Custom vocabulary techniques in ETL to OMOP CDM

An overview of aspects to consider when creating and maintaining a custom vocabulary

Table 1: Comparison of STCM and C/CR

<table>
<thead>
<tr>
<th>Parameter</th>
<th>STCM</th>
<th>C/CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>General description</td>
<td>Creation of a mapping from source_code to a standard_concept_id</td>
<td>Creation of a custom concept with concept_id (Standard to non-Standard) and mapping it to a standard_concept_id</td>
</tr>
<tr>
<td>Loading</td>
<td>Easy: table is designed for insertion of records</td>
<td>Requires alteration of vocabulary tables from Athena</td>
</tr>
<tr>
<td>Destiny of unmapped codes</td>
<td>Live only as source_codes</td>
<td>Live as custom concepts without mapping, i.e. as event_source_concept_id and can be queried by Atlas</td>
</tr>
<tr>
<td>Maps to value relationships</td>
<td>Does not support</td>
<td>Supports</td>
</tr>
<tr>
<td>ETL logic</td>
<td>Requires additional querying of the STCM</td>
<td>Allows for a usual lookup of a concept in the Concept table</td>
</tr>
<tr>
<td>Research use</td>
<td>None: data is not visible in Atlas or the Concept table</td>
<td>Custom concepts can be used in network studies through Atlas</td>
</tr>
<tr>
<td>Conclusion</td>
<td>Easy to implement, but not sharable</td>
<td>More complex, but sharable and usable in network research</td>
</tr>
</tbody>
</table>

Methods: The results are based on our experience of converting 34 unique datasets and their variations with all types of US- and ex-US data sources: EHR (e.g. Flatiron, Epic), claims (e.g. JMDC, CPRD-family), registry (e.g. NAACCR, disease- or study-specific), SDTM, etc. In our experience, the number of newly created custom vocabularies ranged between 4 to 57 per dataset (with a median of 16).

Results: When starting work on a vocabulary and choosing entities to define source_codes, these are some of the most crucial points:

- Domain integrity among codes in the same vocabulary
- Creating codes as event-value pairs (i.e. post-coordinate) vs defined facts (i.e. pre-coordinate) (Figure 1)
- What entities would enrich the definition (units, dosages/concentrations, attributes)
- What are the desired implications in the phenotype construction process?

When the vocabulary is compiled, QA is the next step consisting of three parts:

- Structural integrity checks, e.g.:
  - Are concept codes unique within the custom vocabulary?
  - Is the structure compliant with DDL and the content is referenced to the concept table?

- Semantic integrity checks, e.g.:
  - Are mappings to Standard concepts correct?

- Content integrity checks, e.g.:
  - Do the concept domains and concept classes of target concepts comply with the CDM conventions?
  - Are the post-coordinated (e.g. 1-to-many, event-value pairs) mappings logically justified and structurally supported?

Discussion: To fully enable the benefit of the Common Data Model and portable analysis across datasets, semantic standardization (usage of a thoughtfully selected set of Standard concepts as targets) is equally important and intertwined with structural standardization (column and table structure). Emerging use cases and the overall complexity of the tasks require for the development of new approaches to custom vocabulary construction. Examples of such new approaches are:

- Wide mapping table (1) – alternative mapping table format to allow for multiple target entities
- Logic groups (2) – addition to the concept_relationship table which connects ‘Maps to’ relationships with their respective attributes
- Value_source_concept_id field – approach for better standardization of survey data
- Common Data Environment (2) – approach for mapping harmonization and enrichment

Limitations: OMOP vocabularies, as well as the CDM itself, are always the work in progress. In order to benefit from new Vocabulary, ETL or CDM developments, it is crucial to constantly re-evaluate and adjust the existing approaches and mappings.

References: [1] [2]

Tatsiana Skuhareuskaya, Vlad Korsik, Vojtech Huser, Alexander Davydov