SNOMED overhaul and its impact on ETL and phenotyping

Masha Khitrun[1], Alexander Davydov[1], Oleg Zhuk[1]

[1]Odysseus Data Services Inc., Cambridge, MA

Background: Over the years, the Vocabulary team has been working on integrating SNOMED CT into the ecosystem of OHDSI Standardized vocabularies. However, due to its comprehensive structure, multiple adjustments to SNOMED vocabulary ETL logic and interventions on the content level have been necessary, leading to the accumulation of bugs and discrepancies over the years. The SNOMED load_stage script that integrates the SNOMED into the OMOP vocabularies, has grown larger and more complex than anticipated, resulting in significant delays of OHDSI releases and a time lag between the OMOP version of SNOMED and SNOMED sources.

We present the results of a comprehensive overhaul of SNOMED in OHDSI vocabularies. This overhaul included both technical changes to the load_stage, aimed at simplifying future releases, and content changes designed to optimize cohort creation and ETL process.

Methods: The vocabulary development follows the guiding principles outlined in the Book of OHDSI[1], ensuring adherence to established standards and practices within the OHDSI framework. Developer documentation is maintained and made publicly accessible on GitHub[2], allowing for transparency and collaboration within the community.

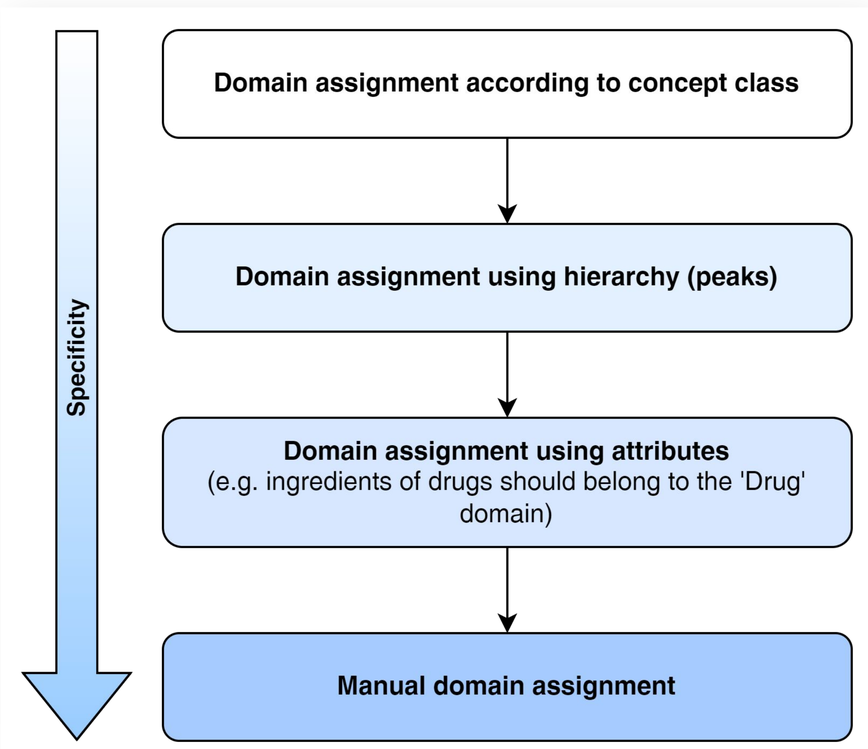
To assess the impact of vocabulary changes on ETL processes, we conducted a comprehensive analysis leveraging completed and ongoing ETL projects. This analysis provided insights into the challenges posed by vocabulary modifications.

Recognizing the significant impact of vocabulary changes on ETL[3], we employ analytical methods tailored to mitigate these challenges. These methods[4], accompanied by our internal quality control approach which includes the collection of vocabulary statistics, and a bunch of specific vocabulary checks enable us to address the implications of vocabulary updates on ETL workflows proactively.

Results:

- **Domain assignment was improved** in its stability and consistency (Figure 1). As a result of this change, you may need to change the tables of interest (eg. querying **Condition_occurence instead of Observation**) in the process of cohort creation.
- We have moved the **pre-coordinated SNOMED measurements** from the Condition to the **Measurement** domain and split them **in a post-coordinated way**. Thus, if you have previously used pre-coordinated SNOMED measurements, you should now start looking into the **Measurement** table where these concepts live as the Measurement / Value pairs or pre-coordinated concepts.
- **Secondary neoplasms** were mapped to Standard concepts in **Cancer Modifier** vocabulary that belong to the **Measurement** domain. Thus, now you should look into the **Measurement** table to find the Secondary neoplasm concepts.
- We improved the **creation of 'Maps to'** relationships following the SNOMED sources' **replacement links**. As a result of this change, more concepts are now mapped to Standard, and the **number of events in cohorts may increase**.

Figure 1. Domain assignment algorithm in SNOMED



- SNOMED concepts in the **Race** and **Provider** domains were **mapped to Standard** concepts in the respective domains.
- Concepts that belong to **Attribute, Location** (except countries), **Social Context** (except concepts that carry the semantics of relatives, religion, occupation), **Physical Force**, and **Physical Object** (except concepts in the Device domain) concept classes have been de-standardized in the course of the overhaul.
- We have performed the **retirement of the UK Drug Extension** module aimed to declutter the vocabulary, with concepts deprecated and linked to their equivalents in the dm+d vocabulary. On the other hand, **dm+d concepts in the Device domain**, previously mapped to the SNOMED UK Drug Extension Module, **have become standard**. **Gemscript Devices**, previously mapped to SNOMED, **have been remapped to dm+d**. UK Drug Extension concepts in the **Route domain have been de-standardized and mapped to standard Routes** that belong to other SNOMED Modules.

Conclusion:

The overhaul of the SNOMED vocabulary in OMOP has yielded significant improvements in ontology structure, cohort creation, and mapping efficiency. These enhancements contribute to more accurate data analysis, better research outcomes, and increased interoperability within the healthcare ecosystem.

However, it is essential to consider their impact on ETL processes and phenotyping algorithms. The adjustments made to domain assignments and concept mappings may require updates to existing ETL workflows, and researchers should carefully review their phenotyping algorithms to ensure compatibility with the updated vocabulary structure and content.

References:

1. Observational Health Data Sciences and Informatics. The Book of OHDSI.
2. https://github.com/OHDSI/Vocabulary-v5.0/tree/master/SNOMED
3. https://forums.ohdsi.org/t/cpt-hierarchy-errors-lost-children-in-2023-and-changed-domains/18383
4. Dmitry Dymshyts, Frank DeFalco, Anthony Molinaro, Clair Blacketer. An Evaluation and maintenance of cohorts and concept sets in the OMOP Vocabulary Evolution. July 2023, Conference: OHDSI European Symposium 2023.