Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Towards Semantic Interoperability in Secondary Use of Health Data: Development and Evaluation of a HealthDCAT-AP-Based Metadata Model

Jaleh Shoshtarian Malak

jaleh.malak@uni.lu

University of Luxembourg

Somayeh Abedian

Ludwig Boltzmann Institute for Digital Health and Prevention

Leonie Berlin

Fraunhofer Institute for Translational Medicine and Pharmacology

Hanna Ćwiek-Kupczyńska

University of Luxembourg

Francisco J. Núñez-Benjumea

IBIS/Virgen Macarena University Hospital, CSIC/University of Seville

Beatriz Garcia Santa Cruz

University of Luxembourg

Jens Declerck

Ghent University

Carlos Lopez Gomez

Instituto de Investigación Sanitaria La Fe (IIS La Fe)

Lourdes Martínez-Martínez

IBIS/Virgen Macarena University Hospital, CSIC/University of Seville

Alberto Moreno-Conde

IBIS/Virgen Macarena University Hospital, CSIC/University of Seville

Rada Hussein

Ludwig Boltzmann Institute for Digital Health and Prevention

Philip Gribbon

Fraunhofer Institute for Translational Medicine and Pharmacology

Venkata P. Satagopam

University of Luxembourg

Research Article

Keywords: Metadata, HealthDCAT-AP, Secondary use of Health Data, European Health Data Spaces (EHDS), OMOP-CDM, Data Quality

Posted Date: June 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9827602/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Aligned with the European Health Data Space (EHDS), the HealthDCAT-AP metadata model supports the semantic description of health datasets, making them findable and interoperable across EU-level catalogues. However, project-specific requirements driven by disease areas and data models should be incorporated on a per-project basis. In the context of the IDERHA project, which aims to establish a federated data space for the secondary use of heterogeneous health data with an initial focus on lung cancer, we designed a customized metadata model extending HealthDCAT-AP to meet these requirements.

Methods

We derived the IDERHA metadata profile based on HealthDCAT-AP by: (i) adopting all mandatory classes and properties and selecting a preferred subset of recommended and optional properties based on working-group input and data holders' requirements; (ii) modeling legal bases and conditions of use using DPV and mapping them to ELIXIR DISH concepts; (iii) extending the profile with OMOP CDM–derived statistics and data-quality indicators; and (iv) adding clinical and imaging measures. The metadata model is represented as a JSON Schema, and Protégé 5.6 was used to develop a semantic ontology. A sample metadata instance and a preliminary usability assessment were developed to evaluate adoption feasibility.

Results

The proposed model comprises 116 properties organized into five categories — HealthDCAT-AP-derived dataset descriptors, OMOP statistics and data quality, data privacy and conditions of use, clinical statistics, and study metadata. A key contribution is the adaptation of HealthDCAT-AP to project-specific requirements, incorporating OMOP CDM–based aggregate statistics and data quality indicators, as well as tailored data privacy and study-level properties. The JSON-LD based samples were validated against HealthDCAT-AP SHACL shapes, and the usability assessment confirmed strong discoverability (avg. 4.5) and adoption potential (avg. 4.3). However, the complexity and effort required to complete the metadata remains a concern (avg. 3.5).

Conclusion

This work presents a tailored, HealthDCAT-AP-compliant metadata model addressing real-world gaps in federated health data environments through project-specific extensions. The proposed methodology

introduces a reusable framework for federated secondary-use health data infrastructures aligned with EHDS, adaptable to other disease areas and advancing FAIR principles.

1. Background

Rich metadata plays a crucial role in healthcare to ensure data are Findable, Accessible, Interoperable, and Reusable (FAIR principles)[1]. Improving metadata models to support health research and incorporate disease-specific metadata enhances accessibility for research and analysis[2]. There have been several efforts to develop actionable designs for Health Metadata Catalogs (HMDCs) to address metadata harmonization gaps in the European health care industry[3].The European strategy for data aims to make datasets more widely available and reusable across the EU[4]. As part of that strategy, the European Health Data Space Regulation (EHDS)[5] intends to enable and empower researchers and innovators for the secondary use of health data. A key component of EHDS is the creation of an EU-level dataset catalog. The DCAT Application profile for data portals in Europe (DCAT-AP) uses Data Catalog Vocabulary (DCAT) to exchange dataset specifications and make public sector datasets more findable and accessible. HealthDCAT-AP[6], on the other hand, is an extension of DCAT-AP, adding new classes, properties, constraints and controlled vocabularies- based on health data requirements, aiming to harmonize cataloguing and metadata specification within European data space. It enables the integration of metadata across local, national, and EU-Level catalogues including HealthData@EU. To address the legal dimensions of data processing, particularly those related to the General Data Protection Regulation (GDPR) and conditions of dataset processing use in a machine-readable format, the Data Privacy Vocabulary (DPV)[7] was introduced. HealthDCAT-AP leverages DPV to formally represent legal bases and metadata related to personal data. The Data Quality Vocabulary (DQV)[8] provides a specification for describing dataset quality metadata. It is an extension of DCAT and supports the representation of user feedback, dataset standards, data quality policies, agreements, and metrics related to both quantitative and qualitative aspects of data quality. Another metadata model is the Data Tag Suite (DATS)[9], designed to consistently describe datasets for discovery, particularly in the biomedical and life sciences. It structures dataset metadata into core, nested entities (such as project, study, dataset) and is compatible with standards like DCAT and schema.org to ensure interoperability[10]. Furthermore, the Data Information Sheet (DISH)[11], used by ELIXIR Luxembourg[12] in translational biomedicine projects, serves as an information collection instrument for documenting data use restrictions and GDPR requirements in projects that typically involve multiple clinical study sites with several cohorts.

To facilitate collaborative research across the Observational Health Data Sciences and Informatics (OHDSI)[13] community, OHDSI has conducted studies to investigate data characterisation and quality of different OHDSI datasets. As part of this effort, they have proposed a set of aggregate statistics and data quality measures to assess the Observational Medical Outcome Partnership (OMOP) Common Data Model (CDM)[14] based datasets' feasibility and readiness to run studies. The European Federation for Cancer Imaging (EUCAIM)[15] proposed a metadata model for describing cancer imaging datasets and

registering them in their public catalog. They considered the cancer-related metadata requirements of the AI4HI[16] projects and incorporated them into their metadata model.

The Integration of Heterogeneous Data and Evidence towards Regulatory and HTA Acceptance (IDERHA) [17] project establishes a federated data space that enables standardized, secure access to heterogeneous health data sources, with an initial focus on lung cancer. The metadata of lung cancer-related datasets (in conformance with OMOP-CDM) will be stored in a central data catalog and will support data discovery for the purpose of executing artificial intelligence and machine learning algorithms for data analysis in secondary use of health data. Beyond enabling discoverability within EHDS-aligned infrastructures, the need for harmonized and interoperable metadata is also driven by the increasing reliance on ML methods in biomedical research. As shown in previous work[18], insufficient documentation of dataset generation processes and the absence of standardised, machine-readable metadata hinder the identification of confounders, promote spurious associations, and ultimately reduce the robustness and generalisability of ML models. In federated environments such as IDERHA, where several types of heterogeneous data sources exist, there is a need for a common and standardised way to share metadata transparently and in a reusable manner. Despite these efforts, existing metadata models and standards do not fully address the requirements of federated health data environments that handle heterogeneous, disease-specific datasets. In particular, there is a lack of project-tailored metadata profiles that combine HealthDCAT-AP compliance with OMOP CDM-based data quality indicators, machine-readable data privacy specifications, and domain-specific clinical and imaging properties within a single interoperable framework. Furthermore, the absence of standardized, machine-readable metadata hinders the identification of confounders, promotes spurious associations, and reduces the robustness of machine learning models in federated settings. Building on existing standards and initiatives, this paper presents the development and evaluation of a tailored, HealthDCAT-AP-based metadata model for the IDERHA project, designed to improve semantic alignment and support reliable secondary-use analytics across heterogeneous health datasets while ensuring compliance with EHDS specifications and advancing FAIR principles.

2. Methods

In this section, we describe the main steps we took to build the IDERHA metadata model. Figure 1 provides a summary of these steps.

2.1 Building a derived metadata model based on HealthDCAT-AP

HealthDCAT-AP extends DCAT-AP by introducing new classes, properties, and constraints, thereby increasing health data accessibility while providing better privacy and security for sensitive information. To comply with HealthDCAT-AP Release 5(September 2025), it is essential to meet both the DCAT-AP conformance requirements and controlled vocabularies, along with the additional specifications introduced by HealthDCAT-AP. Since IDERHA provides closed access to users, we chose the HealthDCAT-

AP restricted access level profile, and we have utilized all mandatory properties related to the key entities, including Datasets, Catalogues, Distributions, Data Services, and Agents, as well as the corresponding controlled vocabularies.

HealthDCAT-AP defines three profiles for public, restricted, and non-public data access, each with different mandatory, recommended and optional properties. Although some properties are identified as recommended in the HealthDCAT-AP specification, the provider should implement them if they exist at the data holder's site. Additionally, for conformance to HealthDCAT-AP, the receiver must support both the recommended and optional properties. In this context, the provider (a data holder) is responsible for exposing required and recommended metadata, while the receiver (e.g., a catalogue) is responsible for consuming and interpreting these properties.

At IDERHA, we have selected certain recommended and optional properties to include in our metadata model based on the following criteria:

- Selected by the Data Holders Working Group
- Compatibility with DATS and DISH specifications
- Compatibility with the GA4GH specification
- Compatibility with the metadata from the data holders' synthetic data

Within the IDERHA Data Holders Working Group, representatives from hospitals and project work packages discuss various use case requirements, metadata specifications, and implementation guidelines for the project. Currently, some of the data holders produce metadata about the synthetic data they have for the project. Some of these properties may correspond to the recommended and optional properties of HealthDCAT-AP, while others may be added to the project metadata model as custom properties (described in the following section).

Additionally, we included some optional properties to comply with the GA4GH Passport[19] specification, which is the standard used in the project to manage security for authorizing data access. In case of any conflicts, we tried to discuss them in group meetings and finalize the decision on whether to include or exclude a property based on the majority's opinion.

2.2 Define Legal Basis and Condition of Use

The legal basis and conditions of use for the project were specified based on the project's data privacy requirements and the level of detail required for implementation in a machine-readable format.In HealthDCAT-AP, the Data Privacy Vocabulary (DPV) standard is used to represent the metadata related to these aspects. Currently, the DISH specification is used by project partners within the ELIXIR infrastructure as an established model for gathering information on data privacy and conditions of use. Therefore, we mapped it to the DPV specification to create a comprehensive, DPV-compatible model. The main effort during the definition and mapping to the DPV specification was identifying the most suitable matching concepts and representations. Concepts related to personal data, data anonymisation,

GDPR articles, the processing of data subjects, and other processing limitations were assessed, and corresponding concepts from the DPV were assigned. The detailed mapping between the DISH condition of use variables and the related DPV concepts can be found in Additional File 1[20].

2.3 Mapping/Adding properties from existing standards/ models

Existing standards and metadata models, such as DATS and DISH, have been assessed, and some properties have been included based on them. The mapping process is described in the following section. We assessed the working version of the DATS metadata schema and attempted to map its properties to HealthDCAT-AP. Properties that did not have a corresponding concept for mapping were identified and added as new entities to the HealthDCAT-AP base model as custom properties. Considering data privacy and conditions of use, we noted that there is currently no clear specification for the requirements in federated data analysis environments, and HealthDCAT-AP considers this a project-dependent, non-normative concept. We also aimed to adopt the DISH data model, mapping its items to DPV concepts, and incorporating them into our metadata model. The list of the mappings from DATS and DISH to HealthDCAT-AP is provided in Additional File 1 [20].

2.4 OMOP CDM statistics and data quality properties

The OHDSI community evidence network study protocols have established data aggregation benchmarks and characterization methods that summarize the contributing OMOP CDM databases[13]. Several R packages, such as DbDiagnostics, built upon Achilles[21] and the Data Quality Dashboard(DQD)[22], generate descriptive statistics and apply parameterized checks to an OMOP CDM instance to evaluate data quality. The OMOP-CDM data quality assessment framework categorizes quality aspects into three dimensions: conformance, completeness, and plausibility[23]. For IDERHA, we have incorporated these statistics as metadata properties, enabling users to gain a clearer understanding of the data characteristics provided by each contributing data source. Some of these metrics are evaluated by different dimensions and forming a data cube structure. The Data Cube Vocabulary[24] and Statistical Data and Metadata eXchange (SDMX) specification are used to represent these measures.

2.5 Clinical and imaging metadata

IDERHA's initial focus is on the analysis of diverse health data from individuals with lung cancer. The EUCAIM project, focused on cancer image data, has defined a metadata model and a hyper-ontology model at three different TIERs for metadata, federated search, and processing[15]. At TIER 1, the Public Catalogue Metadata Model extracted clinical and imaging metadata requirements from the AI4HI projects and built its TIER 1 domain-specific metadata model. Consequently, at IDERHA we aimed to comply with and align to their metadata model, incorporating their cancer-related, domain-specific and imaging metadata properties.

2.6 Metadata validation and Evaluation

We use SHACL shape constraints [25] to validate that the metadata we create, based on our defined schema(JSON Schema), complies with our conditions and constraints. HealthDCAT-AP provides SHACL shapes for different metadata profiles (public, restricted, non-public). These RDF-based shapes also define class and range constraints that we use to assess conformance with the recommended properties. We used the SCHAL shape validator tool provided by the Interoperability Test Bed by the European Commission ‘s DG DIGIT[26] to validate our metadata (JSON-LD file) against the SHACL shape files. A preliminary usability assessment of the IDERHA metadata model was conducted. This assessment evaluates, for each metadata category, the clarity, understandability, and distinction between fields. Additionally, the overall usability of the model and the feasibility of its adoption were assessed. The participants included 4 people from 2 of our data holder sites and 2 researchers.

3. Results

3.1 Proposed Metadata Model

We have created a metadata model that allows all data providers to define their metadata, enabling interoperability at the dataset metadata level in compliance with EHDS specifications. The metadata is categorized into five main categories (Fig. 2): HealthDCAT-AP derived, OMOP Statistics and Data Quality, Data Privacy, Clinical Statistics, and Study Metadata. Figure 2 provides a list of IDERHA metadata categories and their related properties. The full list of the properties in each category with associated details can be accessed in Additional File 2, and the JSON schema of the model is provided in Additional File 3[20].

The main HealthDCAT-AP classes included in our metadata model were the Dataset, Catalog, Distribution, and Agent class. The OMOP CDM statistics include summary statistics obtained from the Person, Visits, Measurements, and Observations, Drugs and Procedures OMOP-CDM tables. The data quality measures are represented using DQV for each measure (qb:measureType) and the dimensions (qb:structure, qb:observation) in the form of objects and values (sdmx measure:obsValue). As shown in Fig. 3, each measure—for instance, “Number_of_persons_by_gender”—is stored as object values for each observation. Similarly, “Number of persons by observation period” represents object values for each time period. Considering the clinical and imaging-related variables, we included all the EUCAIM domain specific properties from their hyper-ontology (number of studies, number of series, image modality, image vendor, condition, body part).

In conformance with the DPV specification, we have used the related dpv:Process concepts to represent categories and types of personal data (dpv:hasPersonalData), anonymization/pseudonymization (dpv:hasTechnique dpv:Anonymization), the collection of sensitive personal data, GDPR Article 6.1 (dpv:hasLegalBasis dpv:Consent), the collection of special categories of data, GDPR Article 9.2 (dpv:hasLegalBasis dpv-gdpr:A9-2-a), and the scope of the research (dpv:hasScope). Figure 4 shows a section of a sample JSON representing the conditions of use based on DPV.

The clinical statistics section indicates the number of individuals, typical age, birth, sex, number of studies,number of records, study series, conditions, and image modality. To formalize the metadata model, we have created an ontology using Protégé version 5.6 to better describe the related concepts and ontologies. We imported three existing external metadata vocabularies, including DCAT-AP version 3, DPV, and DQV, and merged them into our IDERHA ontology model, and related HealthDCAT-AP properties and constraints were added to the model. Figure 5 demonstrates the ontology overview in Protégé ontoGraf class view Additional File 4. The selected aggregated properties related to clinical statistics and OMOP CDM data quality were defined as data properties.

3.2 Coverage Analysis and Comparison

The proposed metadata model contains 116 properties organized into five categories. We assessed the coverage of HealthDCAT-AP Release 5 properties by the proposed model, DATS, and DISH, focusing on the main Data Catalog and Dataset properties and evaluating their coverage of mandatory, recommended, and optional elements (see Table. 1). The results show that the IDERHA model fully covers all mandatory properties (100%), ensuring strong compliance with HealthDCAT-AP, while selectively including recommended (41.38%) and optional (25%) fields. DATS achieves good but lower mandatory coverage (76.2%) with a higher proportion of optional fields, whereas DISH has the lowest overall coverage, as it is designed to focus more specifically on data privacy and conditions of use.

Table.1 Presents a comparative coverage analysis of HealthDCAT-AP Release 5 mandatory, recommended, and optional properties across three metadata models: the proposed IDERHA metadata model, DATS, and DISH.

Proposed-IDERHA Metadata Model	**# of properties**	**Mandatory Coverage (21)**	**Recommended Coverage (29)**	**Optional Coverage (35)**
	116	**100% (21)**	**41.38% (12)**	**25% (10)**
DATS	57	76.2% (16)	41.38% (12)	34.3% (12)
DISH	67	38.1% (8)	13.8(4)	20% (8)

3.3 Experimental use case: building a sample metadata

We have developed a metadata instance based on the proposed metadata model and vocabulary, in collaboration with one of the project's data holders. To extract statistics and data quality metadata related to their OMOP database, we used a synthetic OMOP CDM database derived from their hospital EHR system for lung cancer patients. Considering the initial project arrangements, the database did not include all patient-related data, specifically, data related to observations, measurements, and procedures. Therefore, we included only the properties corresponding to the available data. We used the Achilles and DQD packages to generate statistical summaries and data quality measures, and subsequently extracted the metadata from their result tables stored in a PostgreSQL database. Currently, a metadata catalogue is under development based on the proposed metadata model, through which metadata will be

submitted, managed, and shared. However, for now, we used Visual Studio Code to construct sample metadata (Additional File 5).

In addition to this sample, we also developed a second sample that includes all the proposed metadata variables to showcase the model, which was shared with the data holders.

Both the samples were validated using the SHACL shapes provided by HealthDCAT-AP Release 5 (attached). The first sample (developed based on synthetic data holders' data) was validated against the Non-Public shapes, while the second sample was validated against the Restricted shapes. The validation process was performed iteratively, as some directives and metadata structures did not conform to the shapes. Therefore, we revised the sample multiple times until the validation succeeded (Additional File 6).

3.4 Preliminary usability assessment of the metadata model

As demonstrated in Fig. 6 The overall assessment results indicate that the proposed model is perceived as valuable in terms of usability and discoverability (avg. 4.5), with a high likelihood of adoption (avg. 4.3). However, a main concern relates to the effort required to complete the metadata (avg. 3.5). Considering the core metadata model (derived from HealthDCAT-AP), some ambiguity and understandability issues were reported for fields such as theme, health-theme, healthCategory, and rights/access rights. Additionally, some respondents reported difficulty understanding the abstract concepts related to GDPR and DPV, which requires more clarification and specialised training. The OMOP statistics and data quality checks require prior knowledge of OMOP concepts and the DQD tool to avoid confusion. It is also recommended to include data provenance information prior to the ETL process, in order to distinguish between a clinical gap (expected data missing due to patient state) and a systemic omission (data loss during ETL). Furthermore, some data holders may not currently have data available for all OMOP CDM entities. Querying and manually extracting these measures is complex and imposes a significant workload, highlighting the need for tools to automate this process. During the review, several properties were suggested for inclusion in the Study category, namely study duration and status, protocols, and study population. The details of the questionnaire and the related heatmap can be found in the Additional File 7.

4. Discussions

Defining and implementing a well-structured, standardized, and vocabulary-enriched metadata model supports realizing the FAIR principles. Currently, there are several aspects under discussion to generate high-quality metadata. For instance, capturing information regarding data provenance is usually difficult, since there is no clear documentation, and sometimes complex ETL processes are involved. Enriched metadata that contains contextual and domain-specific information increases the reusability, accessibility, and maintainability of data in federated health data spaces[27]. However, there is a trade-off between the level of detail and complexity that we decide to include from a medical knowledge

domain and the need for data minimization strategies to exclude unnecessary data[28], while researchers prefer to zoom into the data as much as possible[29]. For instance, while extracting data quality metadata according to OMOP DQD, we had a long list of metadata, including measures reported based on different concept dimensions, which overcomplicated the metadata instances. Another challenge we faced in IDERHA was determining the extent to which we should describe data privacy and conditions of use in a machine-readable format using the DPV specification, so that the system can automatically process them. On the other hand, the EU AI Act extends the DPV specification to cover requirements specific to the AI Act by defining concepts such as systems, purposes, risks, and roles. We can build on this to further extend our metadata model[30].

In IDERHA, we aimed to align with the EUCAIM project's specification at their TIER1 metadata level (high level aggregated dataset metadata) dataset metadata level, taking into account clinical and imaging properties related to lung cancer. The EUCAIM project also proposes a CDM and a HyperOntology to enable semantic and synthetic interoperability[31]. However, at the federated search and processing level, EUCAIM handles patient-level data by harmonizing OMOP CDM and FHIR-based data sources through its own CDM and HyperOntology. In contrast, within IDERHA, patient-level data follows the OMOP CDM and its related developed by the OHDSI community, including those developed by Oncology, imaging and GIS workgroups[32], [33]. Regarding the data-quality aspects of the datasets in this research, we considered measures derived from the OMOP CDM. However, we have an ongoing subtask within IDERHA to define a data-quality framework and to advance our data-quality metadata metrics to cover more details on data provenance and the sources of integrated or transformed data. The output of these efforts will be further incorporated into our metadata model.

5. Limitations

This work is subject to some limitations. The metadata model is complex to complete, particularly for OMOP CDM-based data quality indicators, which cover a large number of variables and require prior knowledge of OMOP concepts and the Data Quality Dashboard tool. This highlights the need for user-friendly interfaces and automated metadata extraction pipelines to reduce the burden on data holders and improve adoption feasibility. Furthermore, the validation and usability evaluation were conducted on synthetic data with a limited number of data holders and participants, which does not fully reflect the complexity of real-world health datasets at scale. A more comprehensive evaluation involving real patient data and a larger number of data holders across multiple institutions is needed to validate the model's robustness in production environments.

Finally, both HealthDCAT-AP and the TEHDAS2 technical guidelines are still under active development. New releases of HealthDCAT-AP may introduce changes to mandatory properties or conformance requirements that could necessitate revisions to the IDERHA metadata profile, while TEHDAS2 does not yet fully address all aspects regarding data quality metadata, AI Act requirements, and federated infrastructure specifications. Continuous monitoring of these evolving standards, combined with a

versioning strategy for the metadata model, will therefore be essential to maintain long-term compliance and interoperability.

6. Conclusion

This work presents a tailored, HealthDCAT-AP-compliant metadata model that addresses real-world gaps in federated health data environments through project-specific extensions covering OMOP-based data quality, data privacy, and study-level metadata. Full compliance with mandatory HealthDCAT-AP requirements ensures interoperability with EU-level catalogues, while domain-specific properties enhance dataset discoverability and support reliable secondary-use analytics. The model is formalised as a JSON Schema, enabling machine-readable, semantically consistent metadata descriptions across heterogeneous data holders. A usability assessment confirmed strong discoverability (avg. 4.5) and adoption likelihood (avg. 4.3); however, the main concern relates to the complexity and effort required to complete the metadata. This calls for strengthened provenance tracking, further training, and the development of semi-automated tools to extract metadata. The proposed methodology and profile provide a reusable framework adaptable to other disease areas and federated health data infrastructures, advancing FAIR principles and the secondary use of health data across Europe.

Abbreviations

ACHILLES
Automated Characterization of Health Information at Large–scale Longitudinal Evidence Systems
AI
Artificial Intelligence
AI4HI
Artificial Intelligence for Health Imaging
CDM
Common Data Model
DATS
Data Tag Suite
DCAT
Data Catalog Vocabulary
DCAT
AP–DCAT Application Profile for Data Portals in Europe
DISH
Data Information Sheet
DQD
Data Quality Dashboard
DQV
Data Quality Vocabulary
DPV

Data Privacy Vocabulary

EHR

Electronic Health Record

EHDS

European Health Data Space

ELIXIR

LU–ELIXIR Luxembourg

EU

European Union

EUCAIM

European Federation for Cancer Images

FAIR

Findable, Accessible, Interoperable, and Reusable

FHIR

Fast Healthcare Interoperability Resources

GA4GH

Global Alliance for Genomics and Health

GDPR

General Data Protection Regulation

GIS

Geographic Information Systems

HMDC

Health Metadata Catalog

HTA

Health Technology Assessment

IHI JU

Innovative Health Initiative Joint Undertaking

IDERHA

Integration of Heterogeneous Data and Evidence towards Regulatory and HTA Acceptance

JSON

LD–JavaScript Object Notation for Linked Data

ML

Machine Learning

OHDSI

Observational Health Data Sciences and Informatics

OMOP

Observational Medical Outcomes Partnership

OWL

Web Ontology Language

RDF

Resource Description Framework
SDMX
Statistical Data and Metadata Exchange
SHACL
Shapes Constraint Language

Declarations

Ethics approval and consent to participate

The Andalusian Ethics Committee is governed by Decree 8/2020 of 30 January, which regulates healthcare ethics and biomedical research ethics in this region. This framework applies to our institution, as it is located in Andalusia. The decree stipulates that clinical research must ensure the protection of participants in accordance with the ethical principles of the Declaration of Helsinki and the relevant Council of Europe Convention. Thus, the study for the generation of synthetic data was developed in accordance with these ethical principles and good research practice. The reference code assigned to the communication/application was: SICEIA-2026-000464. A waiver of informed consent was requested in accordance with the Seventeenth Additional Provision, Processing of Health Data. Section d) of Organic Law 3/2018 of 5 December on the Protection of Personal Data and the Guarantee of Digital Rights.

Consent for publication

Not applicable.

Availability of data and materials

The synthetic dataset used in this work is available upon request. All metadata schemas, sample metadata, and the ontology model developed in this study are provided as additional files and will be publicly available at <link> upon publication.

Competing interests

The authors declare that they have no competing interests.

Funding

This research is supported by the Innovative Health Initiative Joint Undertaking (IHI JU) under grant agreement No 101112135. The JU receives support from the European Union’s Horizon Europe research and innovation programme and COCIR, EFPIA, Europa Bío, MedTech Europe, and Vaccines Europe.

Authors' contributions

VPS, PG and RH conceived the study concept and coordinated the development of the metadata model. JSM, HCK led the development of the HealthDCAT-AP–derived metadata schema. JSM, LVB and BGSC

prepared the initial manuscript draft. SA, FJNB, CLG and JD contributed to the enhancement of the metadata model, considering data quality, interoperability, and conditions of use metadata properties. All authors reviewed the manuscript, contributed to revisions, and approved the final submitted version.

Acknowledgements

We gratefully acknowledge Andreas Kremer, Stephen Sauermann, and Lukas Rohatsch for their valuable contributions during project meetings and discussions. Funded by the European Union, the private members, and those contributing partners of the IHI JU. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the aforementioned parties. Neither of the aforementioned parties can be held responsible for them.

Corresponding Author

Corresponded to Jaleh Shoshtarian Malak, Jaleh.Malak@uni.lu

References

1. Making Metadata Machine-Readable as the First Step to Providing Findable. Accessible, Interoperable, and Reusable Population Health Data: Framework Development and Implementation Study. Online J Public Health Inf. 2024;16. https://doi.org/10.2196/56237.
2. Toward a Domain-Overarching Metadata Schema for Making Health Research Studies FAIR, Findable. Accessible, Interoperable, and Reusable): Development of the NFDI4Health Metadata Schema. JMIR Med Inf. 2025;13. https://doi.org/10.2196/63906.
3. Exploring Metadata Catalogs in Health Care Data Ecosystems. Taxonomy Development Study. JMIR Formative Res. 2025;9. https://doi.org/10.2196/63396.
4. A European strategy. for data | Shaping Europe’s digital future. https://digital-strategy.ec.europa.eu/en/policies/strategy-data. Accessed 5 Aug 2025.
5. European Health Data Space(EHDS). 2022. https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en
6. HealthDCAT-AP. https://healthdcat-ap.github.io/. Accessed 30 May 2025.
7. Data Privacy Vocabulary (DPV). https://w3c.github.io/dpv/2.1/dpv/. Accessed 30 May 2025.
8. Data on the Web Best Practices: Data Quality Vocabulary. 2016. https://www.w3.org/TR/vocab-dqv/. Accessed 7 Aug 2025.
9. Sansone SA, Gonzalez-Beltran A, Rocca-Serra P, Alter G, Grethe JS, Xu H, et al. DATS, the data tag suite to enable discoverability of datasets. Sci Data. 2017. https://doi.org/10.1038/sdata.2017.59.4.
10. Introducing the DATS model. https://fairplus.github.io/the-fair-cookbook/content/recipes/infrastructure/dats-model.html. Accessed 8 Aug 2025.
11. Becker R, Alper P. ELIXIR Luxembourg Data Information Sheet - DISH. 2021.

12. ELIXIR-LU - ELIXIR Luxembourg. https://elixir-luxembourg.org/. Accessed 8 Aug 2025.
13. ohdsi-studies. OhdsiEvidenceNetwork/docs/EvidenceNetworkStudyProtocol_v1.0.pdf at main · ohdsi-studies/OhdsiEvidenceNetwork. GitHub. https://github.com/ohdsi-studies/OhdsiEvidenceNetwork/blob/main/docs/EvidenceNetworkStudyProtocol_v1.0.pdf. Accessed 31 July 2025.
14. Data Standardization – OHDSI. https://www.ohdsi.org/data-standardization/. Accessed 7 Aug 2025.
15. Martí-Bonmatí L, Blanquer I, Tsiknakis M, Tsakou G, Martinez R, Capella-Gutierrez S, et al. Empowering cancer research in Europe: the EUCAIM cancer imaging infrastructure. Insights into Imaging. 2025;16:47. https://doi.org/10.1186/s13244-025-01913-x.
16. El Ghosh M, Kalokyri V, Sambres M, Vaterkowski M, Duclos C, Tannier X, et al. From Syntactic to Semantic Interoperability Using a Hyperontology in the Oncology Domain. Stud Health Technol Inf. 2024;316:1385–9. https://doi.org/10.3233/SHTI240670.
17. IDERHA. (Integration of Heterogeneous Data and Evidence towards Regulatory and HTA Acceptance). 2023.
18. Garcia Santa Cruz B, Vega C, Hertel F. The Need of Standardised Metadata to Encode Causal Relationships: Towards Safer Data-Driven Machine Learning Biological Solutions. In: Chicco D, Facchiano A, Tavazzi E, Longato E, Vettoretti M, Bernasconi A, et al. editors. Computational Intelligence Methods for Bioinformatics and Biostatistics. Cham: Springer International Publishing; 2022. pp. 200–16. https://doi.org/10.1007/978-3-031-20837-9_16.
19. Voisin C, Linden M, Dyke SOM, Bowers SR, Alper P, Barkley MP, et al. GA4GH Passport standard for digital identity and access permissions. Cell Genom. 2021. https://doi.org/10.1016/j.xgen.2021.100030. 1:None.
20. uniluxembourg / LCSB / Bioinformatics Core / IDERHA / IDERHA-metadata · GitLab. GitLab. 2026. https://gitlab.com/uniluxembourg/lcsb/BioCore/iderha/iderha-metadata. Accessed 26 May 2026.
21. Automated Characterization of. Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES). 2015.
22. Execute and View Data Quality Checks on OMOP CDM Database. https://ohdsi.github.io/DataQualityDashboard/. Accessed 26 May 2026.
23. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. EGEMS (Wash DC). 2016;4:1244. https://doi.org/10.13063/2327-9214.1244. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data.
24. The RDF Data Cube Vocabulary. https://www.w3.org/TR/vocab-data-cube/. Accessed 19 Nov 2025.
25. Shapes Constraint Language (SHACL). 2017. https://www.w3.org/TR/shacl/. Accessed 19 Nov 2025.
26. SHACL Validator. https://www.itb.ec.europa.eu/shacl/any/upload. Accessed 25 May 2026.
27. Lario R, Kawamoto K, Sottara D, Eilbeck K, Huff S, Del Fiol G, et al. A method for structuring complex clinical knowledge and its representational formalisms to support composite knowledge

interoperability in healthcare. J Biomed Inform. 2023;137:104251. https://doi.org/10.1016/j.jbi.2022.104251.

28. Huang Y-N, Munteanu V, Love MI, Ronkowski CF, Deshpande D, Wong-Beringer A, et al. Perceptual and technical barriers in sharing and formatting metadata accompanying omics studies. Cell Genomics. 2025;5:100845. https://doi.org/10.1016/j.xgen.2025.100845.
29. Brandt O, Gauza H, Kaltenbach J, Müller ME, Schneider G, Zinn C. A Minimal Metadata Schema and Its Tool to Improve the Searchableness of Research Data in Bioinformatics. J Libr Metadata. 2024;24:165–88. https://doi.org/10.1080/19386389.2024.2338314.
30. Golpayegani D, Pandit HJ, Lewis D. To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In: 2023 ACM Conference on Fairness Accountability and Transparency. Chicago IL USA: ACM; 2023. pp. 905–15. https://doi.org/10.1145/3593013.3594050
31. Laboratory of Medical Informatics and Knowledge Engineering in e-Health (LIMICS). EUCAIM's Hyper-Ontology_V1.3. 2025. https://doi.org/10.5281/ZENODO.10777925
32. OHDSI GIS WG. https://ohdsi.github.io/GIS/. Accessed 10 Aug 2025.
33. Park WY, Jeon K, Schmidt TS, Kondylakis H, Alkasab T, Dewey BE, et al. J Imaging Inf Med. 2024;37:899–908. https://doi.org/10.1007/s10278-024-00982-6. Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension.

Figures

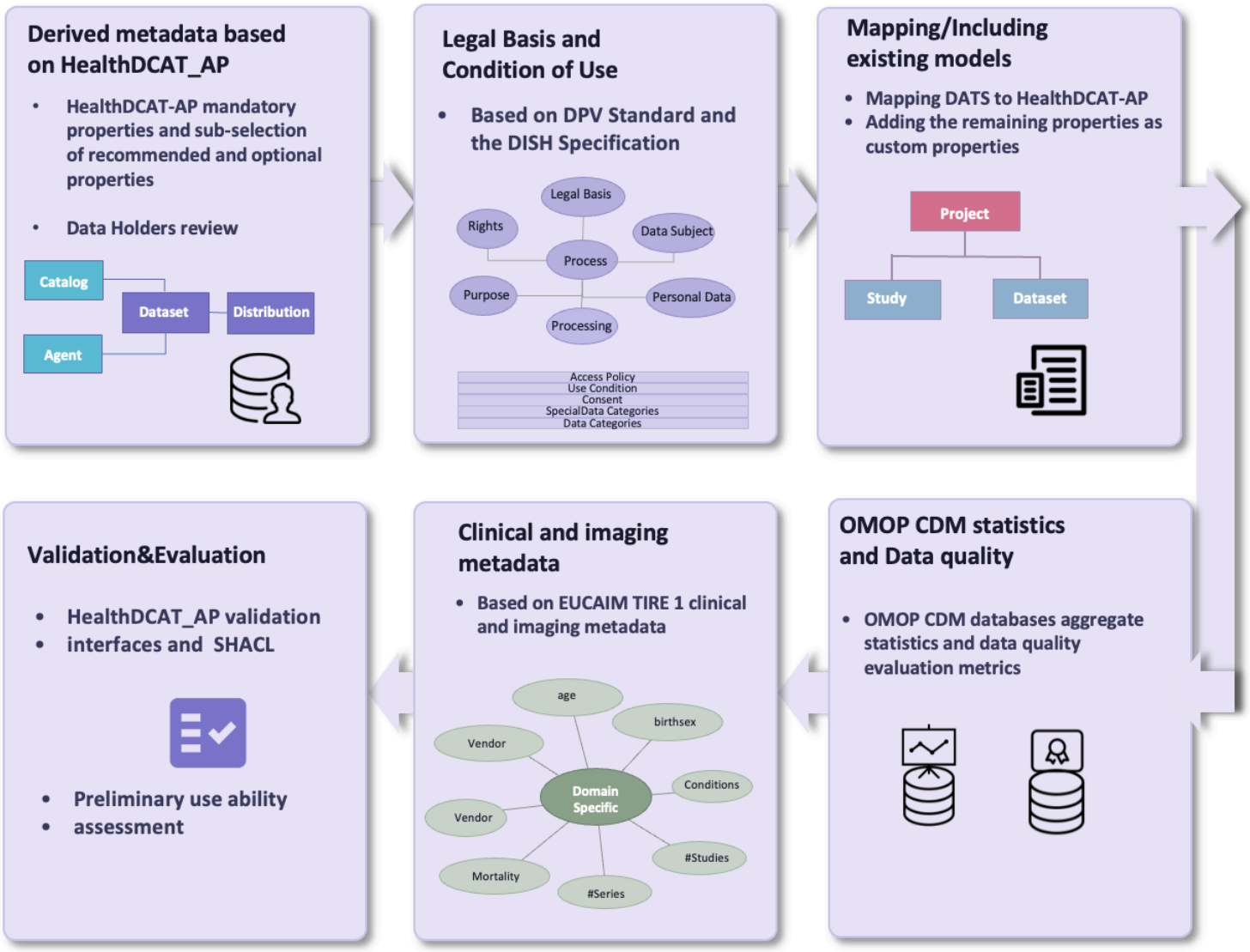


Figure 1

Steps of the IDERHA metadata model development process: building a derived model based on HealthDCAT-AP specifications, defining legal basis and conditions of use requirements, mapping properties from existing models (DATS and DISH), incorporating OMOP CDM statistics and data quality metrics, adding clinical and imaging metadata from EUCAIM TIER 1, and performing validation and usability evaluation.

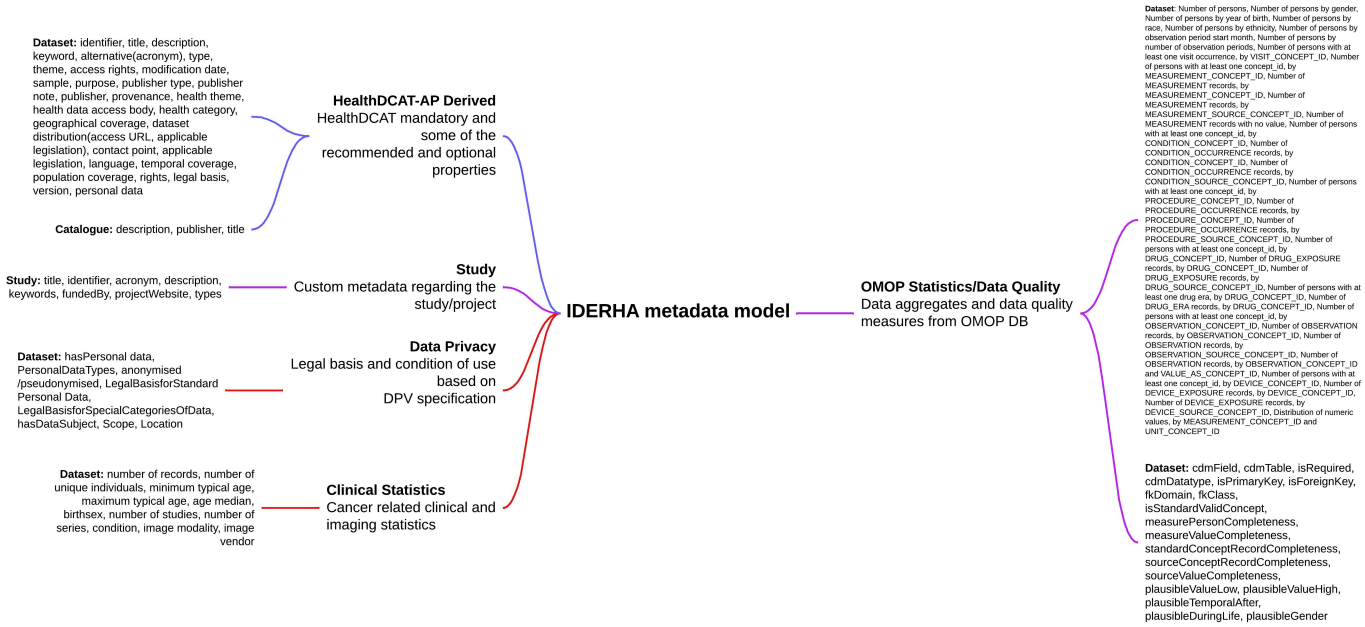


Figure 2

Mind-map overview of the IDERHA metadata model, organized into five categories: HealthDCAT-AP-derived dataset and catalogue descriptors, OMOP CDM aggregate statistics and data quality measures, data privacy and conditions of use based on DPV, cancer-related clinical and imaging statistics, and study-level metadata.



```
"IDERHA:Number_of_persons_by_gender": {
 "qb:measureType": "sdmx-measure:obsValue",
 "qb:structure": "http://IDERHA.org/structure/gender",
 "qb:observation": [
  {
   "sdmx-measure:obsValue": 7450,
   "IDERHA:gender": "male"
  },
  {
   "sdmx-measure:obsValue": 4750,
   "IDERHA:gender": "female"
  },
  {
   "sdmx-measure:obsValue": 145,
   "IDERHA:gender": "unknown"
  }
 ]
},
```

```
"IDERHA:Number_of_persons_by_observation_period_start_month": {
 "qb:measureType": "sdmx-measure:obsValue",
 "qb:structure": "http://IDERHA.org/structure/observation_period_start_month",
 "qb:observation": [
  {
   "sdmx-measure:obsValue": 1500,
   "IDERHA:observation_period_start_month": "2020-01-01"
  },
  {
   "sdmx-measure:obsValue": 1600,
   "IDERHA:observation_period_start_month": "2021-01-01"
  },
  {
   "sdmx-measure:obsValue": 950,
   "IDERHA:observation_period_start_month": "2022-01-01"
  }
 ]
},
```

Figure 3

Sample JSON-LD representation of two OMOP CDM aggregate statistics encoded using DQV and SDMX vocabularies: number of persons by gender (left) and by observation period start month (right).

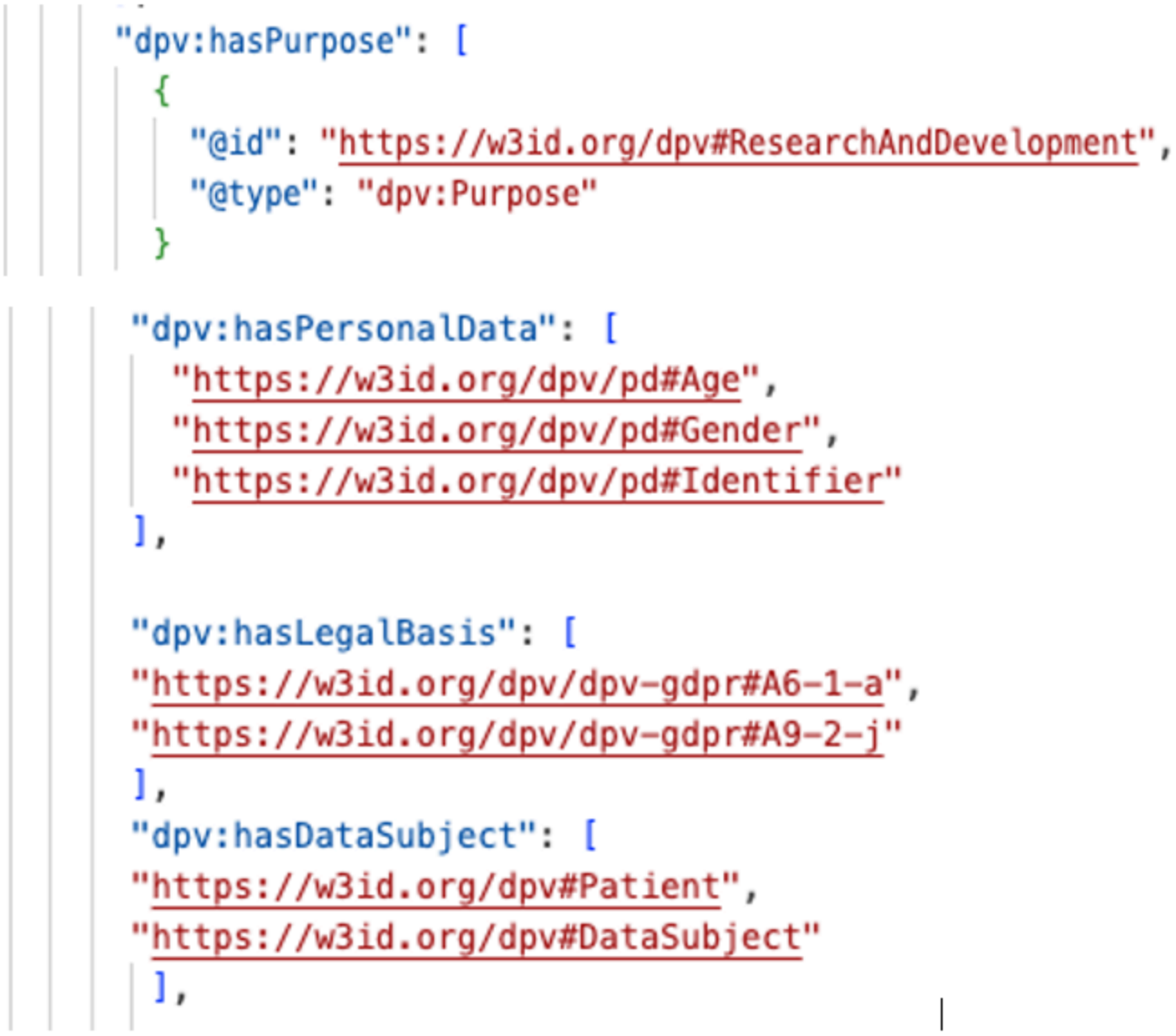


Figure 4

Sample JSON-LD extract illustrating the representation of conditions of use using DPV concepts, including processing purpose, personal data categories, GDPR legal bases, and data subject types.

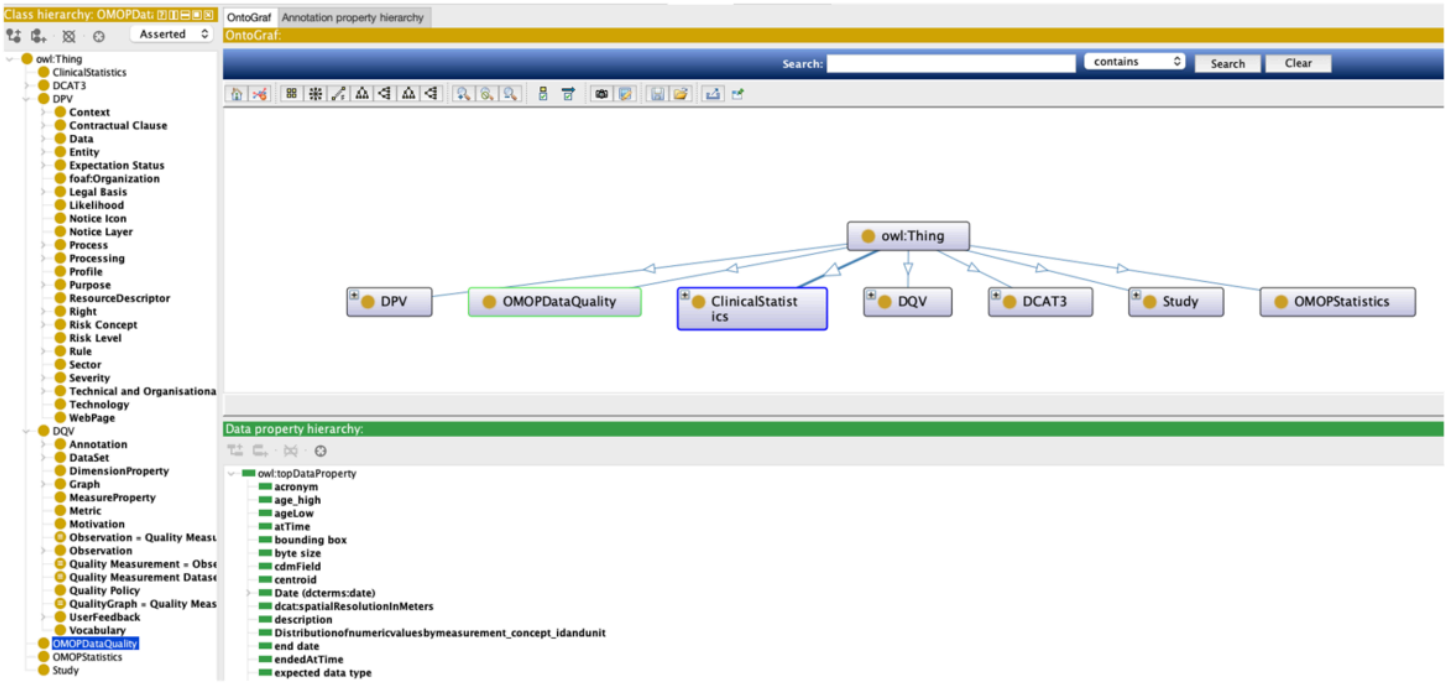


Figure 5

IDERHA metadata ontology class hierarchy visualized in Protégé 5.6 ontoGraf view, illustrating the main classes and their relationships. The ontology integrates concepts from DCAT-AP, DPV, and DQV and is extended with domain-specific classes for OMOP CDM aggregate statistics, data privacy, clinical statistics, and study-level metadata.

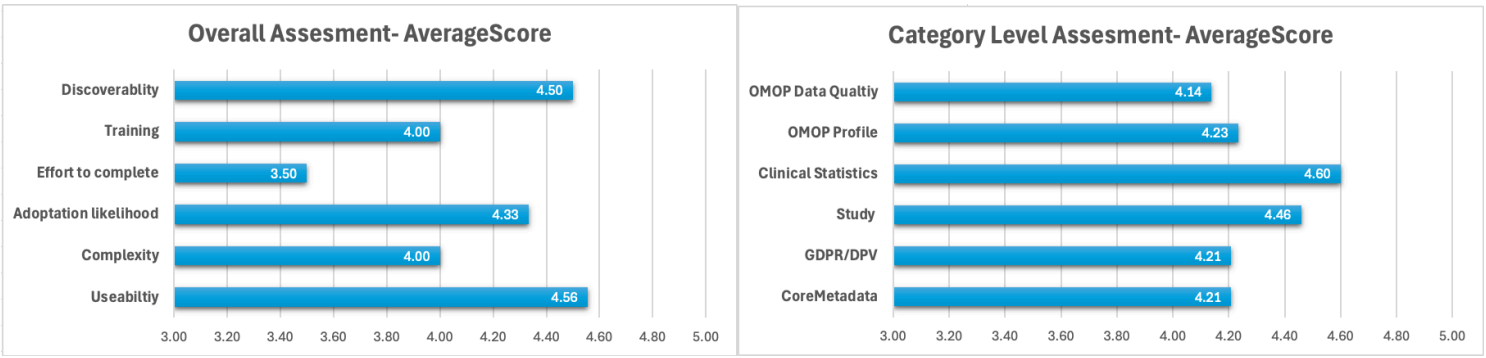


Figure 6

Preliminary usability assessment results. Panel (A) shows average scores (1–5 scale) for overall assessment dimensions, including discoverability, adoption potential, and effort required to complete the metadata. Panel (B) shows average scores per metadata category.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFileslegends.docx
- AdditionalFile1HealthDCATDATSDISH.xlsx
- AdditionalFile2IDERHACustomMetadataV1.03.xlsx
- AdditionalFile3IDERHAMetadataSchema.V2.0.json
- AdditionalFile4IDERHAOntology0.1
- AdditionalFile5SampleSyntheticSAS.jsonld
- AdditionalFile6IDERHAMetadataSampleFull.jsonld
- AdditionalFile7AssesmentAnalysisResults.xlsx