can be a good reference platform, especially for maintaining code lists, for simulating a control vocabulary for health themes, making metadata more interoperable and actionable. However, it's essential to ensure it covers your specific requirements. Using Wikidata to simulate a control vocabulary makes sense, and it's sustainable for long-term use, but governance will be necessary to prevent inconsistencies. Similar work has been done by SEMIC (https://joinup.ec.europa.eu/collection/semic-support-centre/wikidata-and-wikibase)."

Andrea: Introduces a project called " Kohesio Project" (https://kohesio.ec.europa.eu/en/) where similar work was done involving the loading of NUTS data into a dedicated instances of Wikidata (https://linkedopendata.eu/wiki/The_EU_Knowledge_Graph). He emphasizes that

Wikidata is collaborative and can be useful for multiple stakeholders. It can help with interoperability and collaboration across organisations, provided there is proper governance to prevent issues.

8. **Sensitive Data and Distribution Strategies**
   Pascal discusses how to handle sensitive data that isn't publicly accessible. He asks Andrea for solutions on how metadata catalogues can represent sensitive datasets without disclosing personal data. He explains that for sensitive datasets, the distribution users receive through the healthdata@EU infrastructure will often be a landing page for a national node, not direct access to the data. He mentions reusing the adms:sample property in DCAT for non-publicly accessible datasets, allowing the provision of sample data or structural information.

Andrea: "You can handle sensitive data by using sample distributions or anonymised versions of the data to avoid disclosure. Creating synthetic data for testing and training AI models is a common practice. It's important to provide structure and schema to describe datasets, even if the actual data cannot be accessed directly.

   Pascal adds that most data in health are tabular, and encourage the use of CSVW.

Andrea: "I agree that CSVW could be promoted within your community, unless RDF data cubes are preferable. However, it's essential not to enforce it universally, as not all data are tabular. Encouraging stakeholders to adopt best practices like CSVW can improve the reuse and interoperability of datasets."

9. **Dashboards**
   Pascal discusses how certain data holders provide access to dashboards (e.g., ECDC) for sensitive datasets, allowing users to access metrics without requiring access to sensitive information. He explains that HealthDCAT-AP added a new property (analytics), similar to adms:sample or dcat:distribution, to enable users to discover datasets via analytics interfaces. This helps reveal insights about the dataset without disclosure of sensitive information.

Andrea: "So, the distribution provides a URL, and when clicked, it shows only analytics about the dataset, correct? Based on what you're describing, it makes sense. So, this new property, healthDCAT analytics, is treated the same as distribution and sample and is not a subproperty of distribution? Okay, so it's not linked, but you could make it a subproperty of distribution. This reminds me of how geospatial information is linked to WMS services, where users receive images as a type of view or visualisation. Your case could be handled similarly, but it's fine to treat it as a separate distribution. The key here is that you never get the actual data, just the metrics.

Pascal: In the healthdata@EU infrastructure, users follow a process that includes discovering datasets, selecting variables, applying for access, and obtaining a minimized version of the

dataset in a secure processing environment to generate insights. With analytics interfaces, we believe that DCAT can streamline the EHDS user journey, improving dataset discovery while safeguarding sensitive data from exposure.

Andrea: "Okay, it doesn't seem like a strong requirement, but it makes sense in the context you've described."

10. **Health Data and Quality Labels**
    Pascal explores the idea of using DCAT to assess health datasets' quality.

Andrea: While DCAT doesn't enforce strict standards, it offers flexibility for integrating quality metrics, such as conformity with standards. There are existing vocabularies, like the Data Quality Vocabulary (DQV), to manage these metrics. However, I'm not fully aware of the specific quality metrics required for health data."

8.2. Functional analysis of implementing HealthDCAT-AP in the EU health dataset catalogue

8.3. EUPHA Slides