Call: HORIZON-HLTH-2021-TOOL-06
Topic: HORIZON-HLTH-2021-TOOL-06-03
Funding Scheme: HORIZON Research and Innovation Actions (RIA)

Grant Agreement no: 101057062

AIDAVA

AI powered Data Curation & Publishing Virtual Assistant

Deliverable No. 2.1

AIDAVA Reference Ontology: the foundation of a Global Data Sharing Standard

Approval by the European Commission Pending

Contractual Submission Date: 31/05/2023

Actual Submission Date: 01/06/2023

Responsible partner: P5- Sirma AI (Onto)

Funded by
the European Union

Grant agreement no.	101057062
Project full title	AIDAVA - AI powered Data Curation & Publishing Virtual Assistant

Deliverable number	**D2.2**
Deliverable title	**AIDAVA Reference Ontology: the foundation of a Global Data Sharing Standard**
Type[1]	DEM
Dissemination level[2]	PU
Work package number	2
Task leader	P5-ONTO
Author(s)	Todor Primov, Svetla Boytcheva (P5-ONTO), Isabelle de Zegher (P2-b!lo), Remzi Celebi, Louis Powell (P1-UM), Dipak Karlra (P6-IHD), Markus Kreuzthaler (P7-MUG)
Keywords	Ontology, interoperability terminology, information model, SNOMED, LOINC, IPS,

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA).

Neither the European Union nor the granting authority can be held responsible for them.

Document History

Version	Date	Description
V0.1	01/02/2023	Deliverable structure and initial draft
V1.0	01/06/2023	Final version of the deliverable document

List of Definitions & Abbreviations

The definitions used in the deliverable are based on the AIDAVA Glossary [ref].

Table of Contents

[1] **Type**: Use one of the following codes (in consistence with the Description of the Action):

R:	Document, report (excluding the periodic and final reports)
DEM:	Demonstrator, pilot, prototype, plan designs
DEC:	Websites, patents filing, press & media actions, videos, etc.

[2] **Dissemination level**: Use one of the following codes (in consistence with the Description of the Action)

PU:	Public, fully open, e.g. web
SEN:	Sensitive, limited under conditions of the Grant Agreement

Executive Summary

Ontologies are increasingly used to support harmonisation of **population** data from heterogeneous data sources in support of clinical research, with a **specific** research question requiring a well defined dataset. AIDAVA is exploring the possibility of using an ontology to harmonise all **patient** data, extracted from heterogeneous data sources, into an individual personal health knowledge graph (PHKG) that can then be reused for **multiple** purposes, in clinical care and clinical research.

The decision to take an ontology approach in AIDAVA, rather than to follow a structural standard such as an information model, was made already at proposal time as ontologies are semantic rich and agnostic of structural and syntactical formats, increasing potentially of interoperability and reuse in compliance to the FAIR principles. Moreover, new knowledge can be added smoothly by extending the ontology concepts with RDF triples and data quality constraints through SHACL rules.

Development of the AIDAVA Reference Ontology followed a structured approach including ideation, requirement analysis, design and development. The requirements took into account the use cases developed in WP1, the requirements extracted from the automation phases described in Task 2.1 and the annotation process described in Task 4.3. The data quality constraints were built in alignment with Task 4.2. We identified 4 Ontology Strategic Requirements and 6 Ontology Requirement Specifications that provided directions for the design and the developement of the ontology.

A critical aspect of an ontology like the AIDAVA Reference Ontology to comply with FAIR principles as effectively as possible is to maximise alignment with emerging and existing standards. While reviewing the work on semantic interoperability of related initiatives, including TEHDAS and the European Electronic Health Record exchange format (EEHRxf), we came to the conclusion that SNOMED CT and LOINC were priority standards to be included. However they need to be completed by other standards to cover additional relationships and other domains. Several candidates were considered and it was decided to include the semantics subsumed in the HL7 FHIR General Purpose Data Types, and relevant HL7 FHIR profiles through the governance process, as second priority. We expect that other semantic standards will be required to achieve the long term objective of the AIDAVA Reference Ontology to cover a majority of medical concepts contained in personal health medical records.

This deliverable also describes the technical specification of the AIDAVA Reference Ontology, which defines the structure, components, and relationships within the scope of the two targeted use cases (Breast cancer registry and Cardiovascular score) and in a broader context (ensuring semantic interoperability across systems). It includes a formal representation of the concepts, entities and their attributes, which are specified in the AIDAVA Dataset.

While developing the ontology, we realised that additional concepts and relationships as well data quality constraints will need to be added when data sources to be curated will be onboarded across sites, and when more narrative texts will be annotated. This requires a governance process to be executed during the project lifetime, as described in Section 3.4. In addition, and assuming the project will be successful, governance will also be needed beyond the project to maximise sustainability and reuse of the results. While is not in scope of this deliverable, the proposed approach is introduced here; it will be discussed extensively during the planned meetings with the Sustainability Advisory Board.

1 Introduction

This deliverable meets a key objective - and second methodological pillar - of AIDAVA : deliver a **universal semantic representation** of an **interoperable and reusable patient longitudinal health record** that is curated from multiple heterogeneous data sources and that can be reused for multiple purpose, by the patients and their treating physicians in clinical care, or shared - with the consent of the patient - for clinical research.

Why an ontology as universal semantic representation : structure versus semantic

We differentiate data standards into two main categories: structural and semantic data standards. **Structural data standards** focus on the syntax and structure of the data such as format of the data elements, their relationships, and their encoding. Part of the semantic behind the structure is implicit, but known in the specific domain in which they are defined and used. Structural data standards enable data exchange and interoperability between systems in a specific context and domain of use. Multiple structural data standards exist in healthcare to cover different needs across clinical care and clinical research. For example, HL7 FHIR is emerging as the European Electronic Health Record exchange format standard across clinical institutions, CDISC SDTM is required as format for submitting clinical trial data in support of market authorization application to regulatory authorities, and OMOP is used around the globe to fuse real-world clinical data for distributed analytics in research. As a result, health data of a single patient collected and stored in different data sources in different formats are not interoperable at the patient level.

Semantic data standards focus on the meaning and context of the data through terminology and ontology that define the concepts, and relationships describing the data. More specifically an ontology can formalise complex and dynamic relationships between real world objects with contextual information; all the semantics are explicit. Standards such as LOINC [1] (Logical Observation Identifiers Names and Codes) and SNOMED CT [2] (Systematized Nomenclature of Medicine Clinical Terms) are terminologies, though only SNOMED CT can be considered as a true ontology. Semantic data standards and ontologies enable systems to process and interpret the data in a consistent, unambiguous manner and have intrinsic features that align with the FAIR principles.

AIDAVA aims to demonstrate that to support generation of an interoperable patient longitudinal health record (extracted from heterogeneous data sources) that could be reusable in different target formats. Therefore, we need a semantic rich, structure-agnostic data representation with a clear and unambiguous definition of concepts, rather than a pre-defined structural data standard. This semantic rich representation must support transformation from and to multiple (structural) data standards, without being constrained by a specific structure and without loss of semantics. This is what we refer to as the "***universal interoperable format for data sharing***" displayed in the middle of Figure 1; we assert that it can be implemented through an ontological approach, the AIDAVA Reference Ontology (RO).

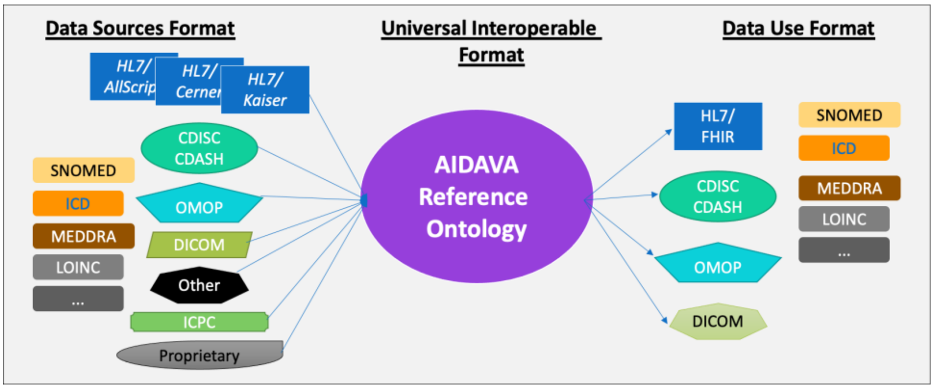


Figure 1. AIDAVA Reference Ontology, a universal format in healthcare

The decision to take an ontology approach in AIDAVA, rather than to follow a structural standard such as an information model, was made already at proposal time for the following reasons.

- While very useful - and performant - for a well defined context or use case, structural data standards and information models are good for specific, well defined data exchanges but are not suitable as a "universal standard for reuse" where the semantic must be precise but independent of the context, as indeed the context may vary widely based on data usage requirements.
- Ontologies support the implementation of the FAIR principles, particularly in the areas of Interoperability and Reuse, which are key objectives of the AIDAVA project. By providing a common, explicit description of the concepts, ontologies limit ambiguity and support semantic interoperability between different systems, while also facilitating the discovery, selection, and reuse of data across a range of use cases.
- Extension of domain concepts can easily be done into an ontology by adding RDF triples; adding new concepts in a schema may involve in depth changes into the different entities (and to related application logic). Extensibility is important to maintain a semantic representation of the longitudinal health record where medical knowledge and its formal representation to meet multiple use cases keep evolving.
- Knowledge graphs, used as machine readable format of ontology, support flexible and easily extensible methods to measure and improve data quality. For example, Shapes Constraint Language (SHACL) can verify the syntactic and semantic accuracy of data; link prediction methods can complete missing links; fact validation methods can detect and clean up any erroneous links.

During the AIDAVA project, the Reference Ontology will be piloted in two main use cases – Breast cancer registry and Cardiovascular Diseases score – and focus on these two therapeutic areas; nevertheless, the ontology will be designed to allow extensibility towards other therapeutic areas.

The Reference Ontology enables an interoperable Personal Health Knowledge Graph

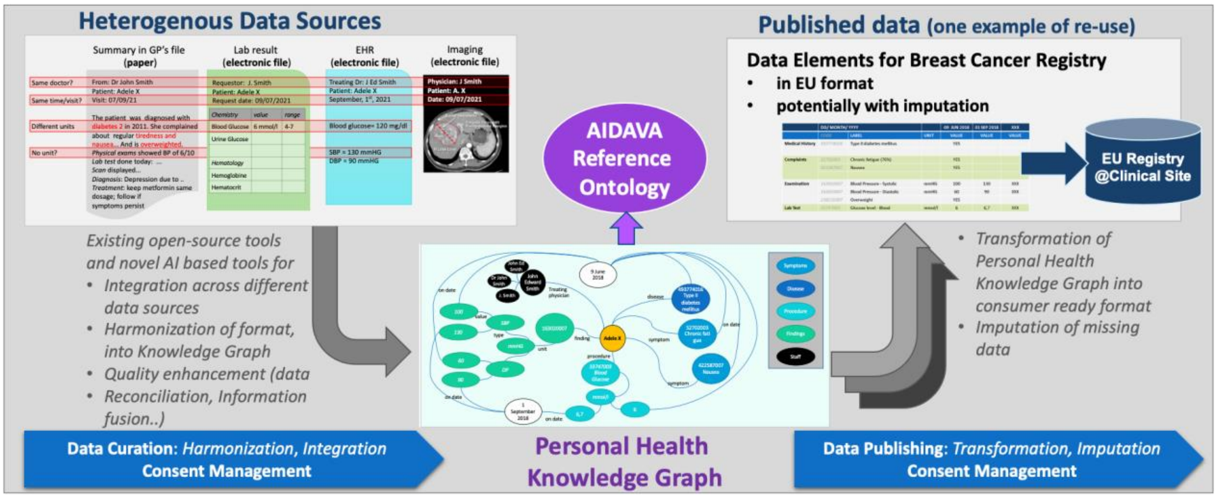

Heterogenous Data Sources
Summary in GP's file (paper)
Lab result (electronic file)
EHR (electronic file)
Imaging (electronic file)
Same doctor?
Same time/visit?
Different units
No unit?
From: Dr John Smith
Patient: Adele X
Visit: 07/09/21
The patient was diagnosed with diabetes 2 in 2011. She complained about regular tiredness and nausea... And is overweighted.
Physical exams showed BP of 6/10
Lab test done today: ...
Scan displayed...
Diagnosis: Depression due to ..
Treatment: keep metformin same dosage; follow if symptoms persist
Requestor: J. Smith
Patient: Adele X
Request date: 09/07/2021
Chemistry
value
range
Blood Glucose
6 mmol/l
4-7
Urine Glucose
Hematology
Hemoglobine
Hematocrit
Treating Dr: J Ed Smith
Patient: Adele X
September, 1st, 2021
Blood glucose= 120 mg/dl
SBP = 130 mmHG
DBP = 90 mmHG
Physician: J Smith
Patient: A. X
Date: 09/07/2021
Existing open-source tools and novel AI based tools for
• Integration across different data sources
• Harmonization of format, into Knowledge Graph
• Quality enhancement (data
• Reconciliation, Information fusion..)
AIDAVA Reference Ontology
John Ed Smith
Dr John Smith
John Edward Smith
J. Smith
Treating physician
9 June 2018
Adele X
on date
value
type
unit
finding
symptom
disease
procedure
1 September 2018
Symptoms
Disease
Procedure
Findings
Staff
Published data (one example of re-use)
Data Elements for Breast Cancer Registry
• in EU format
• potentially with imputation
Medical History
Type II diabetes mellitus
Complaints
Chronic fatigue
Nausea
Examination
Blood Pressure - Systolic
Blood Pressure - Diastolic
Overweight
Lab Test
Glucose level - Blood
EU Registry @Clinical Site
• Transformation of Personal Health Knowledge Graph into consumer ready format
• Imputation of missing data
Data Curation: Harmonization, Integration
Consent Management
Personal Health Knowledge Graph
Data Publishing: Transformation, Imputation
Consent Management

Figure 2. Each PHKG is constrained by the AIDAVA Reference Ontology

The AIDAVA Reference Ontology represents concepts that can be included in any personal health record; it does not include patient data. In AIDAVA, all personal medical information of a single individual will be included in a multimodal Personal Health Knowledge Graph (PHKG) as displayed in Figure 2, enriched and curated gradually as new data sources are added. These PHKG must be compliant with the Reference Ontology to ensure interoperability. Such a multimodal PHKG is ideally positioned to capture the semantics of data from heterogeneous sources, independently of their structure; it can also support data integration, data quality enrichment and correction, based on the context. In addition, a PHKG can also support publishing of data in many different target formats based on the needs.

Description of the content of this document

This document is focusing on the development and maintenance of the AIDAVA Reference Ontology. Section 2 explains the different activities that took place to implement a first release of this Ontology. Section 3 describes in more details the results, including a clarification of the requirements for the ontology, the identification and selection of existing standards to support delivery of the ontology, a description of the ontology itself - managed within an open source repository - and finally a description of the governance process to maintain the ontology during the project and beyond the project.

2 Description of Activities

2.1

2.2 Design and implementation of the Reference ontology

The ontology design has been carried out iteratively with a continuous discussion with the AIDAVA partners involved in the definition of relevant tasks.

- **Task 1.1** *(Detailed description of use cases)* clarified the domain on which the first release of the Reference Ontology (RO) should focus as a priority.
- **Task 2.1** *(Define structured process for automation of data quality enhancement and FAIRification)* clarified how to use the RO as part of the data source onboarding, where the data description or schema of the data source is mapped with concepts of the ontology. This process enables to automatically align patient data - extracted from the onboarded data source and to be stored in the PHKG - with the relevant concept of the RO and ensure interoperability.
- **Task 4.3** *(Manual Annotation of text documents in 3 languages)* provided guidance on how to use the ontology when annotating unstructured medical information in order to ensure that the output resulting from the annotation process contains only concepts defined in the reference ontology. Indeed this output will serve to train the NLP tool developed as part of Task 5.1 that will in turn, extract concepts from patient data available in narrative form and insert it into the PHKG. The training data set must hence be compliant with the Reference Ontology to ensure interoperability and reuse of the extract from the NLP tool[3].
- **Task 4.2** *(Data quality, metadata and open data)* focused on quality metrics to be added to the Reference Ontology .

We are developing the AIDAVA Reference Ontology through five main phases that characterise the development of an ontology [3].

1. ***Ideation***. In this phase, we defined the ontology's goals and the strategic direction that it must follow in terms of alignment with industry standards and established terminologies. We considered the standard practices in ontology development and cross-checked them with the requirements of the project. Much of this activity happened already during the proposal phase; it was reconfirmed when starting this task as part of the requirement analysis.
2. ***Requirements analysis*.** In this phase, we defined the minimum set of requirements that the ontology has to fulfil in terms of extensibility, interoperability, support of the annotation process and data onboarding as well as the quality metrics. We analysed the two use-cases in-depth and ensured coverage for all required classes and their properties. We individuated the main areas of interest in the medical records and the main differences and commonalities amongst the use-cases. The requirements are provided in Section 3.1; additional requirements were as well elicited as part of the design phase and included in Section 3.2.
3. ***Design*.** In this phase, we confirmed the standards to be included in the ontology, as indeed a core objective of AIDAVA is to reuse and integrate existing standards. We consolidated the requirements of AIDAVA with a review of other EU initiatives with the same strategic goals, identified a list of suitable candidates and confirmed the standards to be used in the AIDAVA Reference Ontology as a priority. This is described in Section 3.2.
4. ***Development*.** In this phase, we analysed various medical records examples provided by our partners as part of the definition of the manual annotation guidelines produced in Task 4.3. Similar approach will be applied over the structured data once samples are made available. We proceeded in a bottom-up fashion by individuating the records main concepts and

[3] When using NLP tools that have not been trained by a data set with labels compliant with the AIDAVA Reference Ontology, the output of these tools must be considered as structured data. This structured output must then undergo alignment/mapping with the Reference Ontology.

verifying that selected target medical ontologies (SNOMED CT, LOINC,..) cover them. We decided to connect all the identified classes with synonyms and similar classes in other authoritative ontologies using OWL [4] axioms *equivalentClass* and *equivalentProperty* mappings. This choice will be useful for improving semantic interoperability (entity linking, semantic mappings), as well as to enable cross-lingual features. We decided to employ OWL2 [5] to design and develop the ontology, as it will allow us to define object centric, terminology-based ontology to be developed incrementally based on open-world assumptions. We imported existing classes from external ontology and created customised classes and relationships when necessary. This development model guarantees flexibility and extensibility to meet the project's possible future requirements and potential new use-cases. Development of the ontology is described in Section 3.3.

5. ***Expert interaction*.** In this phase, we presented the ontology draft to the consortium and gathered feedback both from the medical and technical viewpoint. This confirmed the need for adaptation - and mainly extension of concepts and quality rules - of the ontology during the project life cycle. We implemented the required updates but also defined a governance process to be executed during the project. This is described in Section 4.

2.3 Need for Governance

As mentioned above the Reference Ontology will require updates during the AIDAVA project. In addition, as part of the sustainability of the AIDAVA project we have to consider how to expand the ontology to cover other therapeutic areas and other use cases, and also to ensure PHKGs can be widely reused - with patient consent - as part of the European Health Data Space (EHDS) initiative and the Electronic Health Record exchange format (EEHRxf).

3 Results

3.1 Requirements for the Reference Ontology

Requirements for the ontology include strategic ones, part of the ideation phase that was initiated during the proposal stage and confirmed at the start of the tasks, and operational requirements that were gathered through alignment with the different tasks.

3.1.1 Ontology Strategic Requirements

Ontologies can serve multiple purposes. The main purpose of the AIDAVA Reference Ontology is to represent all concepts needed in an interoperable, reusable personal longitudinal health record - in the form of a Personal Health Knowledge Graph - to support its sharing and reuse in accordance with FAIR principles.

The strategic requirements of the AIDAVA Reference Ontology are directly linked to the strategic goals of the project i.e. deliver an **interoperable**, **reusable** PHKG - in compliance with the FAIR principles - in a **cost-effective way, sustainable** beyond the project, to bring a structural solution to the so far unsolved interoperability issue in healthcare, and to support implementation of critical initiatives such as the European Health Data Space.

Out of these strategic goals we derived the following requirements for the RO.

Ontology Strategic Requirement 1 (OSR1). Must support the core objective of AIDAVA to curate once, use many times. This can be decomposed in

- Support ***data interoperabilit***y of the PHKGs beyond the project and
- Maximise ***potential for reuse*** of the PHKGs across a large range of use cases, beyond the ones identified in the project.

This requires that the AIDAVA Reference Ontology can be expanded with any concepts and relationships available in medical records.

As introduced in Section 3.2 and described in more detail in Deliverable 2.2 (*Details on data curation and publishing process)*, interoperability includes multiple layers: in the context of AIDAVA, we need to focus on concrete and solvable data interoperability issues and on 2 specific use cases. However, in the long term, and assume that AIDAVA is successful, we will need to expand the AIDAVA Reference Ontology.

From a technical point of view, reuse can be achieved by transforming knowledge graph format in another - typically structured - format[4]; this transformation however may require some kind of mapping or transformation to convert the graph-based data into a tabular format. Existing tools, such as SPARQL, enable to execute this transformation in a reliable and efficient manner. Thanks to the explicit representation of semantic in the ontology, knowledge graphs support transformation in multiple different formats - without lost of semantics - to meet a wide range use cases

[4] While Knowledge Graph Data stores are increasingly getting traction and query tools are becoming more performant, most use cases still require data in tabular format - supported by structural data standards - to support analytics.

Ontology Strategic Requirement 2 (OSR2). Must support FAIR principles as well as smooth alignment with standards in place, and with emerging ones, to minimise the need and the cost of alignment/mapping from and to these different standards, while maximising reusability during and after the project.

How we intend to meet this requirement is further described in Section 3.2.2.

Ontology Strategic Requirement 3 (OSR3). Must support maintainability and extensibility during the project as well as beyond the project.

While SR1 will require governance beyond the project, it is also important to take into account potential additions during the project when we onboard data sources and annotate narratives. This is further expanded in Section 3.4.

Ontology Strategic Requirement 4 (OSR4). Must support smooth implementation, maintenance and update of data quality constraints supporting correction as well as monitoring of data quality.

This is further expanded in Section 3.3.3.

3.1.2 Implementation and Ontology Requirement Specifications

To demonstrate the approach, the project focuses on a subset of medical records based on the 2 use cases in scope of the project.

The implementation of the first release of the AIDAVA RO is aligned with the requirements coming from the different tasks of the project.

Alignment with Task 1.1 (Detailed description of use cases): domain of focus

The AIDAVA RO must - in time - represent all concepts that are relevant in a medical record. For the project, this is limited to the concepts needed to 2 therapeutic areas with specific requirements.

- Extraction of a set of data elements needed to maintain a federated Breast Cancer registry (BCR). A list of 80+ data elements has been defined by the breast cancer experts. Each hospital will extract and publish these data elements in the required format from the PHKG of the different patients; a federated query will support analytics across the different hospitals demonstrating interoperability and reuse of the data extracted from the PHKG.
- Extraction of a set of data elements needed to monitor the patient and to compute the SMART cardiovascular (CVD) score of a patient. A list of 50+ data elements has been defined by the CVD clinicians. Each hospital will extract and publish these data elements in the required format from the PHKG of a patient to compute the score of that specific patient, and will provide the score to the local treating physician. The same SMART score application will be used across the hospital demonstrating interoperability and reuse of the data extracted from the PHKGs across organisations.
- Extraction of data from the PHKG and transformation into HL7 FHIR IPS which will then be transferred to the relevant Health Data Intermediary that will provide the IPS to the patient for further visualisation and use. By extracting HL7 FHIR IPS from the PHKG with the same transformation programs across all hospitals, we demonstrate interoperability and reuse of the PHKG.

Ontology Requirement Specification 1. To support the AIDAVA use cases, the Reference Ontology must include the concepts related to

- the data elements of the Breast Cancer registry,
- the data elements of the CVD score as well as
- the data elements - and relationship - to generate a patient IPS.

Ontology Requirement Specification 2. The AIDAVA Reference Ontology must also include predefined **mapping** supporting transformation to HL7 FHIR IPS which is an emerging standard for data exchange supported by the European Electronic Health Record exchange format (EEHRxf).

Alignment with Task 2.1 (Automation of data quality enhancement and FAIRification)

Task 2.1 objective is to maximise automation in data curation by describing a workflow to solve each identified data interoperability issue. To enable this, the first set of workflows must support what we refer to as "onboarding of the data sources" i.e. mapping of the data description/ schema of the data sources in scope, with the concept of the ontology.

The following data source documents have been identified. A detailed description of their content is being developed as part of the data transfer agreement; one data source could include several documents mentioned below.

Hospital System	Health data intermediary
● Discharge Summary/Discharge letter ● Medical history ● Progress notes/ICU notes ● Prescribed medications ● Medical imaging reports ● Pathology reports ● Surgical procedure descriptions ● Multidisciplinary meeting reports ● TNM staging ● Patient referral document ● Laboratory reports ● Echocardiography report ● Coronary angiography report ● Ambulance record ● Emergency department record ● ICU/CICU progress notes	● GP record ● Personal App (breast cancer) ● Personal App (CVD) ● Connected medical device

There are two types of data in the data sources: structured and narrative data elements. Metadata extracted from imaging are considered as structured data elements.

- For structured data elements. The concept specified in the data source description/schema must be mapped with one or more concepts defined in the RO.
- For narrative data elements. There are 2 possibilities:
    - As part of Generation 1 of the AIDAVA prototype, we will work with existing tools that have not been trained within AIDAVA; they will extract concepts based on a specific

structured standard; this target structured standards must be mapped to the AIDAVA RO, in the same way than for structured data elements

- ○ As part of Generation 2 of the AIDAVA prototype, the NLP tool will be trained with concepts available in the RO used by annotation tool (see below alignment with Task 4.3); the output can then be directly introduced in the PHKG.

Ontology Requirement Specification 3. The AIDAVA Reference Ontology must include concepts that support mapping - and transformation - with all entities and relationships included in the data schema of the data sources identified in support of the use cases. If concepts are not available - during the data sources onboarding process - they must be added to the Reference Ontology following a strict governance process.

Alignment with Task 4.3 (Manual Annotation of text documents in 3 languages)

The AIDAVA annotation process is expected to extract concepts from narrative in 3 languages (Estonian, Dutch and German); the process is supported by an annotation guideline used across the different teams. The output resulting from the annotation process is serving as a training data set for the NLP tools developed for Generation 2 of the AIDAVA prototype. To ensure that the concepts extracted from the patient record by this NLP tool are relevant for AIDAVA, we must ensure that the concepts used in the text annotation process are defined in the AIDAVA Reference Ontology. This includes coverage for data elements and their attributes in scope of the use cases, as well as representation of complex cases (as for negations, temporality, etc)

Ontology Requirement Specification 4. The AIDAVA Reference Ontology must include concepts that can be expected to be extracted from narratives existing in the data sources document identified above, to ensure that the concepts that are annotated - and will be extracted by the NLP tools from the patient data source - are available in the ontology. This includes data elements identified in the use cases as well as representation of negations, temporality, causality.
If concepts are not available, they must be added to the Reference Ontology following a strict governance process.

Alignment with Task 4.2 (Data quality, metadata and open data)

An ontology defines concepts and relationships. To ensure data quality, it is needed to add different types of data quality checks such as

- schema of the data (e.g. a concept which represent the result of a lab test must include not only the value but also the unit and the reference range, and if possible additional information on the devices used to perform the test),
- semantic consistency of the data (e.g. a birthdate of an alive person should not indicate an age above 130 years; or a height should not be above 2m80),
- scientific/domain soundness of the data (e.g. a biological female should not have diagnostics or procedures related to prostate).

Ontology Requirement Specification 5. The AIDAVA Reference Ontology must be linked with data quality checks typically implemented through SHACL rules. These data quality checks should be governed (during and beyond the project) to support maintenance and extension.

3.2 Candidate standards as support for the AIDAVA Reference Ontology

Per Ontology Strategic Requirement 2 the AIDAVA Reference Ontology must align with emerging/existing standards. In addition, the Reference Ontology must include concepts aligned with the 2 therapeutic areas - including the ones defined in the list of data elements - as well as the International Patient Summary (IPS) standard.

These requirements provide clear guidelines to identify the semantic standards to be considered as candidates for developing the AIDAVA Reference Ontology.

In this section we first review the approach and recommendations from other relevant projects and identify a list of candidate standards that should be either included or aligned with the AIDAVA Reference Ontology. We finally conclude on the priority standards to be used for the ontology.

3.2.1 Overview of selected Initiatives

In this section we review selected EU initiatives of relevance to understand their suggested approach to support semantic interoperability.

The **TEHDAS project** [6] developed joint European principles for the secondary use of health data in support of the implementation of the European Health Data Space. In Deliverable 6.2 *"Recommendations to enhance interoperability within HealthData@EU*" [7], the TEHDAS team used the Common Assessment Method for Standards and Specifications (CAMSS) [8], which is the European guide for assessing and selecting standards and specifications for eGovernment projects. CAMSS evaluates 16 criteria including openness, transparency, reusability, technical neutrality and portability, generic use and how they address users' needs and expectations, potential for cooperation and the principles of interoperability defined in the European Interoperability Framework (EIF).

TEHDAS evaluated standards facilitating programmatic discoverability and findability, standards facilitating interoperable communication and standards facilitating semantic interoperability. In this later category they assessed CDISC SDTM, LOINC, OMOP-CDM, Orphanet standards, SNOMED CT and concluded the following:

- *"All the standards analysed got a similar interoperability overall score (around 80%).*
- *The adoption of these standards is quite uneven. While in the case of SNOMED CT and LOINC there is wide experience across Europe, the case of Orphanet standards is more limited, and the case of OMOP-CDM is essentially linked to research projects on specific domains, CDISC SDTM is not used in the countries that provided insight.*
- *Importantly, SNOMED CT is being mapped to Orphanet standards, OMOP-CDM, CDISC SDTM, and LOINC has joined SNOMED CT. In addition, all the ICD and ICD-O is mapped to SNOMED CT, as well as, is being used in the Human Phenotype Ontology and in the GLobal Alliance for Genomics and Health. Finally, SNOMED CT is ISO-IDMP compliant allowing the extension of EMA case safety reports."*

In the conclusion of the report, TEHDAS is coming with the following recommendations:

- *RECOMMENDATION 5: Although none of the above standards can cover all the data types of interest in the HealthData@EU. As per CAMMS assessment SNOMED CT has been shown to be the best equipped ontology to cover semantic interoperability across controlled vocabularies and taxonomies referred to medical concepts.*
- *RECOMMENDATION 6: As the current mapping of medical concepts from taxonomies and controlled vocabularies to SNOMED CT is not fully completed, we recommend the European Commission fostering this effort and the Member States to progressively deploy SNOMED CT as an ontology of reference for medical concepts.*

- *RECOMMENDATION 7: There will still be a need for development and sharing of semantic maps to other than medical concepts as concepts from other determinants of health (i.e., social, cultural and economic determinants, environmental determinants, genetic determinants).*

The annex of the recommendation for an **European Electronic Health Record exchange format (EEHRxf)** [9] mandates the use of HL7 CDA format but is already mentioning the potential to evolve toward HL7 FHIR to support data exchange of health data in the priority categories (Patient Summaries, ePrescriptions/eDispensations, Laboratory reports, Medical images and reports, Hospital discharge reports) except for Medical images where DICOM is recommended.

The **Swiss Personalized Health Network (SPHN**) [10] is a national initiative under the leadership of the Swiss Academy of Medical Sciences (SAMS) contributing to the development, implementation and validation of coordinated data infrastructures in order to make health-relevant data interoperable and shareable for research in Switzerland. More specifically SPHN developed an Semantic Interoperability Framework based on a strong semantic layer of information, and graph technologies for the exchange layer, which can be extended by the individual projects to fit their purposes. Thus, a universal exchange language for healthcare is established, using the "words'' from various international standard vocabularies (such as SNOMED CT or LOINC), a simple "grammar" (subject-predicate-object; expressed in RDF), and additional SPHN guidelines and rules to establish good practices for FAIR data. To make the SPHN concepts comparable nationally and internationally, they are expressed by using existing semantic standards (controlled vocabularies), and then aligned/mapped wherever possible to SNOMED CT and/or LOINC. Consequently the data element of a concept can be expressed using one or several recommended standards (e.g. LOINC, SNOMED-CT, ICD-10, ICD-O-3, CHOP, ATC). For example, the instance of substance code under the concept “allergy” can be an ATC code, a SNOMED CT code, or a code from another semantic standard.

The Elixir's **FAIRCookBook** proposes a methodology for selecting terminologies and ontologies [11]; the main purpose is to provide guidance on how to select the most suitable semantic artefacts given a specific research context in general, and when it comes to life and biomedical sciences projects, their main themes, i.e. risk assessment, clinical trial, drug discovery or fundamental research. Their first statement in the methodology is "*Context is everything*" i.e. the domain of operation will generally dictate the semantic framework that is most suited to a given dataset. The FAIRCookBook also proposes a set of terminologies and/or classification to be used based on the context and a set of inclusion and exclusion criteria. While their approach is solid, it is mainly applicable in the context of clinical research and specifically around delivering reusable Real World Data.

The **EMA Big Data Task force** data standardisation strategy [12] recommends the use of several pharma related standards such as CDISC SDTM for transfer of clinical data, ICSR for transfer of side effects and IDMP for medical product identification. IDMP implementation is using SNOMED CT and HL7-FHIR.

3.2.2 List of candidate standards

Based on the TEHDAS recommendations, we can infer that SNOMED CT and LOINC should be priority standards, but they are not sufficient as they do not cover all domains (e.g. rare disease, genomics, specialised cancers) and do not include a rich enough structure like in HL7 FHIR disease profiles. There are many other standards in healthcare that should be considered as second priority. While checking on other candidates, it is important to make the difference between different types of standards of relevance for the AIDAVA Reference Ontology. The figure provided below displays the relationships between the different types of standards mentioned in this section and shows that boundaries are not clear. For instance, the HL7 FHIR data exchange standard is based on the HL7 FHIR information model which includes elements of semantic (and ontology) in the definition of the structure and which uses terminologies like SNOMED CT and LOINC as value sets.

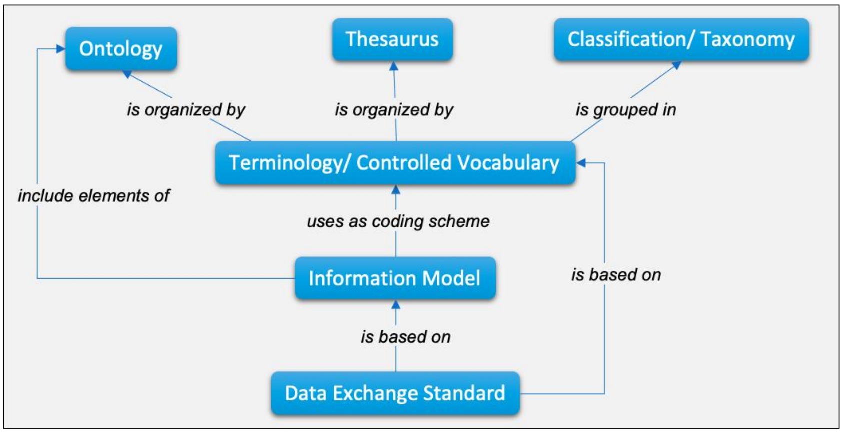


Figure 3. Different types of standards and their relationships

As semantics can be included - more or less explicitly - in different standards, it is relevant to look at all of them and understand how they could contribute to the AIDAVA Reference Ontology. The list provided below is a first attempt to identify the most commonly used standards in Europe; it is not expected to be complete and mostly it is expected that standards will evolve over time and new ones may emerge for specific domains. Identification of appropriate standards - including potentially useful standards from other countries like the US - to cover new needs, will be part of the governance process.

Ontology

Definition. In computer science and information science, an ontology is a formal and explicit specification of a shared conceptualization of a domain of interest. It describes the concepts and categories that are relevant to a particular domain and the relationships that exist between them. An ontology typically consists of a vocabulary of terms and a set of rules for combining these terms to represent knowledge about the domain.

An ontology is expressed in a machine-readable format and consists of a set of concepts (or classes), properties (or relationships), and axioms (or logical statements/constraints). Ontologies are implemented through technologies such as RDF (Resource Description Framework), OWL (Web Ontology Language), and SKOS (Simple Knowledge Organization System). One common use of ontologies is in the development of knowledge graphs, which are large networks of concepts and relationships that can be used for data integration, search, and analysis.

Example:

- SNOMED CT described below is a terminology that is supported and organised through an ontology model.
- LOINC, UMLS, HL7 FHIR (and most information models) described below are not primarily organised through an ontology but include ontology components in the definition of structures and relationships
- Human Phenotype Ontology (HPO) [13] provides a standardised vocabulary of phenotypic abnormalities that are observed in human disease.
- Gene Ontology (GO) [14] provides standardised vocabulary of gene function that is used to annotate gene products in all organisms, including humans.

Initiatives to harmonise ontologies

- National Center for Biomedical Ontology (NCBO) - or BioPortal - [15] is a web-based repository and ontology library that provides access to hundreds of biomedical ontologies, controlled vocabularies, thesauri and classifications. BioPortal includes a suite of tools for searching, browsing, and visualising ontologies, as well as for mapping and integrating data across different ontologies.

- The Mondo Disease Ontology (Mondo) [16] aims to harmonise disease definitions across the world. The name Mondo comes from the latin word 'mundus' and means 'for the world.'

Thesaurus

Definition. A thesaurus (plural thesauri), sometimes called a dictionary of synonyms, arranges words by their meanings, sometimes as a hierarchy of broader and narrower terms, sometimes simply as lists of synonyms and antonyms. Most thesauri do not include definitions, but many controlled vocabularies/terminologies include listings of synonyms.

Although providing unified coding for the terms used in healthcare, the thesauri are not sufficient for providing healthcare interoperability, because they contain only simple semantic relations between terms, like synonyms, but still lack the semantic context of used terms.

Examples

- The **Unified Medical Language System (UMLS)** [2] metathesaurus contains terminology, classification and coding standards.
- The **Medical Subject Headings (MeSH)** [3] thesaurus is a controlled and hierarchically-organised vocabulary produced by the National Library of Medicine. MeSH includes the subject headings appearing in MEDLINE/PubMed.
- The **NCI Metathesaurus (NCIm)** [4] is a wide-ranging biomedical terminology database that covers most terminologies used by NCI for clinical care, translational and basic research, and public information and administrative activities.
- **XEVMPD (eXtended EudraVigilance Medicinal Product Dictionary)** is a database - based on the IDMP information model - created by the European Medicines Agency (EMA) that contains information on medicinal products authorised for use in the European Union (EU); it could be used as thesaurus for the naming of medicinal products as it link active ingredients names - as used in ATC and defined WHODrug - with brand names.

Classification/ Taxonomy

Definition. Classification is the process of identifying and and grouping objects or ideas into predetermined categories. A medical classification is used to transform descriptions of medical diagnoses or procedures into standardised statistical code in a process known as clinical coding.
A taxonomy is a scheme of classification, especially a hierarchical classification, in which things are organised into groups or types. Many taxonomies are hierarchies (and thus, have an intrinsic tree structure), but not all are.

Examples

- The classifications provided below are developed by the World Health Organization (WHO) mainly for (statistical) reporting.
  - International Statistical Classification of Diseases and Related Health Problems
    - **ICD-10.** The latest version [12] was published in 2023. ICD-10 has 22 chapters, where Chapters 1 to 17 deal with a specific type of disease, and Chapters 18 to 22 deal with other types of health problems.
    - **ICD-11** [13] - In 2018 the WHO released a new version of Mortality and Morbidity Statistics (ICD-11 MMS). The active application of this version started on 1 January 2022.

        - International Classification of Diseases for Oncology (**ICD-O)** [14] –includes topography for sites of hematopoietic and reticuloendothelial tumours. It is currently in its third revision (ICD-O-3).
    - International Classification of Functioning, Disability, and Health (**ICF)** [15], includes a list of environmental factors.
    - **ATC** [17], Anatomical Therapeutic Chemical, Classification System is a drug classification system that classifies the active ingredients of drugs based on their main therapeutic use and their mode of action. Each drug in the dictionary is assigned an ATC code that indicates its anatomical, therapeutic, and pharmacological properties.
    - International classification of primary care (ICPC) [18] is a classification method for primary care encounters. It allows for the classification of the patient’s reason for encounter (RFE), the problems/diagnosis managed, primary or general health care interventions, and the ordering of the data of the primary care session in an episode of care structure. It was first published in 1987; a revision and inclusion of criteria and definitions was published in 1998 and was accepted within the WHO Family of International Classifications.
- Other classifications exist to support reimbursement such as the healthcare Common Procedure Coding System (HCPCS) in the US, which is an adapted and extended version of CPT codes by Centers for Medicare and Medicaid Services (CMS). CPT [17] (Current Procedural Terminology) is a Coding System for Patient Care Procedures, developed in 1966 by the American Medical Association (AMA); it is mapped to SNOMED. Similar reimbursement classifications exist in other countries; they are often mapped to medical terminologies - such as SNOMED and/or classification such as ICD.

Controlled Vocabulary/ Terminology

Definition. A controlled vocabulary - or terminology - is an organised arrangement of words and phrases in a particular field, used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms (also called synonyms). Terminologies can be organised through a thesaurus (hierarchy with broader terms at the top of the hierarchy and narrower terms below) or through an ontology (more formal and structured representation of relationships between concepts.

Example:

- **LOINC** [1] (Logical Observation Identifiers Names and Codes): is a standard for coding health observations, measurements, and documents. LOINC provides translations of its documents and terms into various languages. LOINC is primarily organised through a thesaurus but ontological representation; it is fully mapped with SNOMED. LOINC includes UCUM (Unified Code for Units of Measure), a standardised code system used for representing units of measurement.
- **SNOMED CT** [19] (Systematized Nomenclature of Medicine-Clinical Terms) is an organised collection of medical terms - procedures, diseases, and clinical findings.

  The EU4 Health budget (see annex in [20]) allocated to digital health in 2021, includes grants for SNOMED to support convergence towards the use of SNOMED CT and ensure that the EU patients have their data available in all the EU languages for reuse and sharing.

  SNOMED International, as part of its 2020-2025 strategy, has committed to acting as a central hub, or terminology integrator, for healthcare terminologies, and to pursuing alliances and partnerships with other international standardisation bodies to harmonise healthcare terminology across multiple domains. The figure below provides a list of the standards that have been aligned as of 2022.

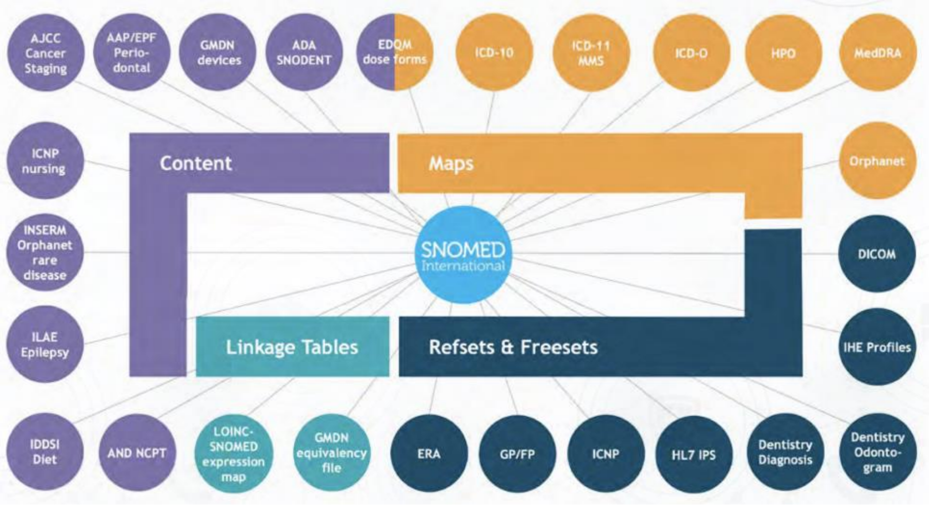


Figure 4. SNOMED CT alignment with other standards

- **ORDO,** or Orphanet Rare Disease Ontology, [21] is a controlled vocabulary and classification system for rare diseases and related phenotypes. It is developed and maintained by the Orphanet database, which is a reference resource dedicated to information on rare diseases and orphan drugs.
- The Medical Dictionary for Regulatory Activities Terminology (**MedDRA)** [22] is a standardised medical terminology to facilitate the sharing of regulatory information internationally for medical products used by humans.
- **WHODrug** Dictionary [23] is a comprehensive global drug terminology database developed and maintained by the World Health Organization (WHO). It is a standardised dictionary of drug names, which includes information on the active ingredients, strength, dosage form, route of administration, and other drug-related information. The WHO Drug Dictionary uses the ATC classification system to organise drugs according to their therapeutic and pharmacological properties.

Information Model

Definition. An information model specifies the format and meaning of data elements, as well as any relationships or constraints between them in a specific domain or context within an entity-relationship model. The semantic of the data elements and their relationship is typically not explicit or simply not included as not needed for the specific context of use in which the information model is used. An information model can provide a sharable, stable, and organised structure of information requirements or knowledge for a specific domain.

Information models use terminologies to instantiate coded data elements.

Examples

- **OMOP** (Observational Medical Outcomes Partnership) [24] Common Data Model is designed to standardise the structure and content of observational data and to enable efficient analyses that can produce reliable evidence; it is mainly used in clinical research. As an example, the EHDEN project decided to map all data to the OMOP common data model to support interoperability and reuse of healthcare data [25].
- **BRIDG** [26] metamodel for clinical trial representation mainly used by NCI in clinical research.
- **IDMP** (Identification of Medicinal Products) [27] is developed by the European Medicines Agency (EMA) for the identification and description of medicinal products. The IDMP information model defines a set of data elements and relationships that describe the characteristics of a medicinal product, including its active ingredients, pharmaceutical form,

strength, route of administration, marketing authorization status, and other relevant information; it is intended to be used in conjunction with established terminologies, such as the International Nonproprietary Names (INN), the Anatomical Therapeutic Chemical (ATC) classification system, and SNOMED CT.

Data Exchange Standards

Definition. Formal rules for the structure of data elements, typically based on an information model and terminologies. A data exchange specification is a common model used across organisations that standardise the format in which data will be shared. It makes use of terminologies and classification.

Examples

- HL7 CDA - (Clinical Document architecture) Release 2 [28] defines the structure and semantics of medical documents for the purpose of exchanging data. CDA documents are coded in XML. A CDA document includes a Header with metadata about the document, the Clinical Document Body with actual clinical content included into several sections and Signatures and Authentication. HL7 CDA is gradually replaced by HL7 FHIR.
- HL7 FHIR (Fast Healthcare Interoperability Resources) is a standard for exchanging clinical and administrative healthcare data. It provides a flexible, web-based framework with a RESTful API, and standardised data formats like XML and JSON. Within HL7, the General-Purpose Data Types (GDTs) [29] represent standardised sets of objects - building on primitive data types to represent objects such as address, person name, quantity, ….
- CDISC SDTM (Study Data Tabulation Model) [30] is a standard for organising and formatting data collected during clinical trials in a consistent and meaningful way. SDTM is one of the required standards by the FDA (U.S.) and PMDA (Japan) to submit clinical trials data as part of an electronic submission of new drug applications (NDA). It is not in use in Europe as regulators do not require submission of the clinical trial data for market authorisation.

3.2.3 Conclusion: Impact for the AIDAVA Reference Ontology

A first conclusion that can be drawn from the review of the key initiatives and the list of standards mentioned above is that some standards clearly stand out in terms of use and acceptance across the healthcare community in Europe; they should be considered as priority standards to be included in the AIDAVA Reference Ontology.

- SNOMED CT as an overarching "hub",
- LOINC for laboratory procedures,
- HL7 FHIR ontological components, starting with the General Purpose Data Types [29].

A second conclusion that can be drawn is that these standards are not sufficient: they do not cover all medical domains - where specialised standards will be required - and they do not include all the needed relations as depicted in the figure below This example displays the representation of familyCondition in SNOMED CT and in HL7 FHIR: SNOMED CT states that there is a person in the family with a disease; FHIR goes into more details and specifies that there is a person in the family - with a specific relationship who has a disease.

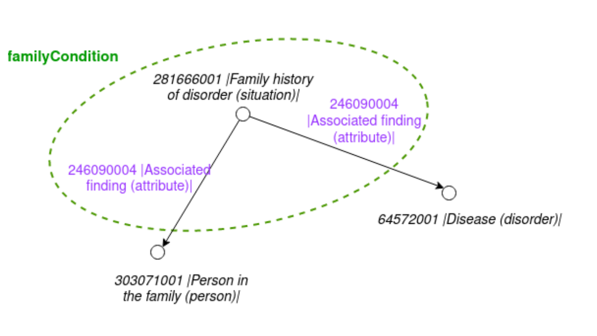
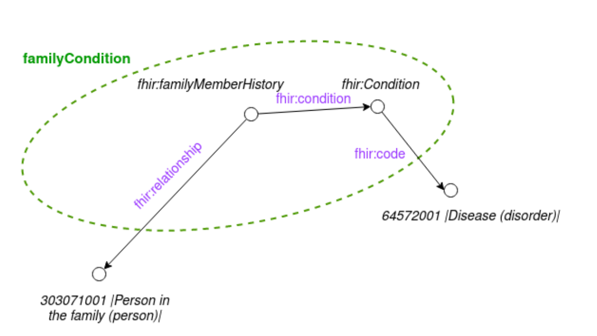
SNOMED CT	FHIR
 familyCondition 281666001 |Family history of disorder (situation)| 246090004 |Associated finding (attribute)| 246090004 |Associated finding (attribute)| 64572001 |Disease (disorder)| 303071001 |Person in the family (person)|	 familyCondition fhir:familyMemberHistory fhir:Condition fhir:condition fhir:code fhir:relationship 64572001 |Disease (disorder)| 303071001 |Person in the family (person)|

Figure 5. Relationships included in FamilyCondition in SNOMED and FHIR

Expanding the AIDAVA Reference Ontology will need to be done at 2 levels

1. Extend the scope of the ontology by adding other terminologies or ontologies related to other domains such as ORDO for rare diseases and HPO for genomics.
2. Enrich the relationships across concepts, extracting this knowledge from information models such as HL7 FHIR profiles - or potentially openEHR/ ISO13606 [31].

This allows to transform Ontology Strategic Requirement 2 (OSR2) into a more concrete requirement. For the scope of AIDAVA it is expected that the priority standards will be sufficient.

Ontology Requirement Specification 6. The AIDAVA Reference Ontology must include as a priority

- SNOMED CT, LOINC (including UCUM),
- Constraints related to HL7 FHIR General-Purpose Data Types, and core HL7 FHIR profiles related to IPS

Additional terminologies and standards should be assessed through a strict governance process, whenever there is an extension of the domain.

3.3 AIDAVA Reference Ontology

AIDAVA aims to build patient health knowledge graphs (PHKG) that follow the FAIR principles [32], with data extracted from multiple sources, both structured and unstructured. It has always been a challenge that data is usually collected from multiple sources and in various formats, meaning that data can be represented and interpreted in different ways. This makes it difficult to effectively use data in a research context - difficult to collect, connect and understand data coming from diverse sources. In order to resolve the above issues and to provide a common harmonisation layer, we implemented the AIDAVA Reference Ontology that will be used for internal semantic alignment between all steps within the process (onboarding of data sources, execution of FAIRification workflows and generation/enrichment and validation of the PHKG) ensuring the exchange of health-related data in an interoperable manner.

In this section, we refer to the AIDAVA Reference Ontology at 2 different levels, following the approach of the SPHN initiative [33].

- A **dataset** is a collection of discrete items (concepts) in the domain in scope that describe the semantic (meaning) of shareable data elements in non expert terms; as such, it offers a flexible way to maintain the domain model by non ontology experts. The meaning of a concept is expressed by value sets, terminologies and ontologies (e.g. SNOMED CT, LOINC).

- A **schema** is the machine usable format - in Resource Description Framework (RDF) [34] - representing the semantics of the concepts defined in the dataset; it is maintained by the ontology experts, based on the concepts identified in the dataset. It is used to ensure interoperability across different data sources and facilitates data exchange.

There is a strict relationship between the dataset and the schema as displayed in Figure 6.

- A concept in the dataset can be mapped to a class, an instance of a class or a property in the RDF schema.
- In the RDF Schema, a class is further refined by properties (object properties and data properties).
- Interchangeable concepts from the dataset representing the same class or the same instance of a class, are mapped using *owl:equivalentClass*;interchangeable concepts representing the same property are mapped to each other using *owl:equivalentProperty*.

3.3.1. AIDAVA Dataset

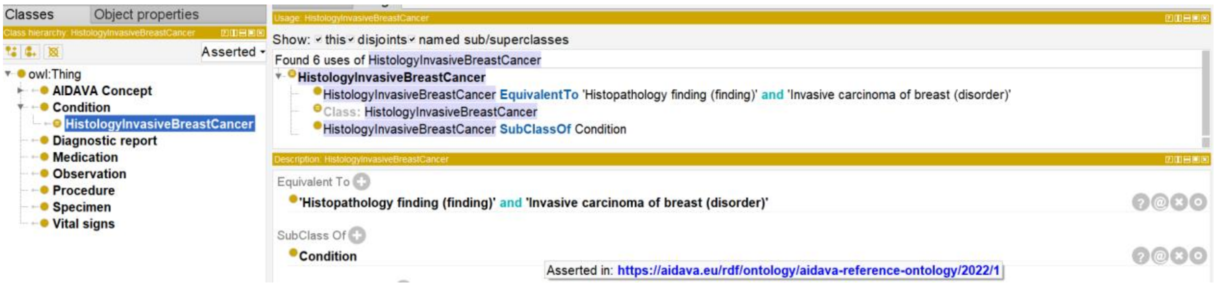
There are many definitions of a "dataset" as a meaning, but the one that we use in the context of AIDAVA project aligns with this definition- *"A dataset is a collection of related to a specific domain, discrete items (concepts) of related data that may be accessed individually or in combination."* The AIDAVA dataset describes the domain in scope.

In alignment with Ontology Requirement Specifications 1,2,3,4 the AIDAVA initial dataset includes health-related concepts (or terms) required by the two use cases – Cardiovascular (CVD) and Breast Cancer Registry (BCR) - and for delivering IPS. More precisely this includes the 80+ data elements for Breast Cancer and BC and 50+ data elements for the CVD score as well as data elements and relations needed to support IPS. As the scope of IPS is broad we will focus on the HL7 FHIR profiles that are most relevant for the 2 use cases, including

- Condition
- VitalSigns
- Medication
- Observation (e.g. Tobacco use)
- Procedure
- DiagnosticReport
- Specimen

Alignment of AIDAVA Reference Ontology with established international ontologies, such as SNOMED CT and LOINC, allows to leverage the specific domain knowledge that is expressed in these ontologies. Even more, the re-use of these internationally recognized standards is supporting the achievement of semantic interoperability with other external systems.

Some of the data elements in scope of the two use cases, are not covered by pre-coordinated concepts from any of the identified terminologies included in the AIDAVA Reference Ontology. For example, there is no single term that can represent "Histology invasive breast cancer", however combining SNOMED CT "250537006 |Histopathology finding (finding)" and "713609000 |Invasive carcinoma of breast (disorder)" captures precisely the meaning of the targeted concept. Usage of such "composite" concepts provides both precision in semantic description of the data elements and flexibility in referring to established terminologies. The composite concept described in the AIDAVA Dataset is represented by a set of concepts in a *owl*:*equivalentClass property* into structures like shown in the example below:


Classes
Object properties
Class hierarchy: HistologyInvasiveBreastCancer
Asserted
Usage: HistologyInvasiveBreastCancer
Show: this disjoints named sub/superclasses
Description: HistologyInvasiveBreastCancer

owl:Thing
- AIDAVA Concept
- Condition
  - HistologyInvasiveBreastCancer
- Diagnostic report
- Medication
- Observation
- Procedure
- Specimen
- Vital signs

Found 6 uses of HistologyInvasiveBreastCancer

HistologyInvasiveBreastCancer
- HistologyInvasiveBreastCancer EquivalentTo 'Histopathology finding (finding)' and 'Invasive carcinoma of breast (disorder)'
- Class: HistologyInvasiveBreastCancer
- HistologyInvasiveBreastCancer SubClassOf Condition

Equivalent To

'Histopathology finding (finding)' and 'Invasive carcinoma of breast (disorder)'

SubClass Of

Condition

Asserted in: https://aidava.eu/rdf/ontology/aidava-reference-ontology/2022/1

Figure 6. An example of defining composite concepts using owl:equivalentClass and owl:objectIntersectionOf

3.3.2 AIDAVA Reference Ontology schema

3.3.2.1 Scope

AIDAVA Reference Ontology schema serves as an interoperable framework for exchange of health-related information and storing it in the corresponding patient health knowledge graphs (PKHGs) using Semantic Web technologies like RDF and OWL. The AIDAVA Reference Ontology schema is based on the AIDAVA Dataset and transforms its elements into a formal ontological structure (see Figure 6).

- All concepts described in the AIDAVA Dataset are translated into classes (*owl:Class*) or properties (*owl:property*) in the AIDAVA RDF schema.
- A value set is represented as an individual (*owl:NamedIndividual*).
- A class (or instance of a class) is represented by a code from one - or potentially several - value sets. When the code from different value sets represent the same class (or instance of a class) - and therefore have a meaning binding - they are mapped using *owl:equivalentClass*.
- A class is further refined by properties. A property can have different types; based on the type, it can be either a *owl:ObjectProperty,* i.e. a relationships between classes (or instance of classes), such as connecting a doctor to their patients, or a *owl:DatatypeProperty* which a characteristics of class, such as the age of a patient.
- A property is also represented by a code from one - or potentially several - value sets. When the code from different value sets represent the same property, they are semantically equivalent and are mapped using *owl:equivalentProperty.*

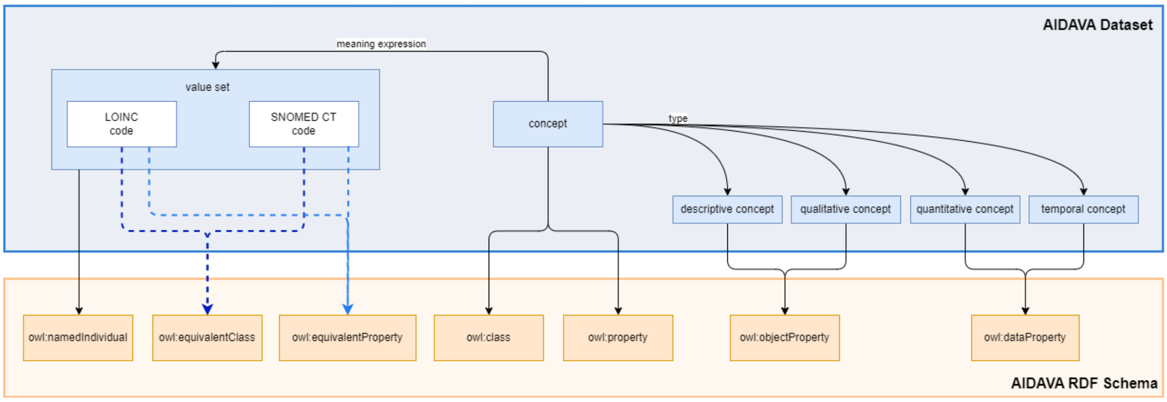


Figure 7. AIDAVA Dataset and reference ontology schema interaction.

The schema facilitates the integration of external resources allowing definition of owl:equivalentClass and owl:equivalentProperty mappings as well as other semantic mappings expressed by SKOS.

3.3.2.2 Technical specification

3.3.2.2.1 Namespace

The namespace of the AIDAVA Reference Ontology can be dual:

- Ontology IRI (*https://aidava.eu/rdf/aidava-ontology/*) - The ontology IRI remains fixed and is defined in the AIDAVA Reference Ontology schema. The ontology IRI is the "base prefix" and will be used to describe the data (during the annotation process) and to query the data in the PHKG (by the data consumers).
- Version IRI (*https://aidava.eu/rdf/aidava-ontology/v1/*) - It is provided for each published release of the AIDAVA Reference Ontology and allows to distinguish different versions of the schema. The versioned IRI can be used to refer to a specific version of the AIDAVA Reference Ontology schema and thus it needs to be included in the header of all datasets generated using this schema version.

3.3.2.2.2 Versioning

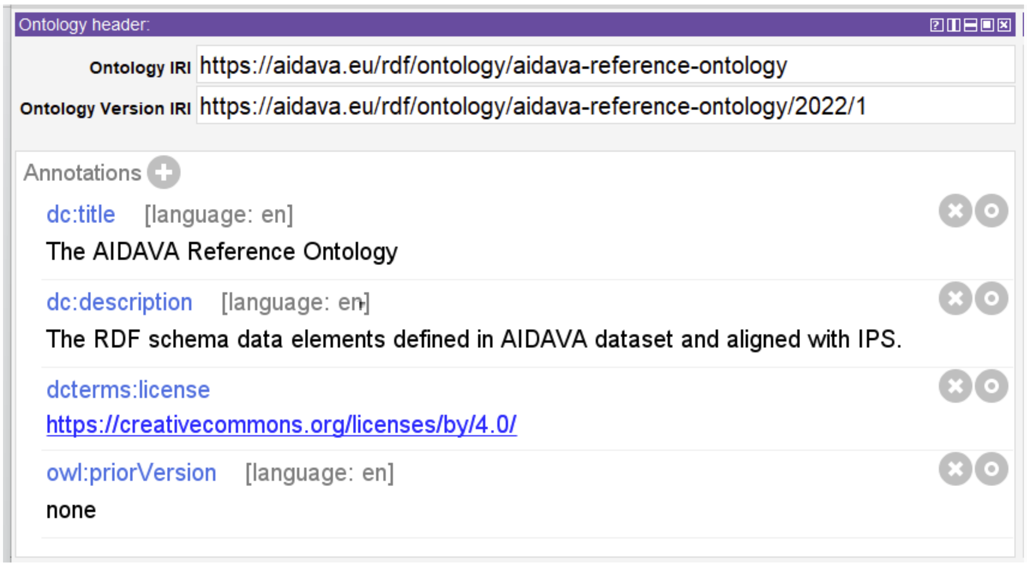
Each release of the AIDAVA Reference Ontology schema has a specific version associated with it. The version, indicated by the tag *owl:versionIRI* in the ontology header section. The published ontology IRI needs always to point to the latest version IRI of the AIDAVA Reference Ontology schema. Each version will include the following information

- Version number data-based, i.e. linked to the the date at which the version was released in a YYYY.MM.DD format
- Standards used, and their version
- Licensing information - as the AIDAVA ontology is based on standards such as LOINC and SNOMED which are under licensing agreement
- Release notes describing the changes
- Description of the AIDAVA Dataset

3.3.2.2.3 Header

The header of the AIDAVA Reference Ontology schema contains the following information:

- *dc:title*- the title of the schema
- *dc:description*- a short description of the ontology
- *dcterms:license* - the license terms and conditions of usage of the ontology
- *owl:versionIRI* - the version of the schema
- *owl:priorVersion* - the previously released version



Ontology IRI https://aidava.eu/rdf/ontology/aidava-reference-ontology

Ontology Version IRI https://aidava.eu/rdf/ontology/aidava-reference-ontology/2022/1

dc:title [language: en]
The AIDAVA Reference Ontology

dc:description [language: en]
The RDF schema data elements defined in AIDAVA dataset and aligned with IPS.

dcterms:license
https://creativecommons.org/licenses/by/4.0/

owl:priorVersion [language: en]
none

Figure 8. Sample ontology header

3.3.2.2.4 Classes

The classes defined in the AIDAVA Reference Ontology schema are derived from the names of the data elements in the AIDAVA Dataset. We aim at one data element to be represented by one concept, which represents one class. The unique identifier of a class corresponds to a concatenation of the words (if more than one) forming the concept written in an UpperCase format. For example, the concept Clinical T Stage is defined as a ClinicalTStage class in the RDF schema.

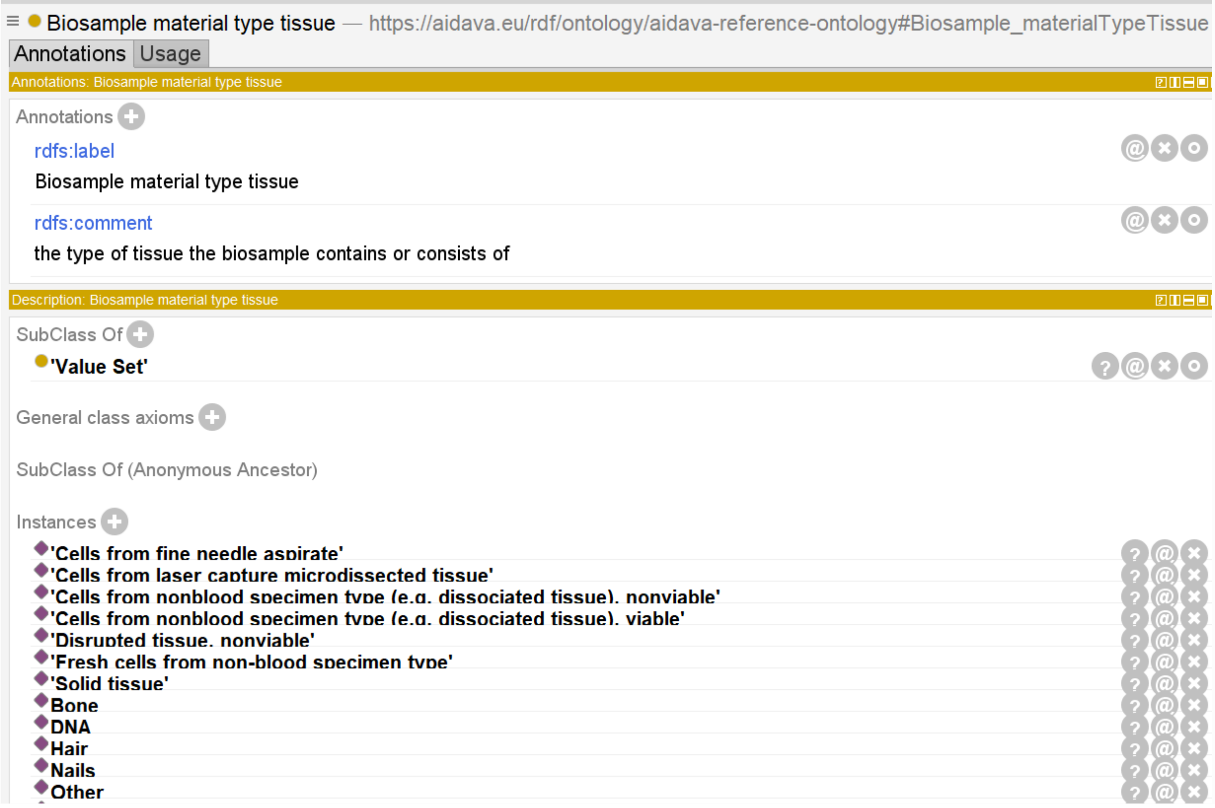
To improve readability in the AIDAVA Reference Ontology few additional properties are introduced:

- An ***rdfs:label*** corresponds to the human readable value for the class name
- An ***rdfs:comment*** represent more detailed description of the class if for some reason it might differ from the associated ***owl:equivalentClass*** concepts (e.g. from SNOMED CT or LOINC)

The majority of the classes have a meaning binding associated with a SNOMED CT, LOINC and other ontologies concept or code. This meaning binding is represented in the schema with the annotation using ***owl:equivalentClass***.

In addition to the main classes defined in the AIDAVA Dataset, there are also additional classes that need to be introduce, in order to refer to other required metadata:

- **High level** classes aligned with the overlapping IPS profiles - condition, diagnosticReport, procedures, etc.
- A **Terminology class** is used to organise groups of classes and individuals originating from external resources (e.g. SNOMED CT, LOINC) used within the AIDAVA. They are required in order to be able to refer to them as equivalent classes (i.e. the SNOMED CT code 184100006 is a an equivalent concept for the AIDAVA **Gender** class), or possible value (i.e. the SNOMED CT code 24028007 is allowed value for AIDAVA **Laterality** class), or even individuals.
- A **ValueSet class**, which groups classes defining specific AIDAVA instances of possible values to use for certain object properties.



Biosample material type tissue — https://aidava.eu/rdf/ontology/aidava-reference-ontology#Biosample_materialTypeTissue

Annotations: Biosample material type tissue

rdfs:label
Biosample material type tissue

rdfs:comment
the type of tissue the biosample contains or consists of

Description: Biosample material type tissue

SubClass Of
'Value Set'

Instances
'Cells from fine needle aspirate'
'Cells from laser capture microdissected tissue'
'Cells from nonblood specimen type (e.g. dissociated tissue), nonviable'
'Cells from nonblood specimen type (e.g. dissociated tissue), viable'
'Disrupted tissue, nonviable'
'Fresh cells from non-blood specimen type'
'Solid tissue'
Bone
DNA
Hair
Nails
Other

Figure 9. An example of a ValueSet class (BioSample Tissue Type)

- A **DataTypes class** is used in order to organise the primitive and general purpose data types defined to represent the data types used in FHIR (e.g. Quantity, Period, Range, Duration).

Some hierarchies are defined in the AIDAVA Reference Ontology schema. For instance, the class **Therapy** is a parent of **Hormone therapy** represented by 169413002 SNOMED CT code. The rules of inheritance of properties is applied over ***owl:subClassOf*** based hierarchies - all properties annotated at a parent class are automatically inherited in the children class. Thus it is only required to define the ***rdfs:Domain*** for the parent class.

3.3.2.3.5 Properties

Each of the data elements in the AIDAVA Dataset have defined attributes that are translated in the AIDAVA Reference Ontology schema to object properties (relationship between individuals) or data properties (relationship of individual to literal values).

The IRIs of properties are named after the general concept name column in the AIDAVA Dataset. The convention used is to define properties with has + <general concept name> as the resultant local name of the IRI is structured in the so-called "camel case" and can be schematically represented as "hasTokenToken".

Object properties in the AIDAVA Reference Ontology schema define relationships between:

- Resources from the clinical data (a given instance of **Laterality** is connected to a patient's **Breast structure**) and
- Resources and elements represented by other, external terminology systems (the **Period** is defined by FHIR's **Period** complex data type)

An additional set of four object properties are generated in the RDF schema, not represented in nor required by the AIDAVA Dataset:

- ***hasPatientPseudoIdentifier***, connecting information to the patient identifier.
- ***hasDataProvider***, connecting information to the data provider.
- ***hasExtractionDateTime***, providing information about the time of extraction of the data.
- ***hasProvenance*** relates to the source of the information.

In the AIDAVA Reference Ontology schema data properties represent literal values of given concepts. These data properties represent values for the common data types, like for example:

- xsd:string - ***rdfs:label*** represents literal values for concept names using xsd:string data type.
- xsd:double – ***ageAtDiagnosis*** represents numerical values for patient's age using xsd:double data type
- xsd:date – ***dateOfDiagnosis*** represent date on which the patient was diagnosed using xds:date type
- and others.

An ***owl:Restriction*** is a class description that is used to define various types of logical rules applied on the values for that property, like value and cardinality constraints. For a single property, there can be multiple constraints that can be used to further specify (to narrow down) the concept, which is described. Value constraints are usually used to restrict the possible types of values of a property, while cardinality constraints restrict the number of instances of a property for this class. The rules for inheritance of restrictions are applied – if there are any restrictions defined on a property, they are automatically applied also on the sub-properties. This means that if a has a sub-property of a certain property, the restrictions of this sub-property must be in the range of the restrictions of the parent property.

- **Value constraint** – usually determines the data type range of a data property.

- **Cardinality constraint** - Cardinalities constraints have been implemented to restrict the number of values an instance of a class may have for specific properties. The ***owl:minCardinality*** and ***owl:maxCardinality*** notation are used.

3.3.2.2.6 Development process

Link to Github with current version of AIDAVA Reference Ontology is available at the following address:

https://github.com/AIDAVA-DEV/AIDAVA-Reference-Ontology

3.3.3 Implementation of SHACL rules

In the AIDAVA project, we will use the SHACL shape language to define data quality checks, which will be partially generated from the AIDAVA ontology. We have identified three types of constraints: cardinality constraints, value constraints, and logical consistencies (including general rules and domain specific rules).

For example, let's consider a datasource that describes a patient's date of birth. The following constraints apply:

- The datasource must provide exactly one value for the patient's date of birth.
- The date of birth value must be in the format "\d{2}/\d{2}/\d{4}$" (e.g., "dd/mm/yyyy").
- The date of birth can be later than the hospital arrival date.

Some of these constraints will be automatically converted from the ontology document using tools such as SHACLer [35] and Astrea[5] - as provided below.

```
shape:PatientShape a sh:NodeShape ;
    sh:targetClass aidava:Patient ;
    sh:property [
        sh:path aidava:dateOfBirth ;
        sh:datatype xsd:date ;
        sh:message "Patients dateOfBirth is not a date value" ;
    ] ;
    sh:property [
        sh:path aidava:dateOfBirth ;
        sh:maxCount 1 ;
        sh:minCount 1 ;
       sh:message "Patient’s dateOfBirth is registered more/less than one
time" ;
    ] ;

  sh:property [
        sh:path aidava:dateOfBirth ;
        sh:lessThanOrEquals aidava:patientAdmissionStartDate ;
        sh:message "Begin Date of Admission must be AFTER Birth Date." ;
    ] .
```

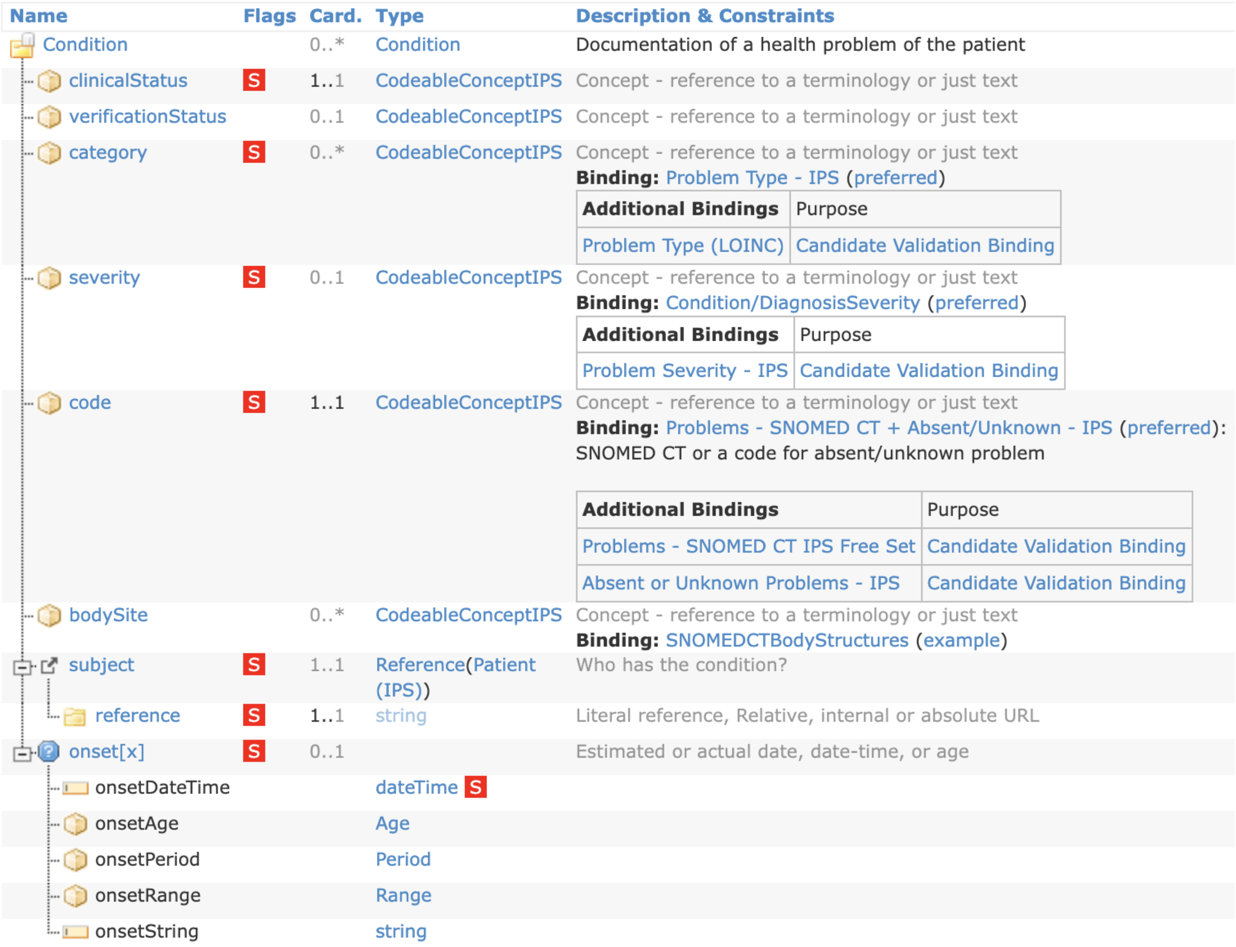
[5] http://astrea.helio.linkeddata.es/

3.3.4 Mapping from the Ontology to target data models

There are 3 target models in the AIDAVA project as described before.

- 80+ data elements for the Breast Cancer registry and 50+ data elements to compute the CVD Smart score . These are included in the initial AIDAVA Dataset, and fully mapped.
- The patient IPS. AIDAVA Reference Ontology schema aligns with FHIR IPS implementation where possible. The initial AIDAVA Dataset covers some of the HL7 FHIR IPS profiles identified above and repeated here.
  - Condition
  - VitalSigns
  - Medication
  - Observation (e.g. Tobacco use)
  - Procedure
  - DiagnosticReport
  - Specimen

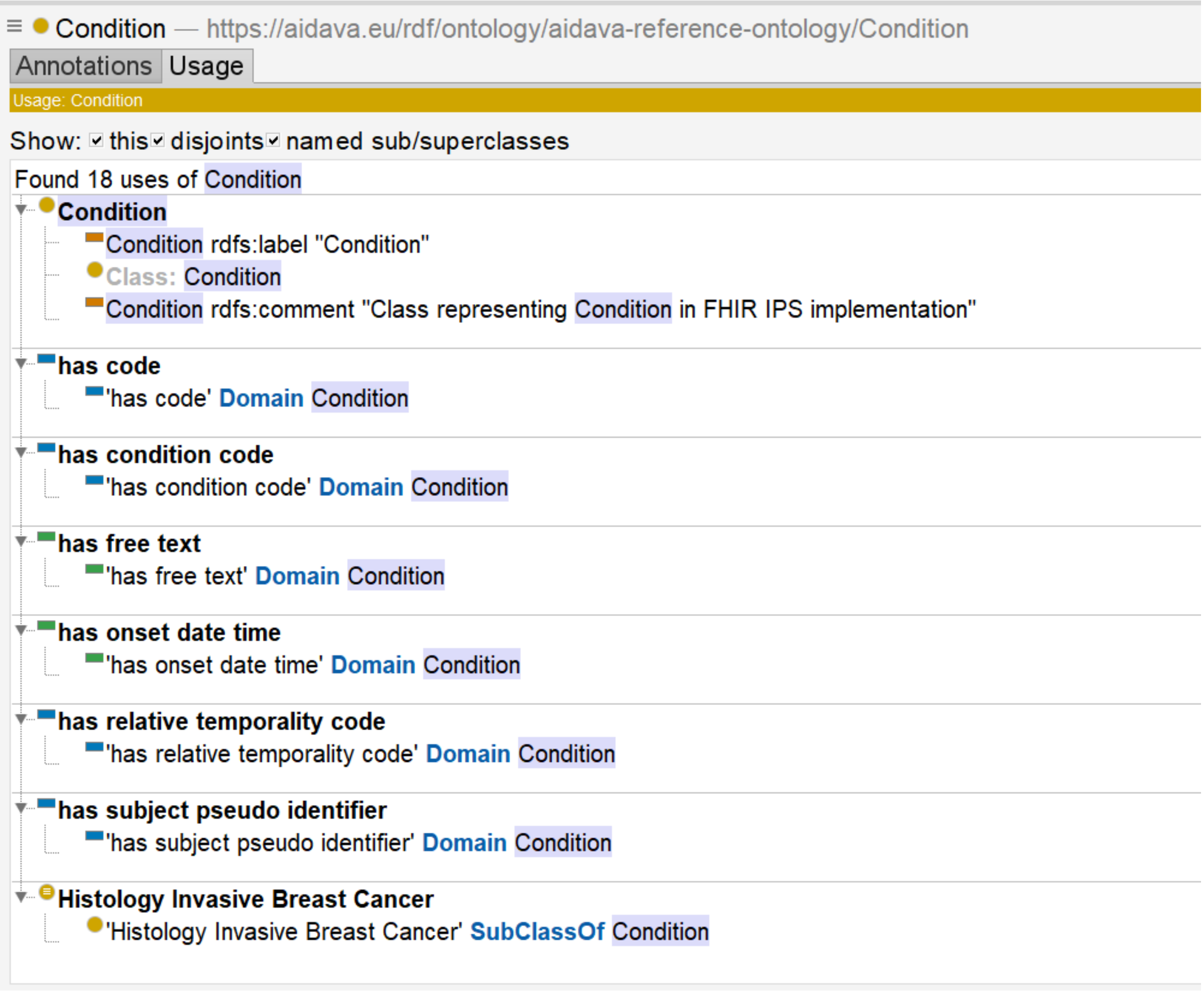
The list of covered IPS profiles can be extended if required through the governance process. AIDAVA Reference Ontology defines a representative class for each of the profiles, which is described with a set of properties as required by the IPS specification. For example the profile for Condition is defined in FHIR IPS implementation with the following set of properties:



Name	Flags	Card.	Type	Description & Constraints
Condition		0..*	Condition	Documentation of a health problem of the patient
clinicalStatus	S	1..1	CodeableConceptIPS	Concept - reference to a terminology or just text
verificationStatus		0..1	CodeableConceptIPS	Concept - reference to a terminology or just text
category	S	0..*	CodeableConceptIPS	Concept - reference to a terminology or just text **Binding:** Problem Type - IPS (preferred) **Additional Bindings** | Purpose Problem Type (LOINC) | Candidate Validation Binding
severity	S	0..1	CodeableConceptIPS	Concept - reference to a terminology or just text **Binding:** Condition/DiagnosisSeverity (preferred) **Additional Bindings** | Purpose Problem Severity - IPS | Candidate Validation Binding
code	S	1..1	CodeableConceptIPS	Concept - reference to a terminology or just text **Binding:** Problems - SNOMED CT + Absent/Unknown - IPS (preferred): SNOMED CT or a code for absent/unknown problem **Additional Bindings** | Purpose Problems - SNOMED CT IPS Free Set | Candidate Validation Binding Absent or Unknown Problems - IPS | Candidate Validation Binding
bodySite		0..*	CodeableConceptIPS	Concept - reference to a terminology or just text **Binding:** SNOMEDCTBodyStructures (example)
subject	S	1..1	Reference(Patient (IPS))	Who has the condition?
reference	S	1..1	string	Literal reference, Relative, internal or absolute URL
onset[x]	S	0..1		Estimated or actual date, date-time, or age
onsetDateTime			dateTime S	
onsetAge			Age	
onsetPeriod			Period	
onsetRange			Range	
onsetString			string	

Figure 10. Definition of the structure of "Condition"in HL7 FHIR IPS.

The reference ontology defines an *owl:Class* for Condition and specifies a set of *owl:objectProperty* and *owl:dataProperty* that will represent the set of properties that belongs to the FHIR IPS implementation profile. In addition the ontology will define similar restrictions that need to be applied on the various object and data properties - e.g. cardinality, valid data types of the values, etc.



● Condition — https://aidava.eu/rdf/ontology/aidava-reference-ontology/Condition

Annotations | Usage

Usage: Condition

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

Found 18 uses of Condition

- ● Condition
  - Condition rdfs:label "Condition"
  - Class: Condition
  - Condition rdfs:comment "Class representing Condition in FHIR IPS implementation"
- has code
  - 'has code' Domain Condition
- has condition code
  - 'has condition code' Domain Condition
- has free text
  - 'has free text' Domain Condition
- has onset date time
  - 'has onset date time' Domain Condition
- has relative temporality code
  - 'has relative temporality code' Domain Condition
- has subject pseudo identifier
  - 'has subject pseudo identifier' Domain Condition
- Histology Invasive Breast Cancer
  - 'Histology Invasive Breast Cancer' SubClassOf Condition

Figure 11. Definition of the "Condition" related concepts in AIDAVA Reference Ontology.

All classes within the ontology can be extended further with mappings to other terminologies and ontology systems using owl:equivalentClass relationships. Thus every class in the ontology can be mapped to SNOMED CT, LOINC or any other relevant ontology. In addition, semantically identical, similar or related concept across terminologies and ontologies can be modelled using W3C Simple Knowledge Organization System (SKOS), which provides the necessary tooling to specify the appropriate semantic of the relation - skos:exactMatch, skos:closeMatch and skos:related.

3.4 Governance Process

3.4.1 Purpose

The first version of the AIDAVA Reference Ontology is focused on the scope identified so far - and specified in the AIDAVA Dataset. This dataset - and the AIDAVA Reference Ontology - will need to be updated, through a strict governance process, at 2 levels.

1. Extension of the ontology - including concepts and data quality rules - delivered in May 2023, will be needed **during the project**, as new requirements appear when developing different components of the systems. Indeed, the first draft of the ontology will be focused on the data elements identified for the use cases, and mentioned before. When we will start to onboard the data sources that have been identified of relevance for these use cases, and when we will annotate text narratives, we may expect to have to include new concepts and/or properties and/or constraints in the ontology.
2. In order to keep the Reference Ontology up to date and to ensure its wider use and acceptance **beyond the project**, we need to align with the latest versions of the foundational ontologies (like SNOMED CT, LOINC) and address evolving needs for already covered and new use cases. In addition, we also need to reach out to external standard activities such as EEHRxF developed in the XPanDH project [36] to align with emerging EU standards. We also need to reach out to non-EU initiatives such as SPHN to maximise alignment toward interoperability with these initiatives.

Requirements for updates into the Reference Ontology may have an impact at different levels.

1. Ontology level
    a. AIDAVA Dataset and the Reference Ontology itself,
    b. Axioms, quality checks and constraints.
2. Related components
    a. AI tools (e.g. NLP) trained with the concepts defined in the ontology,
    b. Catalogue of Data Sources - updated during the Data Source onboarding process - which includes the mapping between source data and the ontology,
    c. All personal health knowledge graphs (PHKGs),
    d. Mapping to transform & publish the content of the PHKG into specific target formats.

3.4.2 Maintaining and expanding the ontology during the project

During the AIDAVA project, we expect that by the end of June 2023 for Generation 1 (G1), that ontology will represent the specified data elements for the use cases in scope and will be aligned to IPS, which corresponds to 90+% completion. Expected changes would mainly be extension, with potential requests for replacement; structural changes are not expected. PHKGs will not be impacted as they will be generated when G1 is deployed and will be aligned with the Reference Ontology version.

We may expect to have additional changes to the ontology for Generation 2 (G2) of the prototype after evaluation of G1. However, we do not expect these changes - mainly related to extension or replacement - to be major for several reasons.

- The AIDAVA Reference Ontology is based on mature terminologies; while the terminologies keep evolving with regular releases, the domains in scope of AIDAVA (Breast Cancer and CVD) are also mature domains where we should not expect much new knowledge - and therefore update in the terminologies.
- The main differences between G1 and G2 lay in the update of curation tools and the addition of a better human computer interaction module; these components are not expected to generate changes into the ontology.

Even if changes are expected to be limited after the first release of the AIDAVA Reference Ontology is delivered, we need to formally manage these changes to be consistent during the project and to learn how to maintain the ontology and related components beyond the project.

3.4.2.1 Roles and responsibilities

The following roles are identified as main actors in the ontology governance process:

- **Subject Matter Experts (SME)** - individuals that have extensive knowledge in the domain that is an object of interest in AIDAVA. They maintain the AIDAVA Dataset.

- **Change Facilitators** - individuals who have knowledge in the domain but also experience with the tools and techniques used to create formalised semantic models. Their responsibilities are to work with the SMEs, model the content as a set of artefacts for review, update the AIDAVA Dataset, and then to interface with the Semantic Engineer. The Change Facilitators are also responsible to address questions and resolve issues during the governance process and maintain the release notes.
- **Semantic Engineers -** individuals who maintain consistency of the AIDAVA Reference Ontology and keep robust and logically sound representation of the ontology concepts in the formalised ontological model. The Semantic Engineers are also responsible to publish new versions of the ontology, in agreement with the Change Facilitators.

3.4.2.2 Managing a change request

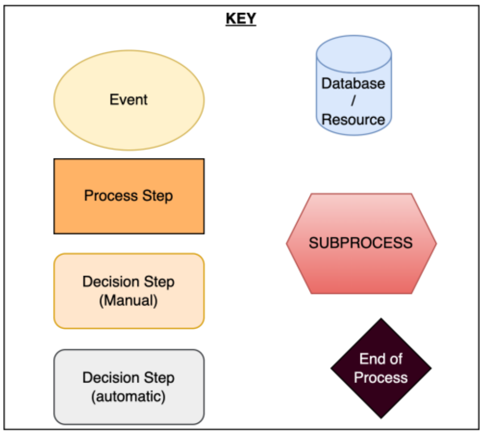


Figure 10. Legend for Governance Process.

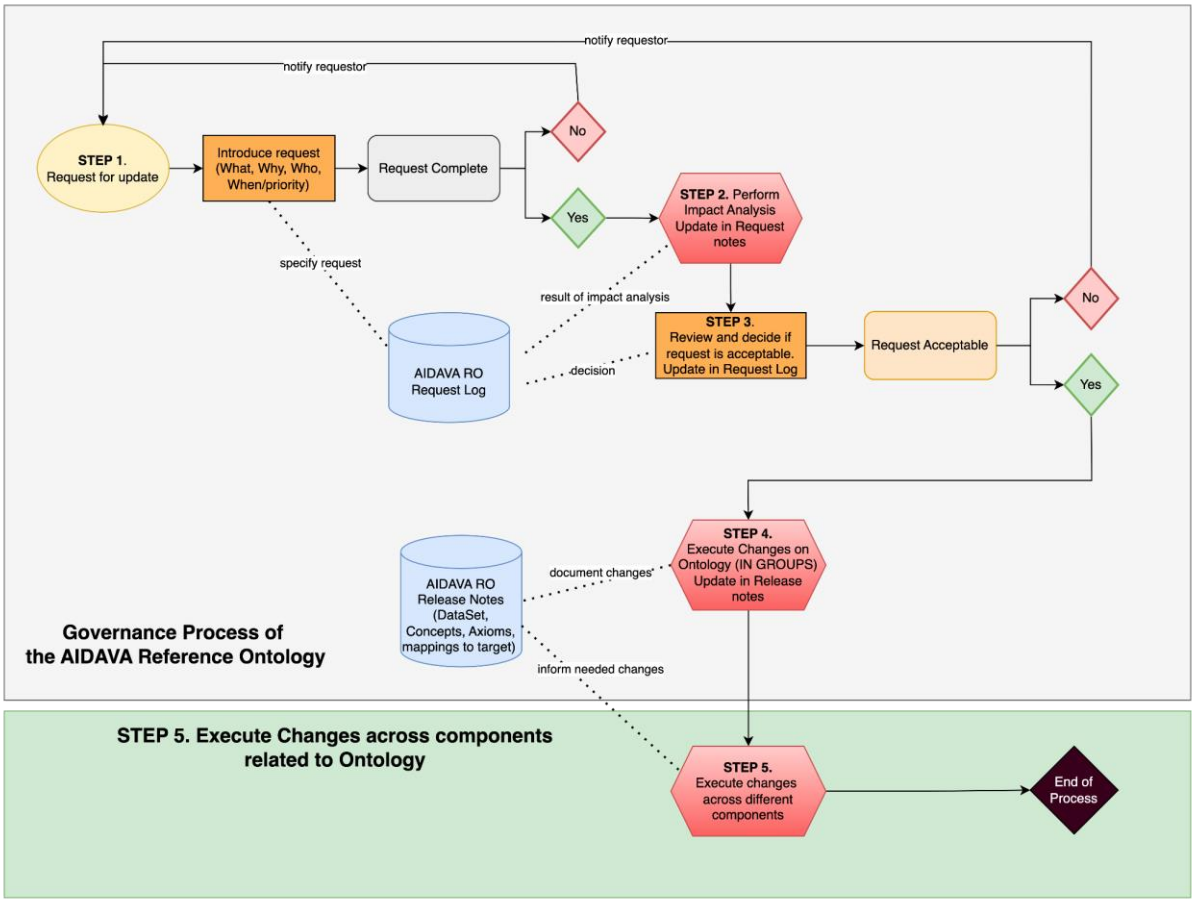

notify requestor
notify requestor
STEP 1. Request for update
Introduce request (What, Why, Who, When/priority)
Request Complete
No
Yes
STEP 2. Perform Impact Analysis Update in Request notes
specify request
result of impact analysis
AIDAVA RO Request Log
decision
STEP 3. Review and decide if request is acceptable. Update in Request Log
Request Acceptable
No
Yes
STEP 4. Execute Changes on Ontology (IN GROUPS) Update in Release notes
document changes
AIDAVA RO Release Notes (DataSet, Concepts, Axioms, mappings to target)
inform needed changes
Governance Process of the AIDAVA Reference Ontology
STEP 5. Execute Changes across components related to Ontology
STEP 5. Execute changes across different components
End of Process

Figure 11. Governance process for AIDAVA Reference Ontology

STEP 1. Define a request - by SME.
Any request will be introduced in the ***AIDAVA Reference Ontology Request Log*** managed through GitHub (see Section 3.4.2.3) with the following information

- What and which component is involved,
- Why - rationale for the request
- When - priority to introduce the change,
- Who is requesting the change)

STEP 2. Perform impact analysis - by Change Facilitators, supported by Semantic Engineers

- *Identify the type of request*
  - *Case 1. Request for ontology extension*
    - Request must specify: New data structure requested
    - Identify relevant terms
    - Identify new data patterns
    - Define ontology model representation (classes, properties, SHACL rules)
  - *Case 2. Replacement of concept in core ontology*

    This process covers the scenario of regular updates of the core ontologies re-used by AIDAVA Reference Ontology.
    - Request must specify: concept to be replaced within the AIDAVA Dataset
    - Ingestion of the new release of the core Dataset in a graph database together with the currently used release
    - DIff comparison and identification of obsolete and new terms across two versions.
    - Identify obsolete terms re-used by AIDAVA Reference Ontology (DataSet, Ontology and Axioms) and their new replacement terms.
    - Specify terms changes that news to be introduced in the ref sets.

    - *Case 3. Structural changes in core ontology*
        - Request must specify: structural changes that need to be done with justification in terms of possible usage scenarios.Identify classes and properties that will be affected by the change and what will be the impact on other downstream steps (onboarding, annotation, ingestion in PHKG, etc).
        - Propose new ontological structure (class, properties, axioms, SHACL validation rules)
        - Specify terms that will be used to represent the instances of the class
- *Consolidation*
    - Assess impact of the proposed change (on ontology level and on related components); conclude with recommendations
    - Update ***AIDAVA Reference Ontology Request Log*** with proposed recommendations

STEP 3. Review and approve proposed changes - by SME, Change Facilitator and Semantic Engineers
Based on the soundness and the impact of proposed changes, requests can be rejected (then the requestor must be notified with justification of the rejection) or approved, which will transition the process to the next step.

STEP 4. Update Reference Ontology and release new version - by Semantic Engineers
Note: As mentioned below the actual implementation of the change will happen in batches.

- *Case 1. Request for ontology extension*
    - The new proposed change needs to be justified in terms of possible usage scenarios
    - The proposed change is analysed in terms of what classes and properties can be used in order to formalise the required logic.
    - Proper term mappings are proposed both for proposed classes and properties if possible.
    - An ontology representation is prepared by the semantic engineers and a minor ontology release is published in the Github repository.
- *Case 2. Replacement of concept in core ontology*
    - If there are changes in the AIDAVA Dataset or in concepts from the source terminologies, this will result in replacement of the *owl:equivalentClass* mappings of the AIDAVA Reference Ontology class to the newly proposed concepts.
    - If the changes concern specific properties, then the *owl:equivalentProperty* mapping will be defined for the new concepts.
    - If changes concern property restrictions rules, SHACL rules or axioms, new rules/axioms definitions will be specified in the reference ontology.
- *Case 3. Structural changes in core ontology*
    - Implement the proposed new ontological structure (class, properties, axioms, SHACL validation rules)
    - An ontology representation is prepared by the semantic engineers and a minor ontology release is published in the Github repository.
- All changes need to be described with the related request in the ***AIDAVA Reference Ontology Request Log;*** each version of the AIDAVA Reference Ontology is documented in a README file in Github, acting as a ***AIDAVA Reference Ontology Release Notes.***

STEP 5. Manage impact on other components
This steps related to the updates required for the components impacted by the changes in the ontology (see below)

3.4.2.3 Supporting tool - Github

To manage changes into the AIDAVA Reference Ontology, we will use GitHUb where the ontology is stored, configured for this purpose.

More specifically we will use the following functionalities of Github

- **Issue Tracking** to support management of the ***AIDAVA Reference Ontology Request Log.*** GitHub's issue tracking system enables teams to create, assign, and track issues and tasks related to the project. It helps capture and manage governance-related requests as mentioned above including the following aspects
    - Specification of the request: why, what, urgency level
    - Results of the impact analysis and recommendations.
    - Information to the requestor on the result.
- **Collaboration and Communication:** GitHub offers features like comments, code discussions, and notifications, which foster collaboration and communication among project stakeholders. This is valuable for discussing and addressing governance-related concerns, documenting decisions, and ensuring transparency in the governance process.
- **Version Control:** GitHub provides robust version control capabilities, allowing teams to track changes, manage branches, and collaborate on code repositories. This is crucial for maintaining a controlled and auditable development process.
- **Code Documentation and Wiki:** GitHub provides space for documenting project-specific guidelines, governance policies, and best practices using the built-in wiki or README files. This facilitates the generation of the AIDAVA Reference Ontology Release Notes, dissemination of governance-related information and ensures that project members are aware of the established rules and processes.

3.4.2.4 Implementation of the change and management of releases

The above tools can be used to address irregular incoming requests for changes that can be distributed quite randomly in time. In order to organise the releases in a more timely manner, we plan to group the proposed changes into releases of the ontology.

To ensure a smooth and organised process, the planned process involves the following steps:

- Review and Finalise the Changes - assess all proposed changes to be made in the ontology. Conduct thorough testing and validation to verify the correctness and coherence of the ontology.
- Versioning - decide on an appropriate versioning increment (minor, major or patch) for each ontology version. We foresee usage of primarily major and probably minor versions of the ontology to be released.
    - A major release signifies significant changes to the ontology that may include major structural modifications, additions, or removal of concepts, and substantial updates to its functionality.
    - A minor release indicates smaller-scale which often include modification of properties and allowed values for certain classes, new relationships, improvements to existing concepts and mappings, or updates to the ontology's documentation.
    - Patch releases do not introduce new features or major changes but aim to provide fixes for specific issues (reported problems, bugs) in the ontology.
- Update Documentation - review and update the ontology's documentation (README files, change log etc)
- Create a Release Branch for the release (major, minor or patch) in scope.
- Make the Release Commit - merge the changes from development branch to the release branch and commit changes.

- Update Release Notes
- Publish the Release - Push the release branch to AIDAVA Reference Ontology GitHub repository.

3.4.2.4 Adapting the components constrained by the AIDAVA Reference Ontology

Changes in the AIDAVA Reference Ontology may impact the other components identified before: PHKGs, Catalogue of Data Sources, mappings to transform & publish the content of PHKG into specific target format and AI curation tools (e.g. NLP) trained with data aligned with the ontology.

Based on the changes introduced in the ontology, we can differentiate the following scenarios to these components

Component impacted	Lead actor	Extension (Incremental update)	Replacement of existing concepts	Remodelling with logical & structural change
		Introduce new classes and properties	*Existing property or class is replaced with another*	*Significant redesign of the logical elements*
"legacy" PHKG	Out of scope of AIDVA project (check SAB recommendations)	No impact on interoperability but on quality i.e. some data may not have been extracted and are missing 1. assess the need/value to extract missing concepts from historical records (and regenerate PHKGs) 2. If yes, decide on when to update PHKGs	Need to define equivalence between the obsolete class/property and the new ones. As the changes are replacement of mapped concepts as owl:equivalentClass or owl:equivalentProperty the PHKG can be easily re-generated.	May require significant updates - based on the type of change (expected out of scope of AIDAVA)
Catalogue of data sources	AIDAVA Site Administrator	Check if this new class/property must be included in the onboarding process	Check if this change has an impact in part of the onboarding process - and notify the changes if any	
Mapping rules for publishing	Semantic data engineer/ Ontologist	Check if this new/updated class/property impact the mapping - if yes update mapping		
AI/NLP tools	Out of scope of AIDVA project (check SAB recommendations)	Update of the annotation of the clinical narratives with the new concepts. Pretraining of the NLP tools for information extraction and text-based classification to cover the new concepts and relations.	If there is a matching rule between the previous and the new concepts - no changes are needed in the NLP tools. Otherwise, updates of the annotation of the clinical narratives will be needed including the new concept and its relations. Followed by fine-tuning of the NLP tools to cover the newly introduced concept.	Significant updates will be needed in the annotation of clinical narratives. Pretraining of the NLP tools will be needed as well.

The process to implement the changes during the project is described in the figure below

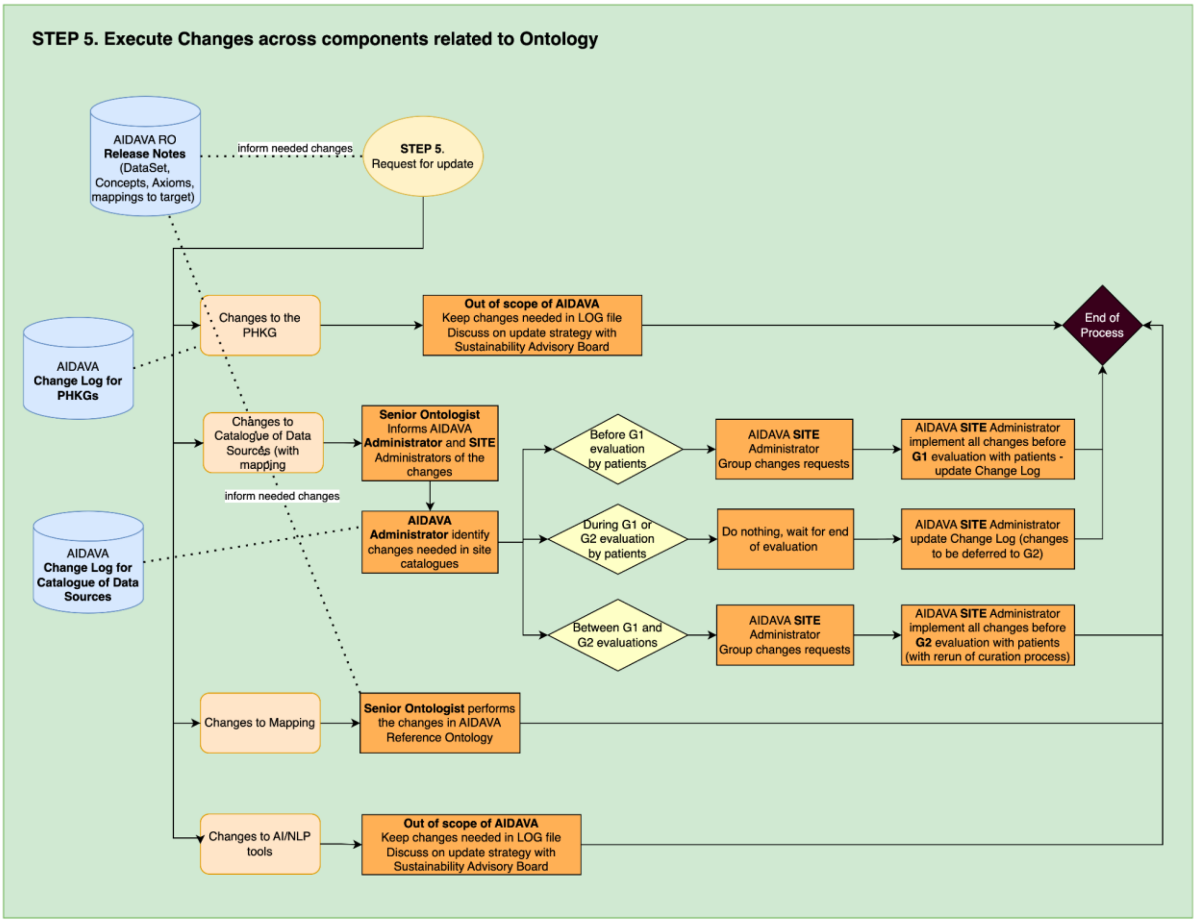

STEP 5. Execute Changes across components related to Ontology
AIDAVA RO Release Notes (DataSet, Concepts, Axioms, mappings to target)
inform needed changes
STEP 5. Request for update
AIDAVA Change Log for PHKGs
Changes to the PHKG
Out of scope of AIDAVA Keep changes needed in LOG file Discuss on update strategy with Sustainability Advisory Board
End of Process
Changes to Catalogue of Data Sources (with mapping
Senior Ontologist Informs AIDAVA Administrator and SITE Administrators of the changes
inform needed changes
AIDAVA Administrator identify changes needed in site catalogues
AIDAVA Change Log for Catalogue of Data Sources
Before G1 evaluation by patients
AIDAVA SITE Administrator Group changes requests
AIDAVA SITE Administrator implement all changes before G1 evaluation with patients - update Change Log
During G1 or G2 evaluation by patients
Do nothing, wait for end of evaluation
AIDAVA SITE Administrator update Change Log (changes to be deferred to G2)
Between G1 and G2 evaluations
AIDAVA SITE Administrator Group changes requests
AIDAVA SITE Administrator implement all changes before G2 evaluation with patients (with rerun of curation process)
Changes to Mapping
Senior Ontologist performs the changes in AIDAVA Reference Ontology
Changes to AI/NLP tools
Out of scope of AIDAVA Keep changes needed in LOG file Discuss on update strategy with Sustainability Advisory Board

Figure 12. Governance process for components impacted by changes of AIDAVA Reference Ontology

Patient PHKGs

The whole purpose of generating - and maintaining - patient PHKGs is to ensure interoperability; there cannot be semantic discrepancy across different PHKGs. To maintain all PHKGs into an interoperable - and reusable - status, it is critical to keep "legacy" PHKGs (created with previous releases of the ontology), compliant with the "current" PHKGs (created with the new release). However, if AIDAVA - or any tool based on the model developed in AIDAVA - is successful, with several millions of PHKGs across the EU, it would be difficult to upgrade all these PHKGs each time there is a new release of the ontology. It is therefore important to have a strategy to ensure that the "legacy" PHKGs generated a few years ago are still interoperable with the ones generated for other patients today; this could happen on demand (when a PHKG is needed) or all at once, in the context of a grouped migration, or through a combination of both. A key component for this would be the ease of migration.

While this is out of scope of AIDAVA, and of this deliverable, this is critical for the sustainability of the product beyond AIDAVA. It will be discussed with the Sustainability Advisory Group (SAB) and will be reported in Deliverables 6.8 and D6.13 with the recommendations of the SAB.

Catalogue of Data Sources

To ensure that new PHKGs generated from the same data sources in a site are compliant with the new release of the ontology, the Catalogue of Data Sources supporting the mapping of the data source with the AIDAVA Reference Ontology must be updated and the data source must be "onboarded" again. A copy of the former mapping must be kept for traceability purposes; this requires strict version control of each site catalogue.

- A first step is for the Senior Ontologist to inform the AIDAVA Administrator, responsible for the overall prototype, that there is a change.
- The AIDAVA Administrator, with the support of the Senior Ontologist can then identify the area of changes that need to be specified to the different AIDAVA Site Administrator and stores this in the Change Log for the Catalogue of Data Source
- Each site Site Administrator will then have to perform these changes, if this is possible i.e. before G1 evaluation and in between end of G1 evaluation and G2 evaluation. During the G1 evaluation changes are not permitted; they need to be stored in the Change Log as not done; they will have to be done after G1 evaluation

Pre-existing mapping for publishing data in target format

As the PHKGs are updated to the new release of the ontology, the mappings that have been defined for publishing data in a specific target format from the PHKG - e.g. in the form of CONSTRUCT SPARQL queries - must be updated by the semantic data engineer/ontologist. For traceability purposes, the former mapping must be stored with the related release of the ontology.

AI/NLP tools

Finally, any AI/NLP tools - trained with concepts related to a specific release of the ontology, may require retraining with the new release of the ontology. Specifically in cases where it is not possible to define a mapping between the former concepts and relations of the ontology to the new ones, it may be required to add and/or update to the annotation of text narratives used to train the NLP tool.

3.4.3 Maintaining and expanding the ontology beyond the project

As mentioned before, the main purpose of the AIDAVA Reference Ontology (RO) is to represent all concepts needed in an interoperable, personal longitudinal health record to support its sharing and reuse. The project is focusing on 2 therapeutic areas (Breast cancer and Cardiovascular disease). Assuming that the project demonstrates its value, it is critical to (I) promote wide acceptance and adoption of the existing ontology content and (II) expand the ontology across therapeutic areas and use cases. While these two sustainable activities are out of scope of this deliverable, it is important to introduce these approaches to sustainability and discuss them with the Sustainability Advisory Board (SAB), and for the tangible sustainable proposals to be reported in Deliverables 6.8 and D6.13.

3.4.3.1 Development of the Reference Ontology across Therapeutic areas and use cases

AIDAVA is a consortium demonstrating a new approach to the yet unsolved data interoperability problem in Healthcare. If the approach is successful and the project can demonstrate the benefit of maintaining PHKGs in compliance with a single Reference Ontology, this Ontology should be

I. formally endorsed and promoted for wide uptake;
II. further developed to cover a wider scope of, and potentially all, health domains.

The former requires collaboration - and eventually transfer of IP ownership - of the AIDAVA Ontology to an existing Standard Development Organisations (SDO).

As a first step we will reach out to the Swiss Public Health Network (SPHN) initiative, which is, to our knowledge, the first initiative maintaining a single ontology across healthcare at national level. This will allow us to understand how they intend to govern their ontology in the long run, and how they plan to align it at international level with similar initiatives.

In parallel we will discuss with the Sustainability Advisory Board (SAB) the most appropriate SDO to be approached.

- Unless advised otherwise by the SAB, MONDO will not be considered as a valid candidate. This is not a SDO and their approach consists in developing new concepts and relationships to provide a comprehensive framework for representing and mapping diseases across different terminologies; this is creating an additional layer. In AIDAVA, the principle is on reuse and developing mappings on top of existing terminologies already in use in healthcare (see concept of *owl: equivalentClass* and *owl: equivalentProperty*).
- SNOMED International may be a potential candidate. As mentioned in their annual report 2022 [37] *SNOMED International, as part of its 2020-2025 strategy, has committed to acting as a central hub, or terminology integrator, for healthcare terminologies, and to pursuing alliances and partnerships with other international standardization bodies to harmonize healthcare terminology across multiple domains*.
- HL7 is another potential SDO, through very focus on data exchange, rather than ontology
- ISO/TC251 Health Informatic could be another SDO, though they have limited focus on ontology so far.
- Finally we should consider the Joint Initiative Council [38] bringing together multiple SDO, including SNOMED, LOINC, HL7, ISO/TC251, CDISC, IHE, …. This might be of interest as the scope of the healthcare ontology should include clinical care as well as clinical research.

The latter requires an ongoing thread of R&D and a maintenance organisation that will act as the curator of new ontological inputs, oversee quality and governance of the complete ontology, and will liaise with the nominated SDO to provide periodic updates and enable revision of the published standardised version. This sustaining entity might be a SDO, an existing AIDAVA partner, or an external entity, in any case with ontology expertise. It is likely that a not for profit entity would be preferred. The key characteristics of this entity, its business model and how future R&D funding to grow and demonstrate new content will be explored with the SAB and reported in future WP6 deliverables.

3.4.3.2 Broaden acceptance and use

Most ontologies in Healthcare - and cross ontologies projects like MONDO - have been focused on secondary use, potentially with different ontologies for each use case.

The principle in AIDAVA is to keep a single ontology across all use cases for primary data use first - at patient record level - to support both primary use and secondary use of these data. Its value will be directly linked with its acceptance across healthcare.

Acceptance - at least in Europe - will be dependent upon integration within the European Electronic Health Record exchange format (EEHRxf), as the core standards for the emerging European Health Data Space (EHDS) . With the support of the SAB, and Partner IHD - who is a member of the project - we will approach the XpanDH Coordination and Support Action [39]. The project is financed by the European Commission, and is developing, experimenting and adopting the EEHRxF toward implementation of the EHDS.

4 Conclusions and Next steps

In this deliverable, we identified the requirements for the AIDAVA Reference Ontology and confirmed - in alignment with the recommendations for implementation of the EHDS and EEHRxF - that SNOMED CT, LOINC, HL7 FHIR General Purpose Data types and HL7 FHIR IPS profiles where the core components of this ontology, required to meet the use cases in scope of the project. A first release of the AIDAVA Reference Ontology was created and is stored in the AIDAVA Github platform, supporting version control and collaboration across the AIDAVA teams.

While developing the ontology, we realised that it required governance during and beyond the project. Governance during the project is key to maintain consistency across tasks that might identify new concepts - including Task 5.2 with data source onboarding and mapping, and Task 4.3 with narrative annotation. Assuming that the overall objectives of the project of (semi) automated data curation of personal health data toward resolution of data interoperability in healthcare is successfull, governance beyond the project will be critical to ensure sustainability of the results.

Implementation of the governance process and supporting tools during the project will be the focus on the coming months, ensuring that we keep the first version of the AIDAVA Reference Ontology up to the date for the duration of the project; discussion on governance after the project will take place through the foreseen meetings with the SAB.

5 References

[1] "Home –," *LOINC*. https://loinc.org/ (accessed May 14, 2023).
[2] "Home," *SNOMED International*. https://www.snomed.org/ (accessed May 29, 2023).
[3] "NeOn book." http://neon-project.org/nw/NeOn_Book.html (accessed May 18, 2023).
[4] M. Dean and D. L. McGuinness, "OWL web ontology language reference." https://www.w3.org/TR/owl-ref/ (accessed May 16, 2023).
[5] "OWL 2 web ontology language document overview (second edition)." https://www.w3.org/TR/owl2-overview/ (accessed May 16, 2023).
[6] "Project," *Tehdas*, Feb. 23, 2021. https://tehdas.eu/project/ (accessed May 14, 2023).
[7] M. Nurmi, "TEHDAS assesses data interoperability standards," *Tehdas*, Dec. 21, 2022. https://tehdas.eu/results/tehdas-assesses-data-interoperability-standards/ (accessed May 14, 2023).
[8] "Common assessment method for standards and specifications (CAMSS)," *Joinup*. https://joinup.ec.europa.eu/collection/common-assessment-method-standards-and-specifications-camss (accessed May 14, 2023).
[9] "Recommendation on a European Electronic Health Record exchange format," *Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/recommendation-european-electronic-health-record-exchange-format (accessed May 14, 2023).
[10] A. K. Lawrence, L. Selter, and U. Frey, "SPHN - The Swiss Personalized Health Network Initiative," *Stud. Health Technol. Inform.*, vol. 270, pp. 1156–1160, Jun. 2020.
[11] "Selecting terminologies and ontologies." https://faircookbook.elixir-europe.org/content/recipes/interoperability/selecting-ontologies.html (accessed May 14, 2023).
[12] EMA, "Big data," *European Medicines Agency*, Feb. 05, 2019. https://www.ema.europa.eu/en/about-us/how-we-work/big-data (accessed May 15, 2023).
[13] S. Köhler *et al.*, "The Human Phenotype Ontology in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1207–D1217, Jan. 2021.
[14] "Gene Ontology resource," *Gene Ontology Resource*. http://geneontology.org/ (accessed May 14, 2023).
[15] "Welcome to the NCBO BioPortal." https://bioportal.bioontology.org/ (accessed May 14, 2023).
[16] "[No title]." https://mondo.monarchinitiative.org/ (accessed May 14, 2023).
[17] WHOCC, "WHOCC - ATC/DDD Index." https://www.whocc.no/atc_ddd_index/ (accessed May 14, 2023).
[18] "International Classification of Primary Care, 2nd edition (ICPC-2)." https://www.who.int/standards/classifications/other-classifications/international-classification-of-primary-care (accessed May 14, 2023).
[19] M. Kapur, *Home*. Faber & Faber, 2012.
[20] "2023 EU4Health work programme," *Public Health*. https://health.ec.europa.eu/publications/2023-eu4health-work-programme_en (accessed May 15, 2023).
[21] "Ontologies." https://www.orphadata.com/ontologies/ (accessed May 14, 2023).
[22] "MedDRA." https://www.meddra.org/ (accessed May 14, 2023).
[23] Uppsala Monitoring Centre, "WHODrug Global." https://who-umc.org/whodrug/whodrug-global/ (accessed May 14, 2023).
[24] "Data standardization – OHDSI." https://www.ohdsi.org/data-standardization/ (accessed May 14, 2023).
[25] "EHDEN certifies 10 SMEs to map data to OMOP CDM." https://ohdsi.org/ehden-certification-aug2020/ (accessed May 13, 2023).
[26] "BRIDG." https://bridgmodel.nci.nih.gov/ (accessed May 14, 2023).

[27] EMA, “Data on medicines (ISO IDMP standards): Overview,” *European Medicines Agency*, Sep. 17, 2018. https://www.ema.europa.eu/en/human-regulatory/overview/data-medicines-iso-idmp-standards-overview (accessed May 14, 2023).

[28] “HL7 CDA® R2 implementation guide: Data provenance, release 1 - US realm.” https://www.hl7.org/implement/standards/product_brief.cfm?product_id=420 (accessed May 18, 2023).

[29] “Datatypes - FHIR v5.0.0.” https://www.hl7.org/fhir/datatypes.html (accessed May 14, 2023).

[30] “SDTM.” https://www.cdisc.org/standards/foundational/sdtm (accessed May 18, 2023).

[31] “[No title].” https://www.iso.org/obp/ui/ (accessed May 14, 2023).

[32] “FAIR principles,” *GO FAIR*, Nov. 23, 2017. https://www.go-fair.org/fair-principles/ (accessed May 14, 2023).

[33] “SPHN Dataset — SPHN Semantic Framework 1.0 documentation.” https://sphn-semantic-framework.readthedocs.io/en/latest/sphn_framework/sphndataset.html (accessed May 18, 2023).

[34] “RDF - semantic web standards.” https://www.w3.org/RDF/ (accessed May 14, 2023).

[35] “SHACLer — SPHN Semantic Framework 1.0 documentation.” https://sphn-semantic-framework.readthedocs.io/en/latest/sphn_framework/shacler.html (accessed May 18, 2023).

[36] “XpanDH project,” *XpanDH Project*, Apr. 20, 2023. https://xpandh-project.iscte-iul.pt/ (accessed May 23, 2023).

[37] “2022 marks a year of continued progress and growth for SNOMED International, SNOMED CT,” *SNOMED International*. https://www.snomed.org/news/2022-marks-a-year-of-continued-progress-and-growth-for-snomed-international%2C-snomed-ct (accessed May 29, 2023).

[38] “JIC - projects.” http://www.jointinitiativecouncil.org/registry/index.asp (accessed May 29, 2023).

[39] “XpanDH project,” *XpanDH Project*, Apr. 20, 2023. https://xpandh-project.iscte-iul.pt (accessed May 29, 2023).