Formal Ontology in Information Systems
C. Trojahn et al. (Eds.)

doi:10.3233/FAIA241305

Towards Semantic Interoperability Among Heterogeneous Cancer Data Models Using a Layered Modular Hyper-Ontology

Mirna EL GHOSH [a,1], Varvara KALOKYRI [b], Melanie SAMBRES [a], Morgan VATERKOWSKI [a], Catherine DUCLOS [c], Xavier TANNIER [a], Gianna TSAKOU [d], Manolis TSIKNAKIS [b], Christel DANIEL [a] and Ferdinand DHOMBRES [a]

[a] *Sorbonne Université, Inserm, Université Sorbonne Paris-Nord, LIMICS, Paris, France*
[b] *Institute of Computer Science, Foundation of Research and Technology Hellas, Heraklion, Greece*
[c] *Université Sorbonne Paris-Nord, Inserm, Sorbonne Université, LIMICS, Paris, France*
[d] *MAGGIOLI S.P.A., Research and Development Lab, Marousi, Greece*

Abstract. Semantic interoperability is a growing and challenging subject in the healthcare domain. It aims to ensure a coherent and unambiguous exchange, use, and reuse of health information among different systems and applications. In the context of the EUCAIM (Cancer Image Europe) project, semantic interoperability among various heterogeneous cancer image data models is required to support the communication, integration, and sharing of data in a standardized and structured way. For this purpose, hyper-ontology is developed as a common semantic meta-model that bridges the disparate imaging and clinical knowledge of the various repositories in EUCAIM and supports their integration. EUCAIM's hyper-ontology is also an application-based ontology targeted for federated semantic querying and image annotation. To facilitate the hyper-ontology building process and ensure the extensibility of the ontology model, an iterative hybrid well-founded approach that divides the ontology structure into layers and modules is established.

Keywords. Semantic Interoperability, Biomedical ontologies, Hyper-Ontology, Cancer image data, Heterogeneous data standards, Common Data Models

1. Introduction

"Rectal cancer", "Cancer of rectum", and "Rectal carcinoma" refer to the cancer type "Malignant tumor of rectum". "Colorectal cancer" and "CRC" are alternative labels for the disorder "Malignant neoplasm of colon and/or rectum". In healthcare, medical terms describing the same cases (e.g., disorders) may have different labels, including synonyms and acronyms. Interoperable health data representations would ideally bridge syntactically different but semantically equivalent expressions in different degrees of structure [2]. Additionally, information models (or Common Data Models) in healthcare are a way

[1] Corresponding Author: Mirna El Ghosh, mirna.el-ghosh@inserm.fr

to standardize data storage (OHDSI-OMOP[2]) or data exchange (HL7-FHIR[3]), but they don't have a shared conceptualization; thus, they create new divergences between identical concepts. For instance, PSA (Prostate Specific Antigen) lab test is viewed as a *Measurement* in OMOP and an *Observation* in FHIR. In this context, semantic interoperability supports medical systems in interpreting the shared meaning of medical terms unambiguously to avoid misunderstandings when describing the same cases using disparate terms/standards. Standardized terminologies and ontologies, agreed at the international level (e.g., SNOMED-CT[4], LOINC[5]), and aligned basically with the FAIR principles [1], support interoperability in healthcare, where each medical term is linked to a shared and controlled vocabulary. However, applying semantic interoperability is challenging due to the large amount of data and the diversity of terminologies/standards.

This study is part of the EUCAIM project[6], a joint effort by the AI for Health Imaging (AI4HI) network [11,12] and major European Research Infrastructures to build up a hybrid (distributed and centralized) infrastructure integrating major existing European Real World Data infrastructures on cancer images, including many types of cancer. EUCAIM aims to integrate and federate large volumes of cancer image data by establishing semantic interoperability among the diverse data provided by the AI4HI network and future data providers. In EUCAIM, clinical and biological data/metadata associated with images are standardized and structured using FHIR and OMOP. For imaging knowledge, the DICOM standard[7] is commonly used to represent image segmentation and studies. Facing this diversity, a common semantic meta-model is required for semantic interoperability among the repositories [11].

Although using/developing ontologies is a vital strategy for enhancing the interoperability of heterogeneous datasets [4], their integration becomes significantly more challenging when annotating with various ontological or terminological references [3,4]. In EUCAIM, we develop a hyper-ontology to ensure and maintain semantic interoperability among multiple independently developed systems used for collecting and describing cancer image data. The hyper-ontology will help, on the semantic level, organize and group common data representations under a common meta-model so that disparate and heterogeneous clinical and imaging data models can easily and unambiguously communicate and interoperate. Also, the meaning of essential medical data is preserved and exchanged in a standardized, consistent, and meaningful way. Therefore, the main challenge of hyper-ontology is to facilitate integration and interoperability among data stored and modeled using diverse clinical and imaging data models and associated terminologies. Hyper-ontology is also an application ontology that permits the exploration of data collections, federated querying, and image annotation. A hybrid approach is proposed to facilitate the hyper-ontology development and ensure its extensibility. Section 2 overviews EUCAIM data models. Sections 3 and 4 present the development process and results, respectively. Section 5 outlines the evaluation and validation. Finally, sections 6 and 7 discuss the related work and conclude the paper.

[2]https://www.ohdsi.org/data-standardization/
[3]https://www.hl7.org/fhir/
[4]https://www.snomed.org/
[5]https://loinc.org/
[6]https://cancerimage.eu/
[7]https://www.dicomstandard.org/

2. Cancer Image Data Models in EUCAIM

In EUCAIM, clinical and imaging data types are considered. All AI4HI projects (CHAIMELEON, ProCAncer-I, EuCanImage, INCISIVE, and PRIMAGE) adopt a common data model (CDM) (e.g., OMOP, FHIR) for the clinical data where standardized vocabularies (e.g., SNOMED-CT, LOINC) are used for harmonizing the data to be stored in the CDM [12]. Using CDM facilitates standardization, ensuring all data nodes adhere to the same structure and represent their information using standardized terminologies. In EUCAIM, OMOP and FHIR are used based on the AI4HI network [11]. For the OMOP approach, oncology extension [14] and imaging extension [15,16,17] are applied in CHAIMELEON and ProCAncer-I. Additionally, the DICOM standard for medical imaging is adopted to represent disparate types of imaging data, including information about imaging studies or image annotations/segmentations. Despite the commonalities, data models/standards and data minimization suffer from diversity [12].

1. Data models and standards: in the AI4HI network, data models and standards are diverse for representing clinical and imaging knowledge.
    - Clinical data: suffers from significant disparity where various data models and ontologies coexist, and the projects combine, among many others, SNOMED-CT with OMOP or FHIR CDMs. Also, multiple terminologies and ontologies are used to standardize and homogenize various fields. For instance, ProCAncer-I and CHAIMELEON both provide clinical data for prostate cancer and adopt OMOP as CDM. However, they differ in choosing terminologies or concepts to standardize the same fields (see Table 1).
    - Imaging data: the DICOM standard is commonly used, but in the ProCAncer-I project, RadLex [13], is also used to standardize and homogenize some imaging metadata (e.g., *Modality*, *Laterality*, *Anatomic Region*).
2. Data minimization: also suffers from variability as each project defines its required clinical/imaging variables differently. There is no standard or consensus on the amount and granularity of data required [12].
    - Clinical data: there is a diversity in specifying clinical variables among projects addressing the same cancer type. For instance, INCISIVE and EuCanImage both have clinical data for breast cancer and adopt FHIR as CDM. However, they differ regarding the amount/type of required data (see Table 1).
    - Imaging data: although the projects commonly adopt the DICOM standard and rely on its representation for imaging knowledge, imaging metadata deemed useful for extracting and being used differs. For instance, CHAIMELEON focuses on extracting imaging metadata from imaging studies (e.g., *image modality*, *body part examined*). In contrast, ProCAncer-I and INCISIVE, apart from the imaging studies, also focus on metadata of image segmentations (e.g., *segment label*, *annotated region*).

The variability and diversity in data models/standards and data minimization on the clinical and imaging levels show that a common semantic meta-model is required to maintain semantic interoperability and enable seamless integration and communication among disparate data models.

Field	Project	Terminology	Standard terms
Therapy	ProCAncer-I	SNOMED-CT	Intensity modulated radiation therapy
	CHAIMELEON	ICD10PCS	Beam Radiation of Prostate
Cancer Staging	ProCAncer-I	Cancer Modifier	AJCC/UICC 7th pathological M1a Category
	CHAIMELEON	NAACCR	pM1a
Diagnosis	INCISIVE	SNOMED-CT	Malignant neoplasm of breast
	EuCanImage	SNOMED-CT	Primary malignant neoplasm of female breast
Medical procedure	INCISIVE	SNOMED-CT	Chemotherapy
	EuCanImage	SNOMED-CT	Chemotherapy cycle
Medication	INCISIVE	-	-
	EuCanImage	RxNorm	Capecitabine
Follow-up	INCISIVE	SNOMED-CT	Treatment changed
	EuCanImage	-	-

Table 1. Examples of diversity of terminologies/standards and required data

3. Hyper-Ontology Development Process

EUCAIM's hyper-ontology is a common semantic meta-model that aims to integrate heterogeneous, disparate, and complex imaging and clinical/biological knowledge represented using various standards/terminologies. Inspired by the Neon methodology [5], and SABiO [7], we propose a hybrid systematic formally and semantically well-founded approach (Figure 1) to develop the hyper-ontology.

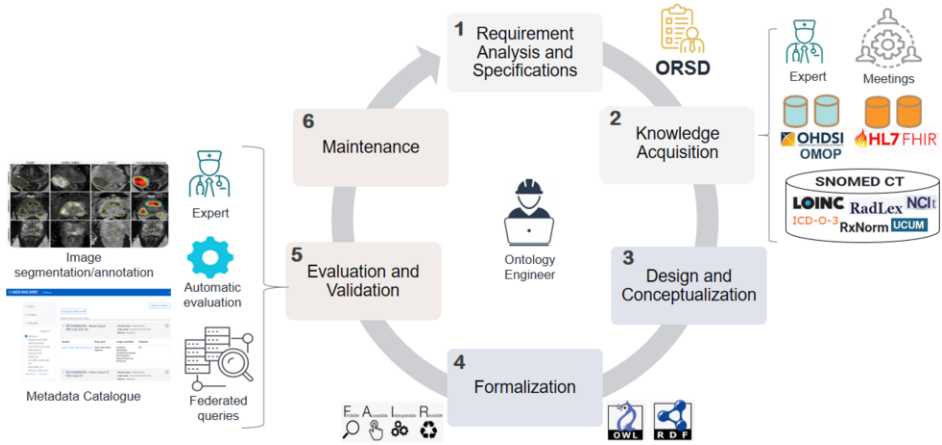


Figure 1. An illustration of the hyper-ontology development process

1- Requirements Analysis and Specifications states why the ontology is being built, its intended uses, who the end users are, and which requirements the ontology should fulfill [6]. After a set of meetings with users and experts from the EUCAIM community, we define the following:

- *Purpose*: EUCAIM's hyper-ontology aims to maintain and support semantic interoperability among diverse cancer image data models by providing the ontology-

based standard and structured vocabulary that integrates and combines these data models in a common meta-model. The hyper-ontology will permit applications to exchange queries and provide the semantic labels required for annotating cancer images. Besides, the integration with the OMOP and FHIR CDMs of the AI4HI network should be ensured, permitting consistent mapping with local nodes to seamlessly explore data collections through the Public Catalogue[8].

- *Scope*: the hyper-ontology covers the oncology domain, mainly cancer imaging data provided by the AI4HI network, including, among many others, the following cancer types: prostate, breast, rectum, lung, colon, colorectal, and liver.
- *Intended uses and users*: the hyper-ontology will support exploring the data collections through the Public Catalogue, Federated Querying (federated search of aggregated data in the collections), and semantic annotation of cancer images. Data users or researchers can use the hyper-ontology to explore or use data in research. We give an example of a simple query that uses terms from the hyper-ontology and will return information from the diverse data collections: *"age at diagnosis"* $\geq$ *50 & "diagnosis" = "prostate cancer" & "modality" = "MR"*.
- *Requirements*: non-functional (NFRs) and functional (FRs) are defined.

1. *NFRs*: such as NFR1- to support English; NFR2- the terminology to be used must be taken from validated biomedical ontologies and standardized terminologies; NFR3- the ontology model should be extensible to include future ontological aspects and new cancer types; NFR4- to support the FAIR principles; NFR5- to comply with the GDPR[9].
2. *FRs*: defined as *quality* requirements that describe the *goals for modeling* [9] and are stated as Competency Questions (CQs). Examples of FRs that the hyper-ontology should fulfill:

   * FR1- to define the different cancer types (e.g., *Breast cancer*, *Prostate cancer*), subtypes (e.g., *Primary malignant neoplasm of breast with axillary lymph node invasion*, *Hormone sensitive prostate cancer*), and their associated morphology (e.g., *Malignant neoplasm*, *Neoplasm, metastatic*);
   * FR2- to define the different body parts related to the identified cancer type/subtypes (e.g., *Breast*, *Axillary lymph node structure*, *Prostate*);
   * FR3- to define the diagnostic and therapeutic procedures, medication, and treatments used throughout and after the diagnosis or treatment of the different cancer types/subtypes (e.g., *Chemotherapy*, *Combined chemotherapy and radiation therapy*, *Androgen deprivation therapy*, *Surgical procedure*);
   * FR4- to define the imaging modalities, procedures, and results (e.g., *MRI*, *MRI of breast for screening for malignant neoplasm*, *MRI scan abnormal*);
   * FR5- to include the semantic labels and qualifier values required for image annotation and segmentation (e.g., *Malignant*, *Benign*, *Automatic*, *Manual*).
   * FR6- to define the semantic relations among the different entities, such as medication *treats* disorder, patient *has undergone* surgical procedure, etc.
   * FR7- to ensure the correspondence of the concepts to their domains/resources in OMOP and FHIR CDMs. For instance, *Primary malignant neoplasm*

[8] https://catalogue.eucaim.cancerimage.eu/

[9] https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en

of breast and *Chemotherapy* have *Condition* and *Procedure* as FHIR `resourceType` and OMOP `domain`, respectively.

* FR8- to represent specific concepts by combining atomic-related concepts, which aims to solve the disparity problem of provided data. For instance, the cancer staging metastasis value of M1 from *TNM (Tumor-node-metastasis)* category could be represented in two different ways: 1) *AJCC/UICC 7th pathological M1a Category* (Table 1, *Cancer Staging*/ProCAncer-I), which is a concept of the *Measurement* OMOP domain; 2) *TNM Path M*, a concept of the *Measurement* domain with value *pM1a* (Table 1, *Cancer Staging*/CHAIMELEON) of the *Meas Value* domain.

Following [5,6], the specifications and requirements are documented in the Ontology Requirements and Specifications Document (ORSD). Examples of knowledge-based CQs classified by cancer types are identified in the ORSD [31]. ORSD will have a key role during the hyper-ontology development process because it facilitates (1) the search and reuse of existing data resources, (2) the search and reuse of terminological/ontological resources, and (3) the verification and evaluation of the hyper-ontology.

***2-** Knowledge Acquisition* aims to collect and organize the knowledge the different projects provide.

1. *Collecting mandatory/required clinical and imaging knowledge*: provided as use cases defined per cancer type. Facing the diversity of knowledge discussed in Section 2, all the acquired knowledge is presented and organized in the ORSD using a set of Competency Questions. For clinical knowledge (following OMOP/FHIR), the standardized terms, their source terminologies/ontologies, and codes are collected. For imaging knowledge, we collect the values associated with DICOM tags and the standardized terms, if available.
2. *Reusing and merging ontological resources*: (scenario defined in Neon methodology [5]) includes establishing semantic and syntactic alignments and mappings to the acquired terms and their ontological resources. Two main mapping strategies are followed (see Figure 2):

   - *Label-based*: aligns the collected OMOP and FHIR standard concepts and DICOM imaging values/labels to existent validated terminological/ontological resources (e.g., SNOMED-CT, LOINC, NCIT, RadLex, ICDO-3) by applying an *exact match* similarity approach. This strategy will enrich the hyper-ontology with synonyms, codes, and definitions.
   - *Hierarchical-based*: extracts the taxonomies (*is-a*) of the OMOP and FHIR standard concepts and DICOM imaging values/labels from the standard terminologies/ontologies and aligns and merges them to construct the hierarchical structure of the hyper-ontology.

   The mappings are performed automatically using the following resources that combine many health and biomedical vocabularies and standards to enable interoperability between computer systems:

   - OHDSI ATHENA[10]: permits harvesting diverse mappings (*exact search* and *is-a*) from various standardized vocabularies.

[10] https://github.com/OHDSI/Athena

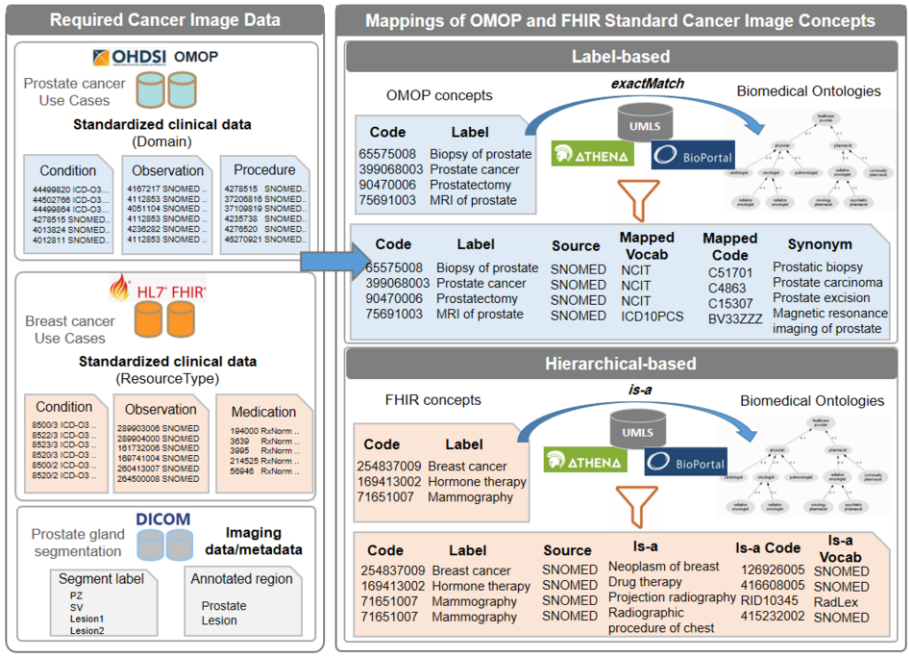


Figure 2. An illustration of *label-based* and *hierarchical-based* mappings

- UMLS (Unified Medical Language System) REST API[11] [19]: provides various endpoints to search and retrieve UMLS content (e.g., atoms, CUIs (Concept Unique Identifiers), definitions, semantic types, parents, children, etc.).
- BioPortal RESTful API[12]: comprises different resources (Ontologies, Classes) and related endpoints to access and browse biomedical contents. BioPortal helps enrich the hyper-ontology with synonyms, codes, and definitions and access the resources not considered in UMLS/ATHENA (e.g., RadLex, NCIT).

3. *Intervention of experts*: to specify application-related knowledge required in the hyper-ontology (e.g., query criteria), preferences in terminologies/ontologies (e.g., RadLex[13] for imaging), and to supervise mappings curation.

***3-** Design and Conceptualization* aims to design and structure the domain knowledge in a conceptual model that describes the problem and its solution regarding the domain vocabulary and mappings identified and harvested in the previous phases. To simplify this activity, we apply ontology *modularization* and *layering* as support processes:

- *Modularization*: this process is needed facing the domain complexity [7]. To simplify the hyper-ontology development, maintenance, and reuse, the content is split into separate generic content-based modules, as *sub-ontologies*, developed independently but connected using semantic relations.
  - ∗ *Clinical*: includes the clinical and biological knowledge, such as cancer types and subtypes, their associated morphology/histology, body parts, family his-

[11] https://documentation.uts.nlm.nih.gov/rest/home.html
[12] https://data.bioportal.lirmm.fr/documentation

tory, medical treatment, surgical/therapeutic procedures, staging/grading systems/methods, specimen, and tumor markers tests/results;

- *Imaging*: covers the essentials of the radiology domain focused on cancer. It defines the imaging knowledge and techniques required during all phases of cancer diagnosis, such as imaging modalities (MRI, PET, CT), procedures, assessment (PI-RADS, BI-RADS), and findings or observations.
- *Common*: includes patient demographics such as, *age at diagnosis*, *sex assigned at birth*, and *gender*. Besides, this module covers *qualifier values* required for image annotation/segmentation (e.g., *benign*, *malignant*, *automatic*, *manual*), tumor staging/grading (e.g., *pT0*, *pM1*, *cM2*), or lab test results, such as absence/presence findings (e.g., *positive*, *negative*).

- *Layering*: is decided to distinguish the domain-specific knowledge provided by the AI4HI network, which will be used for specific applications, and the domain/core knowledge, which will be defined to maintain the hyper-ontology semantic content, permitting its usability and reusability. Classifying ontologies regarding their *level of dependence* on a particular task or point of view is initially proposed by Guarino [8]. Inspired by Guarino's classification, the hyper-ontology structure is divided into four layers located at different granularity levels:

  - *Domain-Specific Layer* (`DSL`): located at the bottom level, this layer includes the domain-specific concepts defined/used in the cancer imaging domain such as *Hormone sensitive prostate cancer*, *Left anterior basal transition zone of prostate*, *pM1a*, *Malignant*, and *MRI guided biopsy of prostate*. These concepts are specified based on the clinical/biological and imaging provided knowledge and their mappings. `DSL` reflects the granularity level of the hyper-ontology.
  - *Domain Layer* (`DL`): defines the domain concepts of the oncology domain such as *Neoplasm of trunk*, *Region of prostate*, *Disease morphology qualifier*, and *MRI of prostate*. These concepts are specified and connected to `DSL` using a *bottom-up* approach considering the *is-a* mappings obtained in the knowledge acquisition phase.
  - *Core Layer* (`CL`): includes the core concepts of the oncology domain reused from conceptual or non-ontological resources (e.g., Minimal Common Oncology Data Elements (mCODE)[13][20]), aiming to ground the hyper-ontology in oncology. For instance, *Cancer Patient*, *Primary Cancer Condition*, *Histology/Morphology*, *Cancer-Related Surgical Procedure*, *Tumor Marker Test*, and *Cancer Stage* are generic entities defined in the conceptual model of mCODE. They must be ontologically analyzed, with the associated relations and properties, in the light of selected foundational ontology [7], ensuring their representation in a core ontological conceptual model that reflects the main ontological semantics of the oncology domain. In this regard, applying an *ontological unpacking* process, which refers to a *process of ontological analysis that reveals the ontological conceptual model*, is a prominent top-down approach that can effectively ensure the semantic interoperability according to FAIR principles [21].

[13] https://build.fhir.org/ig/HL7/fhir-mCODE-ig/

* *Upper Layer* (UL): anchors CL concepts to generic concepts (e.g., *Person*, *Disease*, *Procedure*, *Assessment*, *Treatment*, *Situation*, etc.). The grounding (CL) and anchoring (UL) processes require using foundational ontologies (e.g., UFO [24,25], BFO [23]) as semantic bridges between domain/core concepts, improving the semantic understanding of terms and supporting the mappings of concepts and interoperability [10].

Figure 3 depicts an example of the hyper-ontology structure with examples of concepts selected from the different layers and modules. An additional ontology module, *Correspondence OM*, is envisaged to permit a syntactic integration with OMOP/FHIR by explicitly specifying the correspondences of concepts with OMOP *domain* and FHIR *resourceType*.

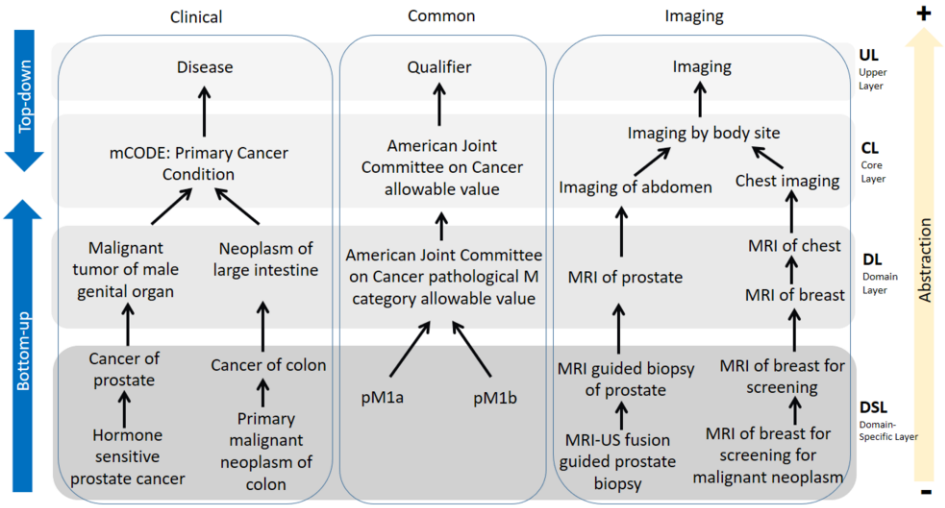


Figure 3. An excerpt of the hyper-ontology structure

Finally, semantic patterns are captured within the collected data to enrich the ontology model semantically. For instance, body locations are linked to cancers (e.g., *Prostate cancer* `SNOMED:Has finding site` some *Prostate*) and imaging procedures (e.g., *MRI of breast* `SNOMED:Has direct procedure site` some *Breast*).

***4-** Formalization* This activity generates a machine-readable and reusable formal version of the hyper-ontology supporting the FAIR principles.

***5-** Evaluation and Validation* The hyper-ontology needs to be evaluated with the help of clinical experts regarding the specified requirements. Also, assessing the integrity and performance of the hyper-ontology in achieving pertinent results in specific tasks of querying and segmentation is required. Besides, automatic methods are used to evaluate the hyper-ontology semantic content (see Section 5).

***6-** Maintenance* Based on the feedback from the evaluation and validation phases, the hyper-ontology will be revised to correct or enhance the ontological semantic/syntactic content.

4. Results

This study takes into consideration 12 use cases of cancer of the prostate, breast, colon, liver, and rectum, provided by the AI4HI network. The clinical/biological and imaging data and metadata provided by the projects are represented in the ORSD and classified per cancer type using a set of competency questions. The hyper-ontology is FAIR-compliant and accessible with the ORSD and the documentation on Zenodo [31]. The version 1.0 defines 2029 concepts, 5395 *is-a* mappings, 1939 *exactMatch* mappings, 2215 synonyms, 275 definitions, and 1755/353 syntactic alignments to OMOP/FHIR.
The hyper-ontology has fulfilled the specified functional requirements (see Section 3). For instance, the requirements related to the clinical/biological (FR1, FR2, and FR3) and imaging content (FR4) are fulfilled in the *Clinical* and *Imaging* modules, respectively, where the essentials of oncology and radiology focused on cancer are semantically represented and aligned to existent terminological/ontological resources (e.g., NCIT, RadLex). Also, the *Common* module covers the required knowledge for image segmentation and federated querying (FR5), including staging/grading values and various modifiers. The modules are interlinked using various semantic relations (n=74), fulfilling FR6. For instance, the object property *Treats* links medication to cancer conditions, *Has Answer* relates staging categories to staging values (Figure 5), and *Unit for* defines the unit measures for tumor marker test (Figure 4). The syntactic alignment to OMOP and FHIR (FR7) is established using semantic annotations (OMOP_Domain_ID, FHIR_ResourceType) and *Has correspondence* relation (see Figure 4). Finally, the hyper-ontology defines various equivalence semantic patterns to overcome the complexity, disparity, and heterogeneity of knowledge, especially for cancer staging, histological grades, tumor marker test results, and specific biological subtypes of cancer conditions (FR8). For instance, Figure 5 shows cancer staging semantic pattern defined using the *owl:equivalentClass* property, representing the specific concept *AJCC/UICC 7th pathological M1a Category* (Cancer Modifier) in the form of atomic concepts, *TNM Path M* (NAACCR:900) and *pM1a* (NAACCR:900@p1A).
Moreover, integrating and abstracting OMOP and FHIR concepts, which support semantic interoperability, are ensured in the upper level of the hyper-ontology using common categories reused from standard terminologies/ontologies. The process will be maintained at the core level by grounding these concepts in the mCODE ontological conceptual model (top-down strategy). We give some examples in what follows.

- Medication standard concepts (e.g., *Capecitabine* (RxNorm:194000)) having *Drug* as OMOP domain and *Medication* as FHIR resourceType are commonly specified as *Medication* (NCIT:C459). *Cancer-Related Medication Administration* mCODE core concept is envisaged to ground the medication in oncology.
- Tumor marker test concepts (e.g., *PSA* (SNOMEDCT:63476009)) having *Measurement* OMOP domain and *Observation* FHIR resource are specified as *Measurement* (SNOMEDCT:122869004, NCIT:C25209). *Tumor Marker Test* core concept will be reused from mCODE to ground lab tests in oncology.
- Cancer staging values (e.g., *pM1a* (NAACCR:900@p1A)), which are *Meas value* in OMOP and *Observation* in FHIR, are specified as qualifier finding values (SNOMEDCT:260245000). *Cancer Stage* and *TNM Stage Group* mCODE core concepts will be potentially reused to classify the staging categories and values for a semantically coherent oncology context.

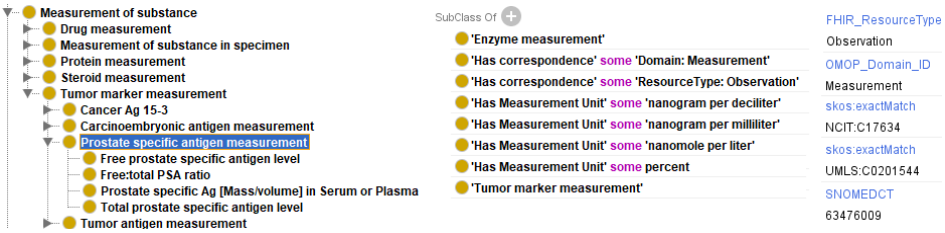


Figure 4. An excerpt of the hyper-ontology around *PSA* fulfilling FR6 and FR7

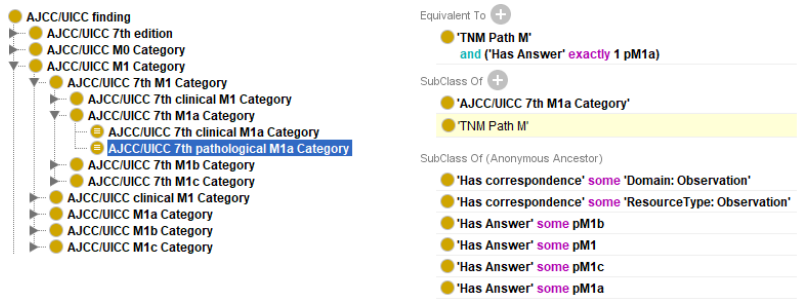


Figure 5. An example of representing specific concepts (TNM staging) fulfilling FR8

5. Evaluation and Validation

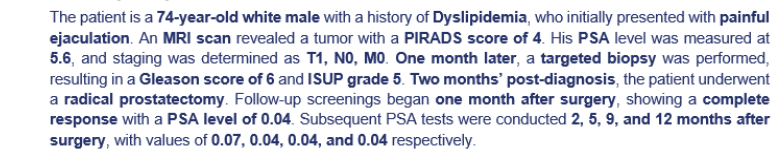
The hyper-ontology is validated as an RDF/OWL formal ontology [31], and its consistency is verified using Pellet[14], an OWL2 inference engine. Medical and ontological experts from EUCAIM's community will assess the hyper-ontology semantic content regarding the specified requirements. Moreover, the hyper-ontology will be evaluated according to its performance in data collection exploration through the Public Catalog[15], federated querying and processing, and cancer image segmentation/annotation tasks. Automatically, we tested two methods. First, to assess the hyper-ontology's correctness in answering specified requirements, generic competency questions (CQs) are translated to SPARQL queries, interrogating the hyper-ontology content regarding the answers given in the ORSD. For instance, (CQ1) *What are the main types and subtypes of breast cancer* (CLIN1000060)? (CQ2) *What body sites are affected by breast cancer*?

```
PREFIX ho: <https://cancerimage.eu/ontology/EUCAIM#>
Q1: SELECT ?p WHERE { ?p rdfs:subClassOf* ho:CLIN1000060. }
Q2: SELECT ?p  WHERE { ho:CLIN1000060  rdfs:subClassOf [ a owl:Restriction ;
owl:onProperty ho:HasFindingSite ; owl:someValuesFrom ?p ]. }
```

Second, to demonstrate the hyper-ontology's completeness in representing real-world scenarios, four use cases around prostate and breast cancers (see Figure 6 for an example) provided by AI4HI projects are considered. A set of individuals (manually) extracted

[14]https://github.com/stardog-union/pellet
[15]https://catalogue.eucaim.cancerimage.eu

from the scenarios, including demographic information (e.g., age, sex, race), symptoms, staging/grading values (e.g., T1, N0, ISUP grade 5), imaging/surgical procedures (e.g., MRI, biopsy, prostatectomy), and tumor marker test (e.g., PSA), is used to populate the ontology model. Semantic relationships are maintained among the individuals considering the scenarios and the object properties specified in the hyper-ontology, permitting the reasoner to deduce the complete diagnosis, including the imaging/clinical results [31].


Patient's journey

The patient is a **74-year-old white male** with a history of **Dyslipidemia**, who initially presented with **painful ejaculation**. An **MRI scan** revealed a tumor with a **PIRADS score of 4**. His **PSA** level was measured at **5.6**, and staging was determined as **T1, N0, M0**. **One month later**, a **targeted biopsy** was performed, resulting in a **Gleason score of 6** and **ISUP grade 5**. **Two months' post-diagnosis**, the patient underwent a **radical prostatectomy**. Follow-up screenings began **one month after surgery**, showing a **complete response** with a **PSA level of 0.04**. Subsequent PSA tests were conducted **2, 5, 9, and 12 months after surgery**, with values of **0.07, 0.04, 0.04, and 0.04** respectively.

Figure 6. A prostate cancer real-world scenario provided by the INCISIVE project

6. Related Work

Various ontologies have been proposed to support semantic interoperability in healthcare, such as Operational Ontology for Oncology (O3) [27,28] (previously OORO), and AIDAVA[16] Reference Ontology (RO) [29,30]. Also, well-founded conceptual models have been proposed, such as the Viral Conceptual Model (VCM) [21], which addresses conceptual uncertainty within the SARS-CoV-2 data and knowledge domain. For oncology, O3 is an expert-centered ontology relying on intervention and expertise from medical experts and stakeholders. RO and hyper-ontology consider health records extracted from heterogeneous data sources in a knowledge-centered approach. Nevertheless, expert intervention is considered in RO and hyper-ontology not only to evaluate the ontology model from the medical and semantic aspects but also to collaborate on some significant decisions regarding the requirements, the selection of terminologies/ontologies, and the curation of mappings. Table 2 presents the main specifications and strategies for building O3 and RO compared to the hyper-ontology.

Ontology	Scope	Approach	Strategy	Content	Alignment/Integration
O3	Prostate, breast, and head and neck cancers	Iterative-deliberations approach	Top-down	Common concepts for all cancer types, including imaging, pathology, medical oncology, surgery, and radiation treatment	SNOMEDCT, NCIT, mCODE, CodeX[17]
RO	Breast cancer registry and Cardiovascular Diseases score	Minimum set of requirements	Bottom-up	covers Observation, Condition, Procedure, Medication, Diagnostic Report (based on HL7 FHIR IPS profiles)	SNOMEDCT, LOINC, and HL7 FHIR ontological components
Hyper-Ontology	Cancer of prostate, breast, rectum, lung, colon, liver, colorectal	Hybrid iterative well-founded approach supported by ontology layering and modularization	Integration of bottom-up and top-down	Common concepts for all cancer types, including histology/morphology, staging/Grading, surgical, diagnostic and therapeutic procedures, tumor marker test, medication, imaging procedures, body parts, time patterns, and units of measure	SNOMEDCT, NCIT, RadLex, ICDO3, ICD10, LOINC, NAACCR, mCODE, foundational ontologies, and OMOP/FHIR

Table 2. A comparison of specifications and strategies for developing O3 and RO, and the hyper-ontology

[16] https://aidava.eu/

7. Conclusion

This study discussed the development process of EUCAIM's hyper-ontology to semantically bridge multiple independently developed systems used for collecting and describing cancer image data. Facing the diversity of healthcare standards (OMOP/FHIR) and the complexity and disparity of imaging and clinical/biological knowledge and associated terminologies, a hybrid systematic and well-founded approach, supported by ontology grounding, layering, and modularization processes, is proposed to simplify the building process. We demonstrated that the hyper-ontology supports semantic interoperability by providing a common terminology and set of concepts that can represent and integrate data from OMOP and FHIR, allowing users to formulate queries without needing to know the specifics of each underlying model. EUCAIM's hyper-ontology has also fulfilled specific requirements to resolve knowledge complexity. Semantic mappings with standardized vocabularies/ontologies and healthcare standards have ensured the reusability and compliance of hyper-ontology with FAIR principles. We also automatically evaluate the hyper-ontology using a dual approach, demonstrating its correctness in answering queries regarding the specifications/requirements and its completeness in representing real-world scenarios. Discussions with oncology/radiology experts, crucial for evolving and enriching the hyper-ontology, will continue. New use cases provided by the AI4HI network or new data providers will be considered. The top-down approach will also be finalized to support semantic mappings and interoperability. Finally, we will address significant challenges, such as extending the hyper-ontology with new cancer types and the evolution and sustainability of the hyper-ontology, mainly after the project completion. This project is co-funded by the European Union under Grant Agreement №101100633.

References

[1] Wilkinson MD, Dumontier M, Aalbersberg I, et al.. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018

[2] Schulz, S et al. Towards Principles of Ontology-Based Annotation of Clinical Narratives. International Conference on Biomedical Ontology (2023).

[3] Oliveira D, Pesquita C. Improving the interoperability of biomedical ontologies with compound alignments. J Biomed Semant 9, 1 (2018). doi: 10.1186/s13326-017-0171-8

[4] Smith B, Arabandi S, Brochhausen M, Calhoun M, Ciccarese P, Doyle S, Gibaud B, Goldberg I, Kahn CE, Overton J, Tomaszewski J, Gurcan M. Biomedical imaging ontologies: A survey and proposal for future work, Journal of Pathology Informatics, Volume 6, Issue 1, 2015, 37, ISSN 2153-3539. doi: 10.4103/2153-3539.159214.

[5] Suárez-Figueroa MC, Gómez-Pérez A, Fernández-López M. The NeOn Methodology for Ontology Engineering. In: Suárez-Figueroa M, Gómez-Pérez A, Motta E, Gangemi A, editors. Ontology Engineering in a Networked World. 2012. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-24794-1_2

[6] Suárez-Figueroa MC, Gómez-Pérez A. Ontology Requirements Specification. In: Suárez-Figueroa M, Gómez-Pérez A, Motta E, Gangemi A, editors. Ontology Engineering in a Networked World. 2012. Springer, Berlin, Heidelberg. p. 93–106, doi: 10.1007/978-3-642-24794-1_5

[7] Falbo R. SABiO: Systematic Approach for Building Ontologies. ONTO.COM/ODISE@FOIS. 2014.

[8] Guarino, N. Semantic matching: Formal ontological distinctions for information organization, extraction, and integration. Information Extraction A Multidisciplinary Approach to an Emerging Information Technology. 1997. Springer. p. 139–170, doi:10.1007/3-540-63438-X_8

[9] Guizzardi G, Proper HA. On Understanding the Value of Domain Modeling. Proceedings of 15th International Workshop on Value Modelling and Business Ontologies (VMBO 2021); 2021.

[10] Bernabé CH, Queralt-Rosinach N, Silva Souza VE et al. The use of foundational ontologies in biomedical research. J Biomed Semant 14, 21 (2023). doi: 10.1186/s13326-023-00300-z

[11] Kondylakis H, Kalokyri V, Sfakianakis S, Marias K, Tsiknakis M, Jimenez-Pastor A, et al.Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. Eur Radiol Exp. 2023 May 8;7(1):20. doi: 10.1186/s41747-023-00336-x. PMID: 37150779; PMCID: PMC10164664.

[12] Kondylakis, H., Ciarrocchi, E., Cerda-Alberich, L. et al. Position of the AI for Health Imaging (AI4HI) network on metadata models for imaging biobanks. Eur Radiol Exp 6, 29 (2022). doi: 10.1186/s41747-022-00281-1

[13] Chepelev LL, Kwan D, Kahn CE, Filice RW, Wang KC. Ontologies in the New Computational Age of Radiology: RadLex for Semantics and Interoperability in Imaging Workflows. Radiographics. 2023 Mar;43(3):e220098. doi: 10.1148/rg.220098. PMID: 36757882.

[14] Belenkaya R, Gurley MJ, Golozar A, et al.: Extending the OMOP common data model and standardized vocabularies to support observational cancer research. JCO Clin Cancer Inform 5:12-20, 2021. doi:10.1200/CCI.20.00079

[15] Kalokyri V, et al. MI-Common Data Model: Extending Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for Registering Medical Imaging Metadata and Subsequent Curation Processes. JCO Clin Cancer Inform. 2023 Sep;7:e2300101. doi: 10.1200/CCI.23.00101. PMID: 38061012; PMCID: PMC10715775.

[16] Park C, et al. Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. Yonsei Med J 63, S74 (2022).

[17] Park WY, Jeon K, Schmidt TS et al. Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension. J Digit Imaging. Inform. med. (2024). doi:10.1007/s10278-024-00982-6

[18] Osterman TJ, Terry M, and Miller RS. Improving cancer data interoperability: the promise of the Minimal Common Oncology Data Elements (mCODE) initiative. JCO Clinical Cancer Informatics 4 (2020): 993-1001.

[19] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.

[20] Osterman TJ, Terry M, Miller RS. Improving Cancer Data Interoperability: The Promise of the Minimal Common Oncology Data Elements (mCODE) Initiative. JCO Clinical Cancer Informatics. 2020; Vol 4:4. doi: 10.1200/CCI.20.0005.

[21] Bernasconi A, Guizzardi G, Pastor O, Storey VC. Semantic interoperability: ontological unpacking of a viral conceptual model. BMC Bioinformatics 2022, 23(Suppl 11):491.doi:10.1186/s12859-022-05022-0

[22] Smith B, Kumar A, Bittner T. Basic Formal Ontology for Bioinformatics; 2005

[23] Arp R, Smith B, and Spear A. Building ontologies with Basic Formal Ontology. 2015. Cambridge, MA: MIT Press

[24] Guizzardi G, Wagner G, Almeida JPA, Guizzardi RS. Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story. Appl Ontol. 2015;10(3-4):259–71.

[25] Guizzardi G, Botti Benevides A, Fonseca CM, Porello D, Almeida JPA, Prince Sales T. UFO: Unified Foundational Ontology. Appl Ontol. 2022;17(1):167–210. doi:10.3233/ao-210256.

[26] Lin, D. An Information-Theoretic Definition of Similarity. In: Jude W. Shavlik (eds.) Proceedings of the Fifteenth International Conference on Machine Learning, ICML'98, pp. 296–304. Morgan Kaufmann Publishers Inc. (1998). doi:10.5555/645527.657297.

[27] Mayo CS et al. Operational Ontology for Oncology (O3): A Professional Society-Based, Multistakeholder, Consensus-Driven Informatics Standard Supporting Clinical and Research Use of Real-World Data From Patients Treated for Cancer. Int J Radiat Oncol Biol Phys. 2023 Nov 1;117(3):533-550. doi: 10.1016/j.ijrobp.2023.05.033. Epub 2023 May 26. PMID: 37244628; PMCID: PMC10741247.

[28] Hong et al. Operational Ontology for Oncology: A Framework for Improved Communication and Understanding in Cancer Care. int J Radiat Oncol Biol Phys. 2023 Nov 1;117(3):551-553. doi:10.1016/j.ijrobp.2023.02.058

[29] De Zegher I, Çeleb, R, Powell L, Bihari B, Dallos D. D2.3 Solution Design Document for G1. 2023. Zenodo. doi: 10.5281/zenodo.10245773

[30] Primov T, Boytcheva S, Celebi R. D2.1 Global Data Sharing Standard. 2023. Zenodo. doi: 10.5281/zenodo.10117718

[31] Laboratory of Medical Informatics and Knowledge Engineering in e-Health (LIMICS). (2024). EUCAIM's Hyper-Ontology_V1.0. Zenodo. doi:10.5281/zenodo.12583826