Development of an OMOP Ontology Application – PROSA – for creation and maintenance of highly granular source concepts within the OMOP vocabulary structure

Jared Houghtaling[a], Emma Gesquiere[a], and Lars Halvorsen[a]

[a] edenceHealth NV

Background:

Preserving the granularity of observational health data following a transformation to the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) is not always possible. This semantic limitation has prompted many data owners to create their own custom OMOP concepts with precise descriptors; integrating these concepts in the relational structure of the standard OMOP vocabularies, however, is nontrivial. In this work, we present our approach to creating an application – named *PROSA*, short for the Greek word for customize, *προσαρμόστε (prosarmóste)* – that can create, maintain, and deprecate source-specific concepts within the standard vocabulary tables. We describe the workings of the application in detail, discuss its benefits and shortcomings, and evaluate its future role and value among OMOP data owners.

Methods:

We developed *PROSA* as a Python application that orchestrates SQL queries across two distinct run modes (**Figure 1**). The application itself interfaces with a standard-format csv file containing source-to-standard mappings and requests for custom concepts. Any requests for custom concepts must have a desired *concept_name* specified, along with any standard parent and/or child concepts and other relationships. Note that we recently added support for handling relationships found within survey-based observational concepts that refer to items like question panels and answers, and we are continuing to add support for more complex relationship types (e.g. drugs) in order to interface effectively with local Athena instances **[1]**.

Once loaded, the application joins the mappings contained within this file against the standard vocabulary tables, and executes the following preliminary diagnostic checks: (1) any newly requested custom concept descriptions are matched (fuzzy string) against all existing standard concept descriptions and similarities above a defined threshold are flagged, (2) source-to-standard mappings and other relationships defined in the file are checked against the CONCEPT RELATIONSHIP table, (3) hierarchies defined in the file are checked against the CONCEPT ANCESTOR table, and (4) any internal references (e.g. custom concepts contained within hierarchy of other custom concepts) are parsed, verified, and ordered appropriately. These diagnostic checks generate a set of staging tables that contain all new items identified in the

mapping file, and importantly, exclude those concepts and relationships from the mapping file that already exist in the current vocabulary version.

Following the preliminary checks, the application either proceeds to producing a diagnostic export of information (Diagnostic Mode) or requests additional user confirmation to apply the updates identified to the core vocabulary version (Execution Mode). Importantly, the application handles the update and maintenance of all non-standard source codes and their associated mappings in the CONCEPT, CONCEPT RELATIONSHIP, and SOURCE TO CONCEPT MAP tables, as well as the creation and assignment of all standard custom codes and their associated linkages in the CONCEPT, CONCEPT RELATIONSHIP and CONCEPT ANCESTOR tables. We have chosen to handle the custom concepts as standard with *concept_id* values greater than 2B for two reasons: (1) descendant concept capture is a core and powerful functionality in cohort creation, and the OMOP vocabulary tables only include standard concepts in the CONCEPT_ANCESTOR table, and (2) relational constraints would otherwise preclude the definition of important relationships between non-standard source concepts, custom concepts, and existing standard concepts. In other words, treating custom concepts as standard makes it possible to utilize those concepts as if they already existed in the complex relational web of the standard vocabulary tables.

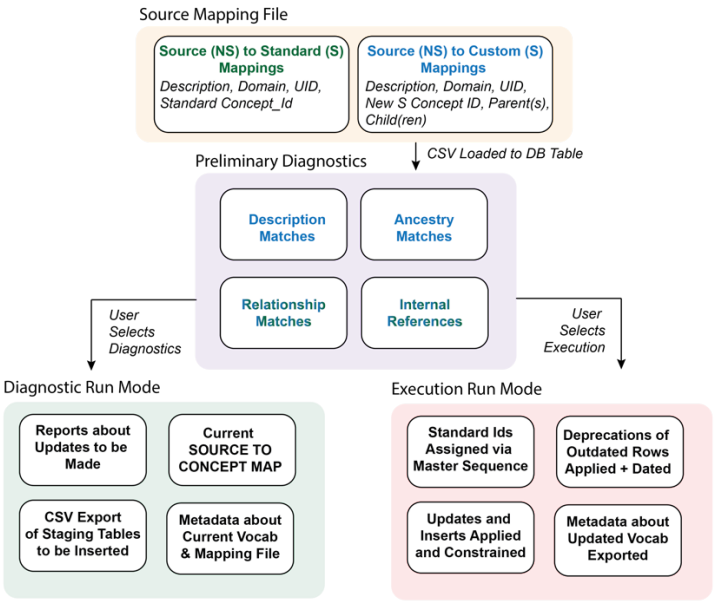

Figure 1. Overview of application process flow, showing load stage and handling of different mapping types; typical source-to-standard mappings shown in green font, and custom concept requests in blue font.

We designed the application to be deployed as a Docker container and executed against a PostgreSQL database containing the standard set of constrained and indexed OMOP vocabulary tables. Execution is currently command-line based, with configurations handled through the docker-compose utility, though we expect to provide a friendlier web interface in future iterations.

Results:

PROSA enables the automated creation and maintenance of OMOP concepts to capture site-specific concept granularity. Thus far, we have used it together with data owners to maintain more than 200’000 unique medical concepts (10M+ unique rows) across five different projects within the European Health Data and Evidence Network (EHDEN). We would like to emphasize that the tool is not designed or promoted as a permanent solution to filling gaps in the OMOP vocabulary; rather, it enables site-specific research using OHDSI tooling on highly granular source concepts in the interim period while those concepts can be formally integrated into the standard OMOP vocabulary through proper channels.

Conclusion:

The work presented above represents a compilation of efforts to solve a common challenge in OMOP harmonization processes. We hope that it may help others to design and implement comparable solutions, and we encourage any and all feedback related to the handling of vocabulary conventions. We expect to continue development on the application in the coming months and aim to share an open-source version shortly.

References:

[1] Houghtaling et al. “Construction of a central ontology platform for semantic mapping coordination and vocabulary augmentation across a multi-partner oncology consortium” OHDSI Europe 2023