Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240670

From Syntactic to Semantic Interoperability Using a Hyperontology in the Oncology Domain

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

[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] *APHP, GRC26 & Fetal Medicine Center, A. Trousseau Hospital, Paris, 75012, France*
[e] *MAGGIOLI S.P.A., Research and Development Lab, Marousi, Greece*
ORCiD ID: Mirna EL GHOSH https://orcid.org/0000-0001-6341-3847

Abstract. Interoperability is crucial to overcoming various challenges of data integration in the healthcare domain. While OMOP and FHIR data standards handle syntactic heterogeneity among heterogeneous data sources, ontologies support semantic interoperability to overcome the complexity and disparity of healthcare data. This study proposes an ontological approach in the context of the EUCAIM project to support semantic interoperability among distributed big data repositories that have applied heterogeneous cancer image data models using a semantically well-founded Hyperontology for the oncology domain.

Keywords. Oncology, Cancer image data, Heterogeneous data models, Syntactic interoperability, OMOP, FHIR, Semantic interoperability, Hyperontology

1. Introduction and Motivation

Interoperability in healthcare is essential and closely associated with health data integration [1]. Several forms, or levels, can be assigned to interoperability considering the heterogeneity in information systems [2]: 1) *structural* that considers heterogeneity involving data storage; 2) *syntactic* that addresses heterogeneity of data formatting/representation; and 3) *semantic* that handles heterogeneity of interpreting the meaning of data. Various international healthcare standards and vocabularies/ontologies are applied to satisfy syntactic and semantic interoperability [3]. The OHDSI-OMOP

[1] Corresponding Author: Mirna El Ghosh, LIMICS, Sorbonne Université, Univ. Sorbonne Paris-Nord, Paris, France; E-mail: mirna.el-ghosh@inserm.fr.

Common Data Model (CDM) [4] and HL7-FHIR [5] are, among others, syntactic standards which define the structure and syntax/format of data being stored and exchanged [6]. They mainly manage data's structural and syntactic heterogeneity by describing/defining the specific health information format to store and exchange. While syntactic interoperability is maintained in these standards by bringing health data stored in different formats using different database systems and information models into a common format or data model, semantic interoperability is supported by standardizing data based on common terminology standards (e.g., SNOMEDCT, LOINC). However, syntactic/semantic interoperability remains challenging, especially when integrating large heterogeneous datasets of existing health data repositories adopting different data standards. In this context, ontologies provide a flexible approach to integrating data and sharing meaning unambiguously [7], helping to overcome syntactic/semantic heterogeneity at certain levels of complexity. By ensuring a common understanding of information and making explicit domain assumptions, ontologies can address the challenges of accessing and querying heterogeneous healthcare data [8]. In the EUCAIM EU-Co-funded project[2], we employ an ontology integration approach toward semantic interoperability among heterogeneous cancer image data models and distributed big cancer data repositories. EUCAIM is a joint effort by the AI for Health Imaging (AI4HI) network [10] and major European Research Infrastructures to build up a hybrid (distributed and centralized) infrastructure integrating many major existing European Real World Data infrastructures (cancer images and accompanying clinical data), including many types of cancer. In the remainder, section 2 introduces the methods, the results are presented in section 3, and section 4 concludes the paper.

2. Methods

This work covers five main projects from the AI4HI network [10]: ProCancer-i, INCISIVE, Chaimeleon, EuCanImage, and PRIMAGE. Our approach aims to ensure the interoperability of cancer data at different levels as a consolidated framework that aggregates health data from multiple heterogeneous sources and integrates them into a common semantic meta-model (Figure 1).

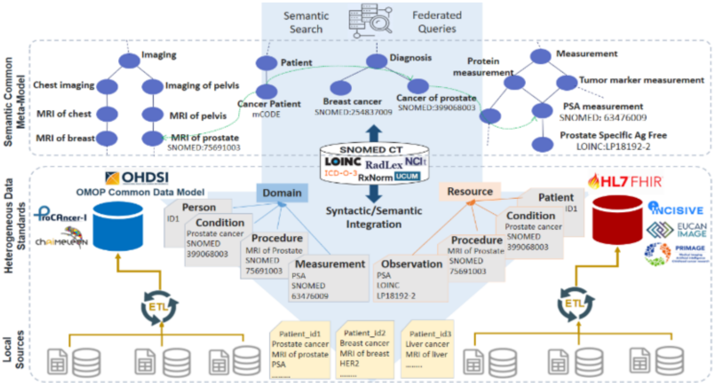


Figure 1. An illustration of the EUCAIM ontology integration approach.

[2] https://cancerimage.eu/, Accessed May 21, 2024

First, the local data sources have performed the standardization process by applying ETL (Extract, Transform, Load) processes to align with OMOP and FHIR data standards. Syntactic interoperability is expected at this level between the repositories following OMOP/FHIR. Once syntactic interoperability has been addressed to an acceptable extent, the Hyperontology integrates heterogeneous data, supporting semantic interoperability among them and the associated standard terminologies. The following subsections briefly describe the syntactic and semantic levels of interoperability.

2.1. Syntactic Interoperability: OMOP/FHIR

While in OMOP, a list of domains (e.g., *Condition*, *Measurement*, *Drug*, *Procedure,* etc.) is defined to which the concepts of the standardized vocabularies can belong, FHIR is built upon a set of resources (e.g., *Patient*, *Observation*, *Medication*, *Procedure*, *Condition,* etc.). Some extent of syntactic interoperability is expected among the resources following a common standard OMOP/FHIR. However, syntactic/semantic heterogeneity remains problematic due to data complexity and disparity. Two levels of syntactic heterogeneity are exposed among the projects adopting: 1) a common data model (e.g., OMOP) or 2) different data models (e.g., OMOP/FHIR). For instance,

1) ProCancer-i and Chaimeleon both adopt OMOP, but defined in two different ways the metastasis cancer staging values of M1 from the TNM (Tumor-node-metastasis) category as follows: 1) *AJCC/UICC 7th pathological M1a Category* which is a *Cancer Modifier* concept of the *Measurement* domain; 2) *TNM Path M*, a *NAACCR* concept of the *Measurement* domain with value *pM1a* of the *Meas Value* domain;
2) ProCancer-i (OMOP) and INCISIVE (FHIR), where the former defined PSA (*Prostate Specific Antigen*) as *SNOMED* concept (*Prostate specific antigen measurement (63476009)*) from the *Measurement* domain and the latter defined it as *LOINC* concept (*Prostate specific Ag [Mass/volume] in Serum or Plasma (LP18192-2)*) from the *Observation* resource (see Figure 1).

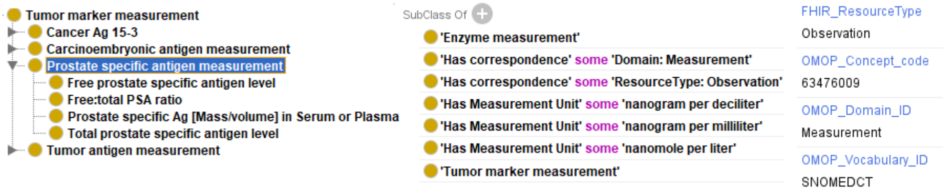
Facing this heterogeneity, formal ontologies can help to handle the automatic recognition and processing of such heterogeneous but *isosemantic* expressions [11].

2.2. Semantic Interoperability: Hyperontology

To ensure semantic interoperability among heterogeneous cancer data models, we propose a semantically and formally well-founded Hyperontology that captures the *real-world semantics* [2] of the oncology domain. To develop the Hyperontology, an iterative hybrid process is established, starting with the analysis of specifications and requirements and knowledge acquisition from the diverse AI4HI data sources and associated terminologies/ontologies in the medical domain (e.g., SNOMEDCT, LOINC, NCIT, ICDO-3). Furthermore, the conceptualization and formalization processes are performed using two different strategies: 1) *bottom-up* that maintains the hierarchical structure of the Hyperontology based on the standard concepts provided by the local sources and their syntactic/semantic mappings and 2) *top-down* that grounds the Hyperontology in foundational ontologies (e.g., UFO [12]) and core conceptual models (e.g., mCODE [13]). Therefore, the Hyperontology structure is divided into layers, facilitating the development process. The integration of these layers is established under the supervision of experts to maintain the Hyperontology structure and content.

3. Results

The formal model of the Hyperontology (beta version) [10], which is developed based on *12* use case studies in the oncology domain, defines *1852* concepts, *4172* subClassOf, and *18* semantic relations. The Hyperontology covers the cancer of breast, prostate, liver, colon, rectum, lung, and colorectal. To facilitate and support the Hyperontology development, an Ontology Requirements and Specification Document (ORSD) is produced regarding the requirements that the Hyperontology should fulfill. In the ORSD [10], the functional requirements are stated as competency questions (CQs) organized by cancer type. In the Hyperontology, the syntactic/semantic heterogeneity is addressed. As an example, Figure 2 depicts how semantically is resolved the PSA diversity of terminologies and OMOP domains (see Section 2.1) by specifying *Prostate specific Ag [Mass/volume] in Serum or Plasma* (*LOINC*) as a *subClassOf* of *Prostate specific antigen measurement* (*SNOMED*).



- Tumor marker measurement
  - Cancer Ag 15-3
  - Carcinoembryonic antigen measurement
  - Prostate specific antigen measurement
    - Free prostate specific antigen level
    - Free:total PSA ratio
    - Prostate specific Ag [Mass/volume] in Serum or Plasma
    - Total prostate specific antigen level
  - Tumor antigen measurement

SubClass Of

- 'Enzyme measurement'
- 'Has correspondence' some 'Domain: Measurement'
- 'Has correspondence' some 'ResourceType: Observation'
- 'Has Measurement Unit' some 'nanogram per deciliter'
- 'Has Measurement Unit' some 'nanogram per milliliter'
- 'Has Measurement Unit' some 'nanomole per liter'
- 'Tumor marker measurement'

FHIR_ResourceType
Observation

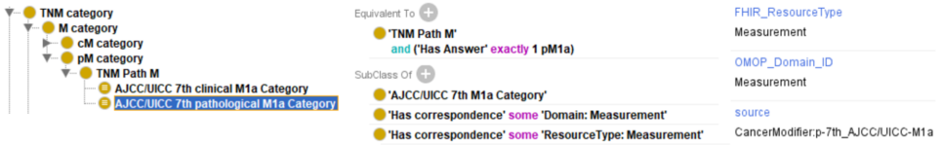
OMOP_Concept_code
63476009

OMOP_Domain_ID
Measurement

OMOP_Vocabulary_ID
SNOMEDCT

Figure 2. Excerpt of the Hyperontology (v0.2 beta) on resolving PSA heterogeneity (Protégé).

Figure 3 depicts an example of resolving the heterogeneity of representing the metastasis cancer staging values (see Section 2.1) by using *owl:equivalentProperty*.



- TNM category
  - M category
    - cM category
    - pM category
      - TNM Path M
        - AJCC/UICC 7th clinical M1a Category
        - AJCC/UICC 7th pathological M1a Category

Equivalent To

- 'TNM Path M'
  and ('Has Answer' exactly 1 pM1a)

SubClass Of

- 'AJCC/UICC 7th M1a Category'
- 'Has correspondence' some 'Domain: Measurement'
- 'Has correspondence' some 'ResourceType: Measurement'

FHIR_ResourceType
Measurement

OMOP_Domain_ID
Measurement

source
CancerModifier:p-7th_AJCC/UICC-M1a

Figure 3. Excerpt of the Hyperontology (v0.2 beta) on resolving TNM staging heterogeneity (Protégé).

The Hyperontology concepts are also syntactically integrated/aligned with their domains/resources in OMOP/FHIR (see Figures 2 and 3). The Hyperontology content is evaluated against the CQs defined in the ORSD with the help of domain experts and validated by embedding it in practical applications, such as semantic search (see examples from the EUCAIM Catalogue[3]). In the following, we give an example of a simple query that uses terms from the Hyperontology and will return information from the diverse local data sources: "*diagnosis*" = "*prostate cancer*" & "*PSA*" > "*20 ng/ml*" & "*modality*" = "*MR*".

4. Discussion and Conclusions

Various studies, such as [1], [14], [7], [15], and [16], have proposed ontologies for integrating heterogeneous healthcare data. In EUCAIM, data heterogeneity is exposed on two levels: syntactic and semantic. Although syntactic/semantic interoperability is

[3] https://catalogue.eucaim.cancerimage.eu/#/, Accessed May 21, 2024

ensured to a large extent by standardizing data using OMOP/FHIR, the complexity and disparity of big data remain challenging when integrating various cancer image data models. A consolidated ontological approach is proposed to semantically integrate heterogeneous data sources under a common semantic meta-model, aiming to resolve the disparity and complexity problems and permit accessing and querying data effectively from local repositories. In further work, we will proceed with the Hyperontology enrichment with the help of clinical experts. Besides, the Hyperontology will be validated using real-world applications for federated querying of data collections, cancer image annotation/segmentation, and federating analysis/learning in the EUCAIM context. Additionally, we are interested in exploring other approaches for enriching the Hyperontology, such as OMOP oncology extension [17]. This project is co-funded by the European Union under Grant Agreement №101100633.

References

[1] Peng C, Goswami P. Meaningful Integration of Data from Heterogeneous Health Services and Home Environment Based on Ontology. Sensors (Basel). 2019;19(8):1747. doi: 10.3390/s19081747

[2] Ouksel A, Amit S. Semantic interoperability in global information systems. ACM Sigmod Record. 1999;28(1):5-12.

[3] Lee A, Kim I, Lee E. Developing a Transnational Health Record Framework with Level-Specific Interoperability Guidelines Based on a Related Literature Review. Healthcare (Basel). 2021;9(1):67. doi: 10.3390/healthcare9010067

[4] Observational Medical Outcomes Partnership. [Accessed March 11, 2024]. Available from: https://ohdsi.github.io/CommonDataModel/.

[5] Fast Health Interoperability Resources. [Accessed March 11, 2024]. Available from: https://hl7.org/fhir/index.html.

[6] Umberfield E, Staes C, Morgan T, Grout R, Mamlin B, Dixon B. Chapter 9 - Syntactic interoperability and the role of syntactic standards in health information exchange. In: Health Information Exchange (Second Edition). Academic Press; 2023. p. 217-36.

[7] Liyanage H, Krause P, De Lusignan S. Using ontologies to improve semantic interoperability in health data. J Innov Health Inform. 2015;22(2):309-15.

[8] Kiourtis A, Nifakos S, Mavrogiorgou A, Kyriazis D. Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching. International Journal of Medical Informatics. 2019;132. doi: 10.1016/j.ijmedinf.2019.104002

[9] 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;7(1):20. doi: 10.1186/s41747-023-00336-x

[10] LIMICS, Eucaim's hyperontology v0.2beta, 2024. doi: 10.5281/zenodo.11109765

[11] Schulz S, Martínez-Costa C. How Ontologies Can Improve Semantic Interoperability in Health Care. In: Process Support and Knowledge Representation in Health Care. ProHealth KR4HC 2013. vol. 8268. Springer, Cham; 2013. p. 1-10.

[12] 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.

[13] 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.

[14] Zhang H, Guo Y, Li Qea. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC Med Inform Decis Mak. 2018;18 (Suppl 2). doi: 10.1186/s12911-018-0636-4

[15] Suárez P, Molina C, Prados, Peña C. On the Use of an Ontology to Improve the Interoperability and Accessibility of the Electronical Health Records (EHR). In: International Workshop on Semantic Interoperability (IWSI-2011); 2011. p. 73-81.

[16] Gonçalves B, Guizzardi G, Filho J. Using an ECG reference ontology for semantic interoperability of ECG data. Journal of Biomedical Informatics. 2011;44:1;126-36.

[17] 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