Demonstration of an Ontology-based Approach Used in the DATOS-CAT Project for Adopting Common Data Models

Santiago FRID [a,1], Guillem BRACONS CUCÓ [a,b]
[a] *Clinical Informatics Service. Hospital Clínic de Barcelona. 08036 – Barcelona. Spain.*
[b] *Institut de Bioenginyeria de Catalunya (IBEC). 08028 – Barcelona. Spain.*
ORCiD ID: Santiago Frid https://orcid.org/0000-0001-8400-5770
Guillem Bracons Cucó https://orcid.org/0000-0003-1274-7403

Abstract. Common Data Models (CDMs) are critical for data exchange and integration in the healthcare domain, although the conversion from local datasets is burdensome. OntoBridge is a scalable tool that facilitates this process by means of a set of ontologies that represent local databases and CDMs, as well as the syntactic mappings between them. It then converts data to RDF using Ontop, extracts it with SPARQL queries and post-processes the resulting CSV file with a Python script. This demonstration carries out the OntoBridge pipeline on a synthetic dataset based on the Hospital Clínic de Barcelona's data warehouse.

Keywords. Ontologies, common data model, OMOP CDM, i2b2, interoperability

1. Introduction

Data exchange in the realm of healthcare organizations is critical, especially for the progression of biomedical research, which relies on the consolidation of vast clinical datasets [1]. Yet, the process of making such data reusable entails considerable efforts in terms of cleaning, merging, and organizing [2], a task made more challenging by the diverse and often proprietary data formats employed by various institutions. This diversity poses a significant barrier to the reproducibility of research findings [3]. A growing trend to mitigate these challenges involves the adoption of Common Data Models (CDMs), which serve to unify data representation, thus enhancing the ability to share and integrate data from multiple sources while maintaining its semantic integrity and context [4-6].

Transforming data from local formats into those compliant with CDMs is typically a challenging task that requires substantial resources. The tools that are either available in the market or published in the literature [7-9] tend to focus on a single part of the ETL process, and are usually centered on a specific CDM.

This manuscript presents OntoBridge, a novel tool devised to simplify the task of converting data into CDM-compliant formats by means of ontologies, which have proven their utility in the biomedical field [2,11–13]. This tool is being used in the DATOS-CAT project, an initiative in Catalonia that employs OMOP for the integration of diverse clinical, phenotypic, genomic and environmental databases.

2. Topic

OntoBridge uses ontologies to represent both the source data and the CDM metamodel, while a third one performs the syntactic mapping between them using the rdf:type property. The semantic equivalence between local and standard concept is established by means of the owl:sameAs property.

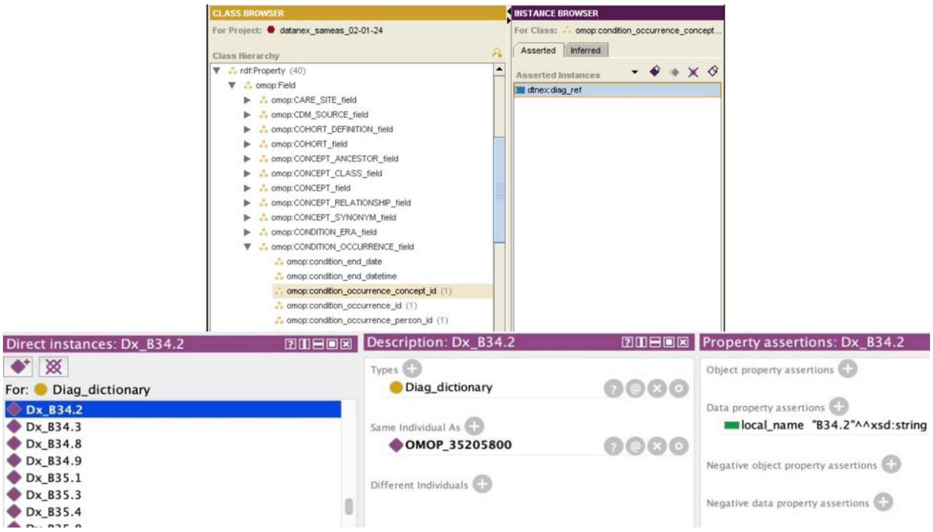
CLASS BROWSER
For Project: datanex_sameas_02-01-24
Class Hierarchy
rdf:Property (40)
omop:Field
omop:CARE_SITE_field
omop:CDM_SOURCE_field
omop:COHORT_DEFINITION_field
omop:COHORT_field
omop:CONCEPT_ANCESTOR_field
omop:CONCEPT_CLASS_field
omop:CONCEPT_field
omop:CONCEPT_RELATIONSHIP_field
omop:CONCEPT_SYNONYM_field
omop:CONDITION_ERA_field
omop:CONDITION_OCCURRENCE_field
omop:condition_end_date
omop:condition_end_datetime
omop:condition_occurrence_concept_id (1)
omop:condition_occurrence_id (1)
omop:condition_occurrence_person_id (1)
INSTANCE BROWSER
For Class: omop:condition_occurrence_concept...
Asserted
Inferred
Asserted Instances
dtnex:diag_ref
Direct instances: Dx_B34.2
For: Diag_dictionary
Dx_B34.2
Dx_B34.3
Dx_B34.8
Dx_B34.9
Dx_B35.1
Dx_B35.3
Dx_B35.4
Description: Dx_B34.2
Types
Diag_dictionary
Same Individual As
OMOP_35205800
Different Individuals
Property assertions: Dx_B34.2
Object property assertions
Data property assertions
local_name "B34.2"^^xsd:string
Negative object property assertions
Negative data property assertions

Figure 1: syntactic (upper part) and semantic (lower part) ontological mappings between local models and CDMs.

Data from relational databases is transformed to RDF using Ontop, an open-source tool leveraging R2RML for mapping to ontology elements. These OWL ontologies with RDF data are uploaded to a Jena Fuseki server for CDM-based SPARQL data extraction, followed by a Python script ensuring CDM compliance of the CSV output.

OntoBridge has proven effectiveness and flexibility in standardizing local datasets across three versions of the OMOP CDM (5.3, 5.4, 6.0) and i2b2, employing a scalable methodology adaptable to various input datasets and output models. Its ontological approach simplifies the management of updates and adjustments in response to changes in local data models and domain knowledge, which are frequent in the biomedical field. Furthermore, the ontologies of each CDM as well as the ontological mappings created between data sources and CDMs are completely reusable for new cohorts.

3. Contents of the demonstration

The demonstration will carry out the whole OntoBridge pipeline within a synthetic dataset of 100 patients mimicking the Hospital Clínic de Barcelona’s data warehouse. It will focus on the following steps:

- Creation of two ontologies: one that represents a data source stored in a relational database, and another one that maps it to the OMOP CDM.

- Exploration of an R2RML file that maps RDB to OWL elements.
- Real-time execution of Ontop to create RDF data from the RDB.
- Loading of the ontologies on a Jena Fuseki server and querying it using CDM-based SPARQL queries.
- Execution of a Python script to post-process the resulting dataset to obtain a final, CDM-compliant CSV file.

4. Brief CV of presenters

Santiago Frid, MD, MSc, PhD(c) is an endocrinologist and specialist in Health Informatics that serves as the head of the Projects section of the Clinical Informatics service at the Hospital Clínic de Barcelona. Guillem Bracons is a biomedical engineer with experience in health information systems. He is a member of the Clinical Informatics service at the Hospital Clínic de Barcelona.

References

[1] Evaluation of OMOP CDM, i2b2 and ICGC ARGO for supporting data harmonization in a breast cancer use case of a multicentric European AI project. J Biomed Inform. 2023 Nov 1;147:104505.

[2] Frid S, Pastor Duran X, Bracons Cucó G, Pedrera-Jiménez M, Serrano-Balazote P, Muñoz Carrero A, et al. An Ontology-Based Approach for Consolidating Patient Data Standardized With European Norm/International Organization for Standardization 13606 (EN/ISO 13606) Into Joint Observational Medical Outcomes Partnership (OMOP) Repositories: Description of a Methodology. JMIR Med Inform. 2023 Mar 8;11:e44547.

[3] Danese MD, Halperin M, Duryea J, Duryea R. The Generalized Data Model for clinical research. BMC Med Inform Decis Mak. 2019 Jun 24;19(1):1–13.

[4] Weeks J, Pardee R. Learning to Share Health Care Data: A Brief Timeline of Influential Common Data Models and Distributed Health Data Networks in U.S. Health Care Research. EGEMS (Wash DC). 2019 Mar 25;7(1):4.

[5] Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016 Dec 1;64:333–41.

[6] FitzHenry F, Resnic FS, Robbins SL, Denton J, Nookala L, Meeker D, et al. Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. Appl Clin Inform. 2015 Aug 26;6(3):536–47.

[7] Ontop: Answering SPARQL queries over relational databases [Internet]. Paperpile. [cited 2024 Feb 23]. Available from: https://paperpile.com/app/p/4346650f-169d-0d7c-aee6-7c79b88a1e41

[8] Wagholikar KB, Ainsworth L, Zelle D, Chaney K, Mendis M, Klann J, et al. I2b2-etl: Python application for importing electronic health data into the informatics for integrating biology and the bedside platform. Bioinformatics. 2022 Oct 14;38(20):4833–6.

[9] Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLoS One. 2019 Feb 19;14(2):e0212463.

[10] OMOP Converter: ETL Automation Software [Internet]. Paperpile. [cited 2024 Feb 23]. Available from: https://paperpile.com/app/p/1dad1f27-15ae-0fa2-83a5-c4f54c5dda7d

[11] Calvo-Cidoncha E, Camacho-Hernando C, Feu F, Pastor-Duran X, Codina-Jané C, Lozano-Rubí R. OntoPharma: ontology based clinical decision support system to reduce medication prescribing errors. BMC Med Inform Decis Mak. 2022 Sep 10;22(1):238.

[12] Frid S, Fuentes Expósito MA, Grau-Corral I, Amat-Fernandez C, Muñoz Mateu M, Pastor Duran X, et al. Successful Integration of EN/ISO 13606-Standardized Extracts From a Patient Mobile App Into an Electronic Health Record: Description of a Methodology. JMIR Med Inform. 2022 Oct 12;10(10):e40344.

[13] Lozano-Rubí R, Muñoz Carrero A, Serrano Balazote P, Pastor X. OntoCR: A CEN/ISO-13606 clinical repository based on ontologies. J Biomed Inform. 2016 Apr;60:224–33.