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

doi:10.3233/SHTI240674

Mapping OMOP-CDM to RDF: Bringing Real-World-Data to the Semantic Web Realm

Achilleas CHYTAS[a,b,1], Nick BASSILEIADES[b] and Pantelis NATSIAVAS[a]
[a] *Institute of Applied Biosciences, Centre for Research and Technology Hellas, Greece*
[b] *School of Informatics, Aristotle University of Thessaloniki, Greece*
ORCiD ID: Achilleas CHYTAS https://orcid.org/0000-0001-8486-011X
NICK BASSILEIADES https://orcid.org/0000-0001-6035-1038
Pantelis NATSIAVAS https://orcid.org/0000-0002-4061-9815

Abstract. Real-world data (RWD) (i.e., data from Electronic Healthcare Records – EHRs, ePrescription systems, patient registries, etc.) gain increasing attention as they could support observational studies on a large scale. OHDSI is one of the most prominent initiatives regarding the harmonization of RWD and the development of relevant tools via the use of a common data model, OMOP-CDM. OMOP-CDM is a crucial step towards syntactic and semantic data interoperability. Still, OMOP-CDM is based on a typical relational database format, and thus, the vision of a fully connected semantically enriched model is not fully realized. This work presents an open-source effort to map the OMOP-CDM model and the data it hosts, to an ontological model using RDF to support the FAIRness of RWD and their interlinking with Linked Open Data (LOD) towards the vision of the Semantic Web.

Keywords. Semantic Web, Real-World Data, OMOP-CDM, Knowledge Graphs

1. Introduction

Real-world data (RWD) - e.g. data from Electronic Healthcare Records – EHRs, ePrescription systems, patient registries, etc. - are underutilized for uses other than its primary [1]. This can be attributed to semantic and syntactic interoperability issues, the fragmentation across institutions, as well as legal and ethical barriers. Additionally, it is common for RWD to face limitations regarding data quality and completeness due to errors and missing information that reduces their reliability and further impede their utilization. Privacy and security issues, such as compliance with regulations like HIPAA and GDPR [2] impose additional limitations on their usage, hindering data sharing among healthcare organizations even within the same country/region.

To address these challenges, Observational Health Data Sciences and Informatics (OHDSI) [3] is a global community maintaining and building open-source tools upon the OMOP Common Data Model (CDM). OMOP-CDM aspires to play a pivotal role as a de-facto standard supporting the harmonization of RWD to facilitate multi-site observational studies. Besides providing a common data framework, OMOP-CDM also

[1] Corresponding Author: Achilleas Chytas; E-mail: achytas@certh.gr.

is enriched with a certain degree of semantic interoperability which is achieved via the use of a plethora of interlinked vocabularies and widely accepted terminologies that cover a wide spectrum of all health-related concepts (e.g., diseases, laboratory exams, drugs, etc.,) and even non-health-specific concepts (e.g. such as geographical location, ethnicities, etc.). However, the OMOP-CDM is built as a relational database schema and therefore it is not aligned with the Semantic Web vision as this was outlined by the use of relevant W3C recommendations which are based on the use of RDF[2].

There have been previous attempts to use OMOP-CDM and RDF together. For example, LAERTES employed an RDF Knowledge Base [4] and there has also been an effort to map the OMOP-CDM vocabularies to RDF [5]. Moreover, Jean-Baptiste Lamy et. al [6] presented previously a related work for mapping OMOP-CDM (v6.0) which is yet not widely adopted. In this paper, we present a pipeline for the conversion of OMOP-CDM v5.4 databases – the widely adopted version of OMOP-CDM.

2. Methodology

To demonstrate the feasibility of our approach, MIMIC-IV (Medical Information Mart for Intensive Care IV) database is used. MIMIC-IV is a large database with anonymized health data for over 40,000 Intensive Care Unit (ICU) patients [7]. We used a version of MIMIC-IV which has already been converted to OMOP-CDM format[3] [8].

R2RML is used to map the OMOP-CDM in RDF and convert relational data to a Knowledge Graph (KG). R2RML is a language for expressing customized mappings from relational databases to RDF datasets [9]. The R2RML-F[4] implementation mappings are shaped as RDF turtle graphs mapping the relational OMOP-CDM data tables to relevant RDF concepts (Figure 1). These mappings were used to convert the OMOP-CDM data to RDF. An R script was also built to fine-tune the results of the R2RML conversion process to create a proper and consistent instance of an RDF KG, i.e. to set proper namespaces, to set object properties, to set the domain and range of the object properties, etc.

Finally, in order to provide a quality assurance method to ensure proper conversion of OMOP-CDM data to an RDF KG, we built also a testing framework based on the rationale of OHDSI HADES framework[5]. HADES is a suite of R packages that provide a variety of functionalities varying from statistical analysis, cohort creation, machine learning algorithms, functions that facilitate data sharing, and also tools that act as quality assurance which can assist in the validation the contents of an OMOP-CDM database. Along this line of thinking, a set of functions in R was created to validate the data conversion in RDF regarding both completeness and accuracy. This testing framework validated the proposed transformation mechanism of MIMIC-IV data to RDF using via

---

[2] https://www.w3.org/RDF/

[3] https://github.com/OHDSI/MIMIC - It should be noted that this MIMIC-IV version in OMOP-CDM format required certain changes to be complied with the current v5.4 of OMO-CDM. The most significant changes had to do with the tables 'event' and 'event_episode' which have been created to better model patients suffering from certain diseases, such as cancer, that require a specific and occasionally long-term treatment that is usually formulated in therapeutic protocols.

[4] https://github.com/chrdebru/r2rml

[5] https://ohdsi.github.io/Hades/

the execution of queries upon both the OMOP-CDM dataset and the KG and afterward a comparison of the results. These queries can be summarized as follows:

1. Count the number of patient records (rows in relational OMOP-CDM format, patient individuals in the case of the KG) and compare them.
2. For each table/class get the IDs of the rows/individuals and compare them.
3. For each non-empty OMOP-CDM relational table get a number of random rows, also get the corresponding individuals from the KG and compare each database field to the corresponding property.

The 3rd step of comparing a random number of rows instead of using the table as a whole was preferred since in a real-world scenario tables might be too large for an exhaustive one-on-one comparison, although such functionality might be added on a future version. Along with the tests, a reporting mechanism was developed to inform the for the test results and identify relevant errors.

```
map:drug_exposure rr:logicalTable [rr:tableName "mimicredux.drug_exposure"];
 rr:predicateObjectMap [
  rr:objectMap [
   rr:template "{drug_exposure_id}";
   rr:termType rr:Literal
  ];
  rr:predicate rdfs:label
 ], [
  rr:objectMap [
   rr:column "dose_unit_source_value";
   rr:datatype xsd:string
  ];
  rr:predicate <drug_exposure#dose_unit_source_value>
 ], [
  rr:objectMap [rr:template "http://omop_cdm/concept/{drug_concept_id}"];
  rr:predicate <drug_exposure#drug_concept_id>
 ], [
  rr:objectMap [rr:template "http://omop_cdm/person/{person_id}"];
  rr:predicate <drug_exposure#person_id>
 ],
…
```

Figure 1. Snippet of the R2RML mapping script for the OMOP-CDM table “drug_exposure”

It should be noted that the conversion of an OMOP-CDM database to a functional RDF Knowledge Graph (KG) was performed in a neutral/‘naïve’ manner, meaning the authors did not create a sophisticated ontological schema but rather kept the OMOP data structure and converted it to RDF as is. OMOP-CDM is a generic schema and we argue that an ontological model in order to provide real value, it should be created based on a specific use-case/observational study scenario. However, such an attempt would inevitably restrict the generalization of the OMOP-CDM to RDF conversion pipeline. As such, it was decided to create a pipeline to facilitate a “plain” representation of OMOP-CDM data in RDF. The idea is that this model could be semantically enriched to support specific study scenarios via the use of relevant ontologies integrated when needed, on top of the presented model.

3. Results

The resulted RDF KG built using MIMIC-IV OMOP-CDM instance as the data source has over 6.4M axioms, 475K individuals, 22 classes and 62 object properties. Figure 2 presents the main classes of the ontological model and their interconnections.

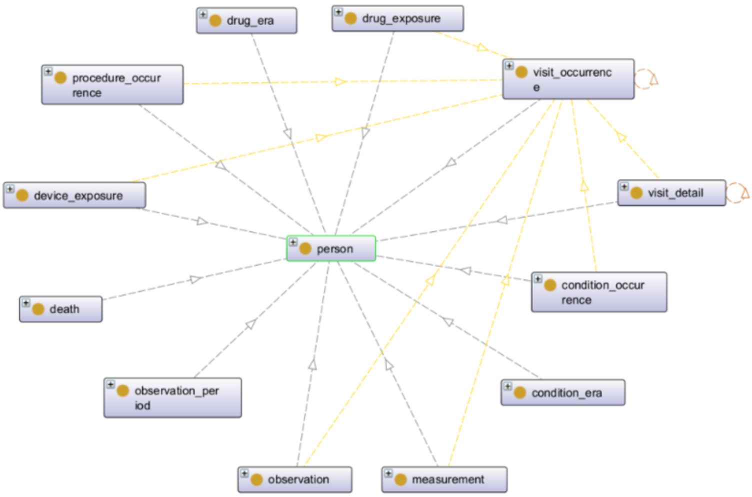

Figure 2. Important nodes of the ontological model

The first 2 tests resulted in a 100% compliance rate, while the 3rd test reported a few errors. Upon inspection of the reports, it was established that all errors on the testing dataset were due to data formatting. More specifically, the errors were located in OMOP-CDM columns that contained verbatim the information from the initial data source (MIMIC-IV in our case), typically named 'X_source_value' where X stands for drug, condition, device, etc., and these inconsistencies occurred since R2RML tried to 'tidy up' the data fields. The error categories that were identified:

- Space padding, i.e. 'R89.4 ' instead of 'R89.4'
- Strings that depict numeric values that during the conversion were transformed to numeric and back to a string, dropping zero padding in the process, i.e. '00160' to '160'.
- Encoding errors, i.e. for Greek or Chinese.

The R2RML mapping along with the quality assurance validation scripts are accessible online in a Github repository[6].

4. Discussion

In terms of potential limitations, we should mention that the MIMIC-IV relational OMOP-CDM dataset did not contain information in all the all tables or columns of the OMOP-CDM. This is a common occurrence as OMOP-CDM is an extensive model, and it is not unusual for a single data source to lack information for specific tables. This

[6] https://github.com/achillec/omop2ttl

means that although the entirety of OMOP-CDM was mapped using R2RML, only the tables contained data would be present in the final ontology as class, properties, and individuals and thus only these were tested in the context of this study.

Regarding future work, the first step is to use a dataset that contains data in all of the tables or use a synthetic dataset data in order to have a complete version of the RDF model. Then, other biomedical ontologies could be used to enrich the produced model and execute an observational study in order to evaluate potential benefits and challenges when using the RDF version of OMOP-CDM.

5. Conclusions

RWD modelled following the Semantic Web paradigm could in principle offer substantial benefits for healthcare research. By integrating RWD data with ontological models, interlinking with other knowledge structures could be facilitated and automatic reasoning could also be employed to support data processing. Converting OMOP-CDM to RDF via an open-source pipeline will enable a large amount of RWD data to benefit from the interlinking with the Linked Open Data realm.

References

[1] Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. JAMA. 2018 Sep 4;320(9):867–8.
[2] Voigt P, Von Dem Bussche A. The EU General Data Protection Regulation (GDPR) [Internet]. Cham: Springer International Publishing; 2017 [cited 2024 Mar 14]. Available from: http://link.springer.com/10.1007/978-3-319-57959-7
[3] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. MEDINFO 2015: eHealth-enabled Health. 2015;574–8.
[4] Boyce RD, Voss EA, Huser V, Evans L, Reich C, Duke JD, 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 Mar 7;8(1):11.
[5] Banda JM. Fully connecting the Observational Health Data Science and Informatics (OHDSI) initiative with the world of linked open data. Genomics Inform. 2019 Jun 11;17(2):e13.
[6] Jean-Baptiste L, Mouazer A, Sedki K, Tsopra R. Translating the Observational Medical Outcomes Partnership – Common Data Model (OMOP-CDM) Electronic Health Records to an OWL Ontology. In: Otero P, Scott P, Martin SZ, Huesing E, editors. Studies in Health Technology and Informatics [Internet]. IOS Press; 2022 [cited 2024 May 20]. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI220035
[7] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation [Internet]. 2000 Jun 13 [cited 2024 May 20];101(23). Available from: https://www.ahajournals.org/doi/10.1161/01.CIR.101.23.e215
[8] Kallfelz M, Tsvetkova A, Pollard T, Kwong M, Lipori G, Huser V, et al. MIMIC-IV demo data in the OMOP Common Data Model [Internet]. PhysioNet; [cited 2024 May 20]. Available from: https://physionet.org/content/mimic-iv-demo-omop/0.9/
[9] R2RML: RDB to RDF Mapping Language [Internet]. [cited 2023 Dec 8]. Available from: https://www.w3.org/TR/r2rml/