edenceMapper: Mapping Suggestion Framework for Non-Standard to Standard Codes

Freija Descamps[1], Shirah Cashriel[1], Isaac Claessen[1], Mythili Palanisamy[1]
[1]edenceHealth NV

Background

Mapping non-standard codes to standardized medical vocabularies is crucial for maintaining data consistency in healthcare systems. While small code lists can be manually mapped using tools like Athena, large-scale mapping requires an automated and efficient workflow.

To address this challenge, edenceHealth NV has developed a mapping suggestion framework called **edenceMapper**. This framework automates the semantic mapping process by utilizing multiple matching algorithms to generate ranked lists of suggested mappings towards standardized vocabularies such as OMOP and various ontologies.

Methods

edenceMapper is a containerized application that maps non-standard source codes to standard codes. It is a powerful recommendation engine that provides relevant suggestions to users by suggesting the best-matching standard concept for given code.

The framework is divided into two primary components: Translator and Mappers.

1. Translator

When the source concepts are in a language other than English, translation is required when mapping to English target vocabularies. Currently, the AWS translator is implemented and can be used for all the mappers, which acts as a wrapper around the Amazon Translate API, a fully managed service that provides text translation. The service supports a wide range of languages and allows custom translation to be uploaded. However, when using the multilingual transformer model, separate translation is unnecessary, as the model itself handles multilingual input processing.

2. Mappers

The edenceMapper code includes a Mapper base class, which serves as a parent class for all specific mapping algorithms (matching source to target terms) implementations. Any new mapper must inherit from the Mapper class, ensuring consistency and adherence to a predefined structure. This design facilitates the seamless addition of new mappers by enforcing a common interface and behavior across all mapper subclasses. To map source codes to standard codes, various mappers have already been integrated into the framework, including:

- **Fuzzy string matching** helps identify similar but non-identical strings using Levenshtein Distance to calculate the differences between two terms.
- **Locality-Sensitive Hashing (LSH)** improves efficiency by significantly speeding up the process of finding similar terms in a large dataset. It does so by grouping comparable items using specialized hash functions.
- The **Lucene algorithm** makes use of the Apache Lucene open-source search engine software library. It provides a robust toolset for text comparison and information retrieval, utilizing indexing, querying, tokenization, and scoring mechanisms to enhance search accuracy and efficiency.

- The **Multilingual-E5-small architecture** refines text matching beyond basic keyword comparison. By leveraging sentence embedding models, textual data is transformed into high-dimensional (vector)^2 representations that encode semantic meaning. These embedding vectors, composed of floating-point values, enable more accurate similarity assessments. The necessary vocabularies are vectorized and stored into a database. The source codes are converted into vector representations and compared with the vectorized standard code to extract better match.

Semantic mapping

Semantic mapping involves aligning local source terms with target vocabulary based on semantic similarity. The edenceMapper framework allows for the implementation of multiple algorithms, each defined as an instance of the specific mapper class. This process goes beyond exact word matching by understanding the contextual meaning of terms and phrases.

To improve mapping accuracy, the initial step includes the following refinement criteria:

- **Target Vocabulary**: Ensures mapped concepts align with predefined vocabulary.
- **Standard Concepts**: Prioritizes standard concepts unless exceptions or specific conventions dictate otherwise.
- **User-Defined Subsets**: Allows users to define and select subsets of concepts relevant to their needs.

Semantic mapping is an iterative and adaptable process, adjusting to different data sources and user-defined criteria. This flexibility ensures accurate mappings that align with both the context of the data source and its intended application. Once the search space is refined, the next step is to map local source codes to target vocabulary codes based on semantic similarity.

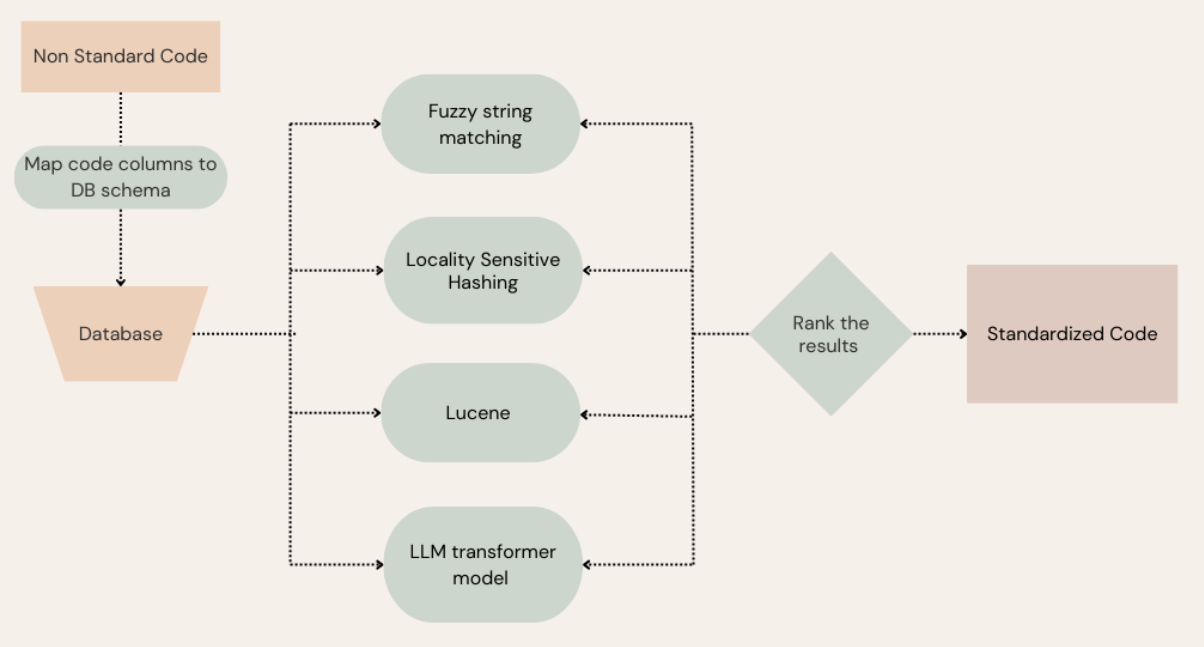

Figure 1. Flowchart that represents the workflow of non-standard code to standard code

Conversion of non-standard code to standard code

Depending on the source file and user requirements, manual steps may be necessary for pre-processing source code information. (Figure 1) At this stage, non-standard code columns are transformed to align with the database schema. The framework utilizes both predefined mappers for conversion and also offers the flexibility to incorporate new mappers. These mappers interact with PostgreSQL databases containing vocabulary files and vectorized standard code files. After the mapping is completed, the results are ranked and stored in database

As a final step, a review by a medical expert is required to validate and confirm the accuracy of the mappings, for which a web-based collaborative review tool developed by edenceHealth, edenceReviewer, can be used.

Results

Table 1 presents the mapping outputs for different source termsacross different mapping models. Each model applies a different technique to determine the best match within a standardized vocabulary.

Source code	**Suggested mapping per model** (concept_id [concept_name])		
	Lucene	**Fuzzy**	**Multilingual LLM**
Haemodialysis	4120120 [Hemodialysis]	4120120 [Hemodialysis]	4120120 [Hemodialysis]
Packed cell transfusion	4125928 [Packed blood cell transfusion]	4125928 [Packed blood cell transfusion]	4125928 [Packed blood cell transfusion]
Ven cath renal dialysis	4146536 [Renal dialysis]	4289454 [Venous catheterization for renal dialysis]	4289454 [Venous catheterization for renal dialysis]

Table 1: the mapping results of different models for different example source codes.

Conclusion

The edenceMapper framework is an ongoing project designed to efficiently map non-standard source terms (regardless of language) to standardized vocabularies. By integrating advanced NLP techniques and multiple mapping algorithms, it provides automated suggestions while allowing for expert validation, ensuring high accuracy and reliable standardization in healthcare data management.

References

1. Multilingual-LLM-model, https://huggingface.co/intfloat/multilingual-e5-small
2. Vector semantic mapping, https://github.com/pgvector/pgvector