Construction of a central ontology platform for semantic mapping coordination and vocabulary augmentation across a multi-partner oncology consortium

Jared Houghtaling[a], Peter Prinsen[b], Maaike van Swieten[b], Chiara Attanasio[b], Lars Halvorsen[a]

[a] *edenceHealth NV*
[b] *Netherlands Comprehensive Cancer Organisation (IKNL)*

Background:

The *Integraal Kankercentrum Nederland*, or IKNL, collects and catalogues data from individuals diagnosed with cancer across the Netherlands [1]. In an effort to enrich this national dataset, as well as to validate and expand upon ongoing analyses, IKNL has joined efforts with the Blueberry consortium alongside other regional and national cancer registries. Much of the critical information the consortium aims to capture exists as part of the ICDO3 vocabulary, which is not comprehensively represented in the standard Observational Medical Outcomes Partnership (OMOP) vocabulary tables as of spring 2023. In order to capture sufficient granularity of source concepts across the network, we have constructed a central ontology platform that provides Blueberry users with the following functionalities: (1) submission of unmapped source codes to a semantic mapping portal for automated standard suggestions (edenceMapper), (2) creation of semantic mapping sets to be collaboratively reviewed, and subsequently approved or modified, in an online web portal (edenceReviewer), (3) maintenance and update of the central consortium vocabulary tables using *PROSA*, an OMOP ontology application, (4) search and reference of the latest central consortium vocabulary version using a centrally deployed Athena web application, and (5) utilities and triggers that help integrate and synchronize each of the independent components. In this work, we describe each of these functionalities and the tools that support them, and we present the benefits and drawbacks of using such an approach to fill gaps (temporarily) in the OMOP vocabularies to support semantic mapping coordination across a targeted federated network.

Methods:

A general overview of the platform services and their interconnections is shown in **Figure 1** below:

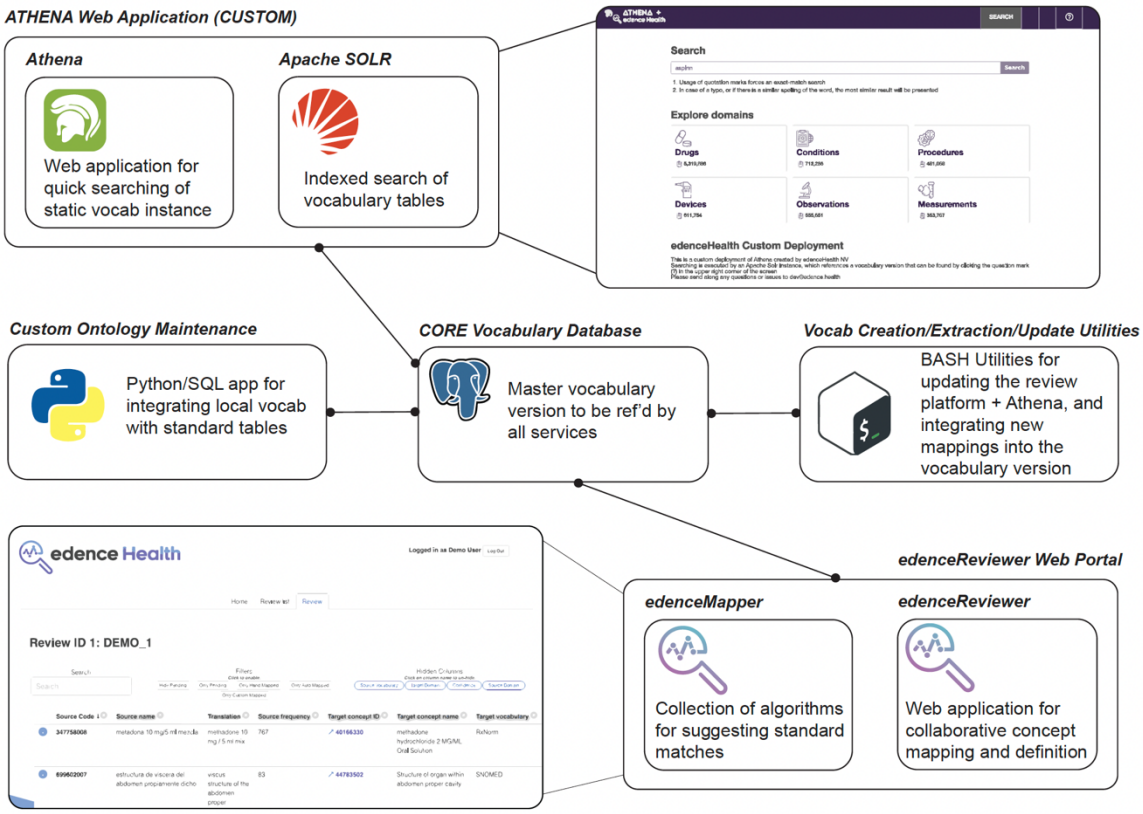

Figure 1: Overview of services available on the ontology platform, which include two web applications (Athena and edenceReviewer) as well as various tooling to support those applications and ensure that the core vocabulary tables are updated and maintained appropriately.

All components shown in **Figure 1** are deployed as Docker containers on a Linux-based virtual machine. The *Core Vocabulary Database Instance* is simply a Postgres (12) container hosting multiple databases – 6 in total – that support each of the services. Note that there is redundancy in standard vocabulary tables across the databases to enable concurrent read/write access. The Athena application requires three distinct databases: (1) an application database that handles user management and permissions, metadata, and vocabulary version tracking, (2) an OMOP CDM v4 version of the vocabularies (required to launch the server, but the vocabulary tables can remain empty), and (3) the core OMOP CDM v5 version of the vocabulary, along with two critical views based on the standard vocabulary tables. The edenceMapper and edenceReviewer applications share a database that supports the back-end processes necessary for collaborative semantic mapping in their web application, and the custom ontology application has both a staging and production database where it updates and modifies the core vocabulary tables referenced by the consortium. The *Custom Ontology/Maintenance Application (PROSA)* is a Python-based application that orchestrates database operations in order to make updates to the standard vocabulary tables (for more details, see **[2]**). The *Athena Web Application* requires two services – the java-based Athena server, and an Apache solr instance – for proper search functionality. Note that we have disabled the user login and vocabulary download functionality because this deployment is not available publicly, and vocabulary downloads are handled through a separate process. The *edenceMapper* and *edenceReviewer* services were built by

edenceHealth as an alternative to the *Usagi* mapping application that enable powerful and customizable suggestion algorithms and a collaborative user interface to review and correct suggestions.

Results:

In comparison with site-specific vocabulary maintenance, the ontology management structure we have established above has several advantages: all partners can implement the same standard vocabulary version in their OMOP CDM instances, all partners can reference the consortium-specific vocabulary through a familiar Athena interface, and all partners can quickly request and update the vocabulary version to incorporate concepts that capture source-specific information, without concern of concept redundancy across the network. One critical consideration with this type of approach is that data partners need to establish consistent ETL deployment schedules to reference a dynamic vocabulary version. We expect to keep track of source releases and referenced vocabulary versions across the network using the Ares network tool **[3, 4]**, which has already been implemented at IKNL to track OMOP version releases internally.

Conclusion:

Custom concepts are generally discouraged within the OHDSI community because they diminish the ability to collaborate on OMOP CDM studies within federated networks. The implementation we present here is not actively promoting the use of custom concepts. Rather, it is intended to be a temporary solution that enables OMOP consortia to pre-coordinate their semantic mapping approach and to maintain necessary granularity until it is possible to formally integrate those concepts into a new standard OMOP vocabulary version. We will coordinate with the OHDSI vocabulary team about best practices for implementing these sorts of ontology solutions at a broader scale, and we hope that the solution proposed can serve as inspiration for other consortia facing similar challenges.

References:

[1] https://iknl.nl

[2] Houghtaling et al. (2023) Development of an OMOP Ontology Application – PROSA – for creation and maintenance of highly granular source concepts within the OMOP vocabulary structure. OHDSI Europe Symposium 2023

[3] Ohdsi. (n.d.-a). *GitHub - OHDSI/Ares: A Research Exploration System*. GitHub. https://github.com/OHDSI/Ares

[4] Ohdsi. (n.d.). *GitHub - OHDSI/AresIndexer: R package that creates the index and relevant files for an Ares deployment.* GitHub. https://github.com/OHDSI/AresIndexer