Custom vocabulary techniques in ETL to OMOP CDM

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

Background: While OHDSI Standardized vocabularies available through Athena (e.g. ICD10CM, NDC) greatly facilitate the mapping of a big portion of source data, other source data encoded in terminologies not included in Athena or free text requires many special considerations. Data such as Medical histories, Allergies and Registries, including Surveys, are valuable for research purposes and need to be captured in the CDM. Currently creating a custom vocabulary with concept_id of >2 billion is a default approach (Table 1). Here we describe the reasoning behind some of the crucial choices made when creating a custom vocabulary - identification of custom concepts using pre- and post-coordination. We also touch on domain drift as an important factor when maintaining a custom vocabulary or developing and supporting custom terminologies.

Parameter	STCM	C/CR
General description	Creation of a mapping from source_code to a standard_concept_id	Creation of a custom concept (with concept_id>2bil) and mapping it to a standard_concept_id
Loading	Easy: table is designed for insertion of records	Requires alteration of vocabulary tables from Athena
Destiny of unmapped codes	Live only as source_codes	Live as custom concepts without mapping, i.e. as event_source_concept_id and can be queried by Atlas
Maps_to_value relationships	Does not support	Supports
ETL logic	Requires additional querying of the STCM	Allows for a usual lookup of a concept in the Concept table
Research use	None: data is not visible in Atlas or the Concept table	Custom concepts can be used in network studies through Atlas
Conclusion	**Easy to implement, but not sharable**	**More complex, but sharable and usable in network research**

Table 1: Comparison of source_to_concept_map and concept/concept_relationship approaches to ETL

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?

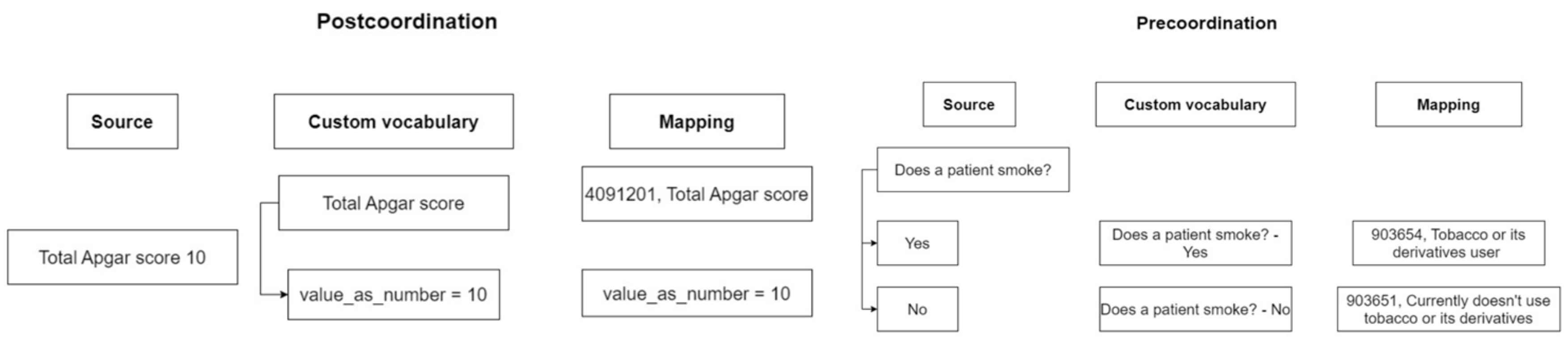
Semantic integrity checks, e.g.:

- 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?


Postcoordination
Source
Custom vocabulary
Mapping
Total Apgar score 10
Total Apgar score
value_as_number = 10
4091201, Total Apgar score
value_as_number = 10

Precoordination
Source
Custom vocabulary
Mapping
Does a patient smoke?
Yes
No
Does a patient smoke? - Yes
Does a patient smoke? - No
903654, Tobacco or its derivatives user
903651, Currently doesn't use tobacco or its derivatives

Figure 1: Examples of pre-coordination and post-coordination when creating a custom vocabulary (left) and changing mappings (right)

Afterwards, the iterative process of **maintaining** the custom vocabulary takes place. When OMOP Vocabularies are updated, target concepts can change their attributes and live cycle (Standard to non-Standard; Domain change is also possible) (Figure 2) and remappings may be necessary. Post-coordination should be taken into account and 'Maps to value' relationships should be created where necessary (Figure 3). This creates a need for additional semantic integrity checks, especially for cases where post-coordination has previously been used.

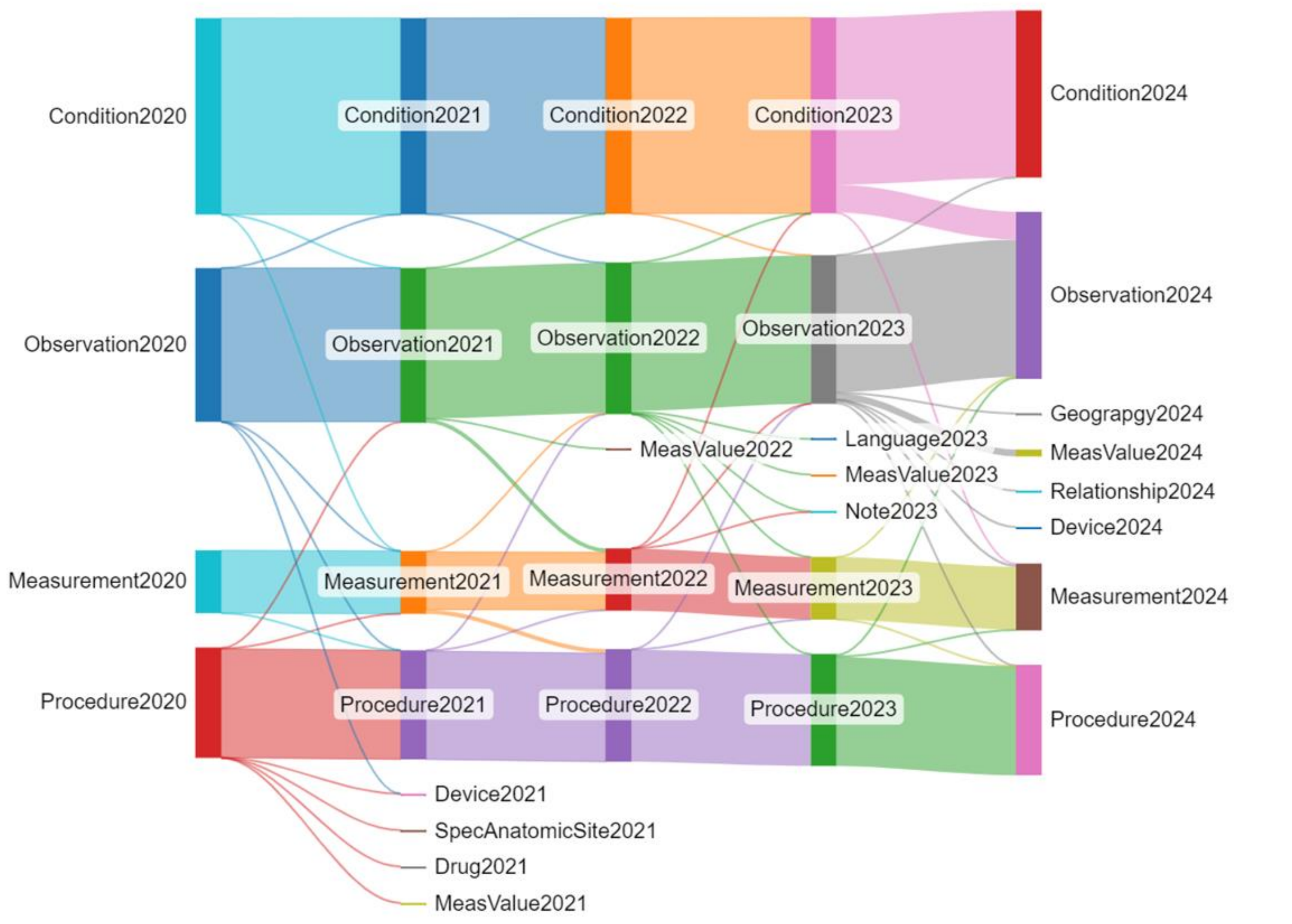


Figure 2: Domain changes for Standard concepts used in ETL across 2020-2024 years

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 [1] - addition to the concept_relationship table which connects 'Maps to' relationships with their respective attributes
- Value_source_concept_id field - approach for better standartization 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

ODYSSEUS DATA SERVICES INC