OMOP-CDM mapping to RDF/OWL: Attempting to bridge the OHDSI ecosystem and the Semantic Web world

Achilleas Chytas [1,2], Nick Bassiliades [2] and Pantelis Natsiavas [1]

[1] *Centre for Research and Technology Hellas| Institute of Applied Biosciences, 6th km Charilaou-Thermi 570 01, Thessaloniki, Greece*
[2] *Aristotle University of Thessaloniki | School of Informatics, Thessaloniki 541 24, Thessaloniki, Greece*

Abstract
Utilizing Real-World Data (RWD) for secondary use is still an open issue. Initiatives like OHDSI aim to tackle it by introducing a common data model (OMOP-CDM) to which data providers can opt to convert their data. While OMOP-CDM supports data interoperability and maintains a degree of intertwined terminologies/vocabularies, does not utilize the benefits of the Semantic Web technical paradigm. This paper presents an effort to convert the OMOP-CDM to RDF format to further enhance its linked data capabilities.

Keywords
OMOP-CDM, ETL, Semantic Web, Real-World Data

1. Introduction

OMOP-CDM (common data model) has been introduced and maintained by OHDSI aiming to support federated observational studies [1] and is used as a common reference to harmonize data from heterogeneous real-world healthcare data (RWD) sources, including electronic health records (EHRs), administrative/insurance claims, etc. A CDM can facilitate large-scale analyses and the use of distributed data without the need to share data, as healthcare (HC) data sharing is a legally, ethically, and technically complex process. OMOP-CDM consists of patient data (e.g., demographics, diagnosis, laboratory results, vital signs, etc.) but also interlinked vocabularies/terminologies, such as SNOMED-CT, WHO-ATC, and RxNorm, to ensure consistency and interoperability across different data sources.

Numerous international initiatives support the OHDSI distributed data network upon OMOP-CDM – EHDEN has been funding the conversion to OMOP-CDM of 187 data sources across Europe. Notably, OMOP-CDM is the main reference data model for the European Medicines Agency DARWIN infrastructure and has been used for many observational studies, including cohort studies, comparative effectiveness studies, etc across large datasets containing potentially millions of records.

Technically, OMOP-CDM is developed as a plain relational database model. It heavily relies on multiple hierarchical interconnected vocabularies and aims to support data interoperability, but it does not at all exploit the Semantic Web paradigm. While the Semantic Web stack could be used to provide a common language and standardized representation to support federated analysis of HC data, and even though ontologies and the RDF-based Knowledge Graphs (KGs) have been used to support HC data interoperability, still, the OMOP-CDM data model remains distant to the Semantic Web paradigm.

There have been attempts to use RDF-based knowledge structures to support activities related to the OHDSI ecosystem, e.g. LAERTES [2] a knowledge base using RDF, or an effort to map the OMOP-CDM vocabularies to RDF [3]. However, to the best of the authors' knowledge, there is no actively maintained full mapping of OMOP-CDM to RDF. This work presents an attempt to map OMOP-CDM to the RDF/OWL realm to bridge the gap between the world of OMOP-CDM and the Semantic Web ecosystem.

15th International SWAT4HCLS Conference, February 26-29, 2024, Leiden, The Netherlands
EMAIL: achytas@certh.gr (A. 1); nbassili@csd.auth.gr (A. 2); pnatsiavas@certh.gr (A.3)
ORCID: 0000-0001-8486-011X (A. 1); 0000-0001-6035-1038 (A. 2); 0000-0002-4061-9815 (A. 3)

2. Methodology

R2RML is a language for expressing customized mappings from relational databases to RDF datasets [4]. The R2RML mappings are RDF graphs in Turtle syntax and can be used to map the relational OMOP-CDM data tables and relevant RDF/OWL concepts.

MIMIC-IV (Medical Information Mart for Intensive Care IV) is a large, and available upon-request relational database that contains anonymized health data for over 40,000 Intensive Care Unit (ICU) patients [5] that is commonly used for exploring research questions and testing HC algorithms. This dataset has been converted to OMOP-CDM format [6] and it was used as the testbed dataset for the described data modelling conversion pipeline.

In general, each OMOP-CDM data table is mapped to a separate OWL class, while each table column corresponds to OWL properties:

1. **Object properties**: foreign keys from the initial source are mapped as object properties using a URI to link to a different individual
2. **Data Properties**: the majority of the numerical, string, date, etc fields from the initial source are mapped as Data Properties of the respective domain
3. **Annotation Properties**: fields that didn't fall in the previous categories and usually contain information like the initial Vocabulary that a term derived from, such as *ATC* or *MedDRA*

Regarding validation, a set of querying scripts was created to compare the source data (MIMIC-IV data in relational OMOP-CDM format) with the target data (MIMIC-IV data in OWL/RDF format).

3. Discussion

Semantic-based ontologies are indispensable in HC for their role in promoting interoperability, supporting clinical and policy decision-making, while advancing medical research. As the HC industry, both applied and research, continues to evolve and embrace digital transformation, the adoption of semantic technologies is vital for unlocking the full potential of the collected RWD that can lead to direct improvements to patient outcomes and enhance the overall efficiency of HC systems.

A seamless transformation of the OMOP-CDM to a semantically enriched format means that all those sources can be easily converted to a format that benefits from capabilities provided by semantic knowledge modelling such as the ease of integration with other diverse data sources such as genetic profiling, signalling pathways, drug biochemistry, could lead to the identification of latent relationships and patterns, elevating the usage of RWD to a higher level.

4. References

[1] OHDSI, *The Book of OHDSI: Observational Health Data Sciences and Informatics*. OHDSI, 2019. [Online]. Available: https://books.google.gr/books?id=JxpnzQEACAAJ

[2] Boyce RD, Voss EA, Huser V, et al. Large-scale adverse effects related to treatment evidence standardization (LAERTES): an open scalable system for linking pharmacovigilance evidence sources with clinical data. Journal of Biomedical Semantics. 2017;8(1):11. doi:10.1186/s13326-017-0115-3

[3] J. M. Banda, "Fully connecting the Observational Health Data Science and Informatics (OHDSI) initiative with the world of linked open data," *Genomics Inform*, vol. 17, no. 2, p. e13, Jun. 2019, doi: 10.5808/GI.2019.17.2.e13.

[4] Das, S., Sundara, S., & Cyganiak, R. (2012). R2rml: Rdb to rdf mapping language. W3c recommendation. World wide web consortium, 9.

[5] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.

[6] Kallfelz, M., Tsvetkova, A., Pollard, T., Kwong, M., Lipori, G., Huser, V., Osborn, J., Hao, S., & Williams, A. (2021). MIMIC-IV demo data in the OMOP Common Data Model (version 0.9). PhysioNet. https://doi.org/10.13026/p1f5-7x35