Introducing Ananke, A Tool for Mapping Between OHDSI Concept Identifiers to Unified Medical Language System (UMLS) identifiers

Juan M. Banda, PhD[1]
[1] **Georgia State University, Atlanta, Georgia**

Abstract

Using clinical narratives available in Electronic Medical Record (EMR) systems has become a defacto task for many knowledge discovery and learning projects. With the addition of the NOTE_NLP table to the OMOP Common Data Model (CDM), OHDSI practitioners now have the option of storing any concepts or terms extracted from their clinical text into structured fields. These enhanced capabilities bring new challenges, particularly the need to consolidate the output of NLP text extraction/annotation systems into concept identifiers within the OHDSI vocabulary. Most widely available tools, like cTakes for example, rely on the Unified Medical Language System (UMLS) to label the extracted terms and concepts. Both the OHDSI vocabulary and UMLS contain some of the same domain-specific vocabularies, but the mapping between unique identifiers is not completely straightforward. With Ananke, we have developed a tool that provides a guided mapping between UMLS concept unique identifiers (CUIs) and OHDSI concept identifiers. This tool guides the user in the task of creating mostly automatic mappings between the OHDSI and UMLS vocabularies.

Introduction

Ananke in Greek mythology is the personification of inevitability, compulsion, and necessity, which very well relates to the need to have the OHDSI vocabulary mapped to other available and more widely used resources, like the Unified Medical Language System (UMLS). The UMLS metathesaurus (1), contains over three million concepts and over 130 English vocabularies, the OHDSI vocabulary on the other hand covers over 70 vocabularies with many of them overlapping. While the vocabulary is the defacto standard in the OHDSI community, UMLS is widely used outside of the community for most natural language processing (NLP) tools and tasks. The software presented in the poster bridges the gaps between the communities, allowing for the previously impossible interoperability between tools.

How Ananke works

We developed Ananke using UMLS 2017AB, which has 131 English vocabularies, and the OHDSI Vocabulary (Jan-2018) which has 71+ vocabularies. This is posing multiple issues. Firstly, and quite evidently, we have plenty of vocabularies in UMLS that are not contained in the OHDSI vocabulary and vice versa. Secondly, internal versions of vocabularies between UMLS and OHDSI vocabulary exist, with some older in UMLS for the most part.

The process of mapping UMLS concept unique identifiers (CUI) begins with selecting the proper vocabulary in UMLS and on the OHDSI vocabulary. This is done via the SAB field in UMLS and the vocabulary_id field on the OHDSI Vocabulary. The vital detail in this step is that the versions of the source vocabularies used usually differ between the resources, with OHDSI Vocabulary having newer versions of them.

Once the UMLS and OHDSI vocabulary names are identified, the mapping relies on using the CODE field in UMLS and the concept_code in OHDSI vocabulary. This is, using their source vocabulary identifiers to get to the CUI and the concept_id respectively. Figures 1 to 3 show one concept and how it is

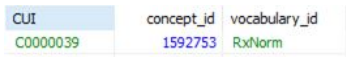
represented in both UMLS (Figure 1) and OHDSI vocabulary (Figure 2), as well as the resulting mapping generated by Ananke.

CUI	LAT	TS	LUI	STT	SUI	ISPREF	AUI	SAUI	SCUI	SDUI	SAB	TTY	CODE	STR	SRL	SUPPRESS	CVF
C0000039	ENG	P	L0000039	PF	S17175117	N	**A28315139**	9194921	1926948	(NULL)	RXNORM	IN	1926948	1,2-dipalmitoylphosphatidylcholine	0	N	(NULL)

Figure 1. UMLS record for an RxNorm concept.

concept_id	concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code	valid_start_date	valid_end_date	invalid_reason
1592753	1,2-dipalmitoylphosphatidylcholine	Drug	RxNorm	Ingredient	S	1926948	2017-08-07	2099-12-31	

Figure 2. OHDSI vocabulary record for the equivalent RxNorm concept from Figure 1.



CUI	concept_id	vocabulary_id
C0000039	1592753	RxNorm

Figure 3. Resulting

Current progress

After overcoming the previously mentioned challenges and developing a strategy to craft the automatic mapping scripts, we have managed to automatically map 1,119,625 UMLS CUIs into OHDSI concept identifiers. We have released these mappings initially as a list (available at https://github.com/jmbanda/OHDSIconceptid2cui/blob/master/CUItoOHDSIv1.zip). The list only includes the UMLS CUI and the OHDSI concept_id in order to avoid any licensing issues with UMLS and their underlying vocabularies. More detailed mappings can be obtained via Ananke, containing all fields, as the licensing issues should be handled by the user of our package. These licensing issues are addressed by the user when downloading and installing UMLS, as well as the OHDSI vocabulary with the restricted vocabularies.

Table 1. Total number of UMLS concepts mapped into their respective OHDSI concept identifiers.

Vocabulary Name	Mapped concepts
CPT4	39,947
HCPCS	6,376
ICD10CM	101,725
ICD10PCS	180,450
ICD9CM	16,376
LOINC	124,813
MedDRA	52,953
NDFRT	23,345
RxNorm	202,627
SNOMED	371,013

Conclusion

The first version of Ananke contains functionality to map between ten major vocabularies: ICD9, ICD10, HCPCS, SNOMED, RxNorm, CPT, MedDRA, NDFRT and LOINC. These vocabularies cover over a million terms, a third of the main OHDSI vocabulary, and it includes the most important domains, which are diagnosis, procedures, observations, measurements, drugs, and devices. While certainly a work in progress, this software is highly customizable as mapping queries as well as individual mappings can be added via simple mechanisms. We are looking forward to releasing this work and having input, as well as improvements by the OHDSI community. Ananke is available at: https://github.com/jmbanda/Ananke.

References

1. Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. Bull Med Libr Assoc. 1993 Apr;81(2):170–7.