ANANKE 2.0: An Update on the Tool for Mapping Between OHDSI Concept Identifiers to Unified Medical Language System (UMLS) identifiers - now with mappings to the Human Phenotype Ontology (HPO)

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

Abstract

Originally introduced at the 2018 OHDSI symposium, Ananke is a tool that provides automated mapping between UMLS concept unique identifiers (CUIs) and OHDSI concept identifiers. While the first version of Ananke satisfied the basic requirements of mappings between popular vocabularies such as: SNOMED, LOINC, RxNorm, ICD9, and ICD10, among others, coverage was not very comprehensive. In this work we introduce version 2 of the software and mappings that incorporate the additions made to UMLS since version 2017AB to version 2020AA and to the OHDSI Vocabulary up to the end of June 2020. This increases coverage by over 300,000 concepts since the first release. We also add vital mappings between OHDSI concept_id's and the Human Phenotype Ontology (HPO) as part of ongoing work being performed on the N3C Tools and Resources group.

Research Category

Tools and resources

Introduction

Ananke in Greek mythology is the personification of inevitability, compulsion, and necessity, which relates to the need to have any other clinical vocabulary mapped into the OHDSI vocabulary. Our tool facilitates the mapping of any vocabulary or ontology that is available in both the OHDSI vocabulary and the Unified Medical Language System (UMLS). The UMLS metathesaurus(1) (version 2020AA) contains 10,953,636 concepts from 150 English vocabularies, the OHDSI vocabulary on the other hand covers over 70 vocabularies with many of them overlapping between the two. 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 here bridges the gaps between the communities, allowing for the previously impossible interoperability between tools. In this release we also bridge mappings between the OHDSI Vocabulary and the Human Phenotype Ontology(2), which is not on the OHDSI Vocabulary, but can be found in UMLS. The need of this mapping is for higher resolution on phenotyping tasks and additional identification of rare diseases. This shows that Ananke can also be used to link further outside of the OHDSI vocabulary through UMLS.

How Ananke mappings work

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. However, these discrepancies are usually resolved, and new mappings can be added when a newer version of UMLS is released (as shown on Table 1). 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. That is, using their source vocabulary identifiers to get to the CUI and the concept_id respectively. Over versions 1 and 2 of Ananke, one can see (table 1) that the number of mappings covered have increased over time (Table 1, columns 2 and 3) with the exception of RxNorm where some drugs have been removed from UMLS and do not have a mapping anymore. All code for the mappings is available in the Github Repository: https://github.com/thepanacealab/OHDSIananke

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

Vocabulary Name	Mapped concepts (Version 1)	Mapped concept (Version 2)
CPT4	39,947	44,083
HCPCS	6,376	7,193
ICD10CM	101,725	104,733
ICD10PCS	180,450	189,499
ICD9CM	16,376	17,553
LOINC	124,813	215,111
MedDRA	52,953	76,241
RxNorm	202,627	198,962
SNOMED	371,013	493,665

Adding HPO mappings to Ananke

Leveraging resources outside of the OHDSI vocabulary has been difficult to achieve, most of the times the need to have an incompatible OHDSI instance is the result of using these vocabularies for either structured data or text annotations. With our addition of mappings between the HPO and the OHDSI vocabulary we leverage UMLS internal mappings between vocabularies that are present in the OHDSI vocabulary to link to one outside of the OHDSI vocabulary. We used the HPO UMLS mappings to SNOMED, ICD9, ICD10, MeSH, and MEDRA to achieve linkage to CUIs that have an OHDSI vocabulary mapping. While this technique is fairly straight forward, the limitation is that if the UMLS mappings are not very complete, there will be plenty of terms without mapping as only 4,808 out of 14,831 HPO unique terms have a mapping with this method. The actual SQL queries to do this are also found on the Ananke GitHub repo. As a work in progress as part of the N3C Tools and Resources groups, other researchers are working on adding more mappings that are not directly found via UMLS. These extra mappings will be added in the future as they become available.

Conclusion

The first version of Ananke showcased the initial versions of mappings between several major vocabularies: ICD9, ICD10, HCPCS, SNOMED, RxNorm, CPT, MedDRA and LOINC. These vocabularies cover 1,403,927 terms, around a third of the core OHDSI vocabulary, including the most important domains, which are diagnosis, procedures, observations, measurements, drugs, and devices. The expansion into adding mappings to the HPO opens a very interesting line of work which can allow others to leverage the internal UMLS mappings to link additional resources not found in the OHDSI Vocabulary in an automated way. 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. Metathesaurus. 2009 Sep 3 [cited 2020 Jul 9]; Available from: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

2. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet [Internet]. 2008 Nov;83(5):610–5. Available from: http://dx.doi.org/10.1016/j.ajhg.2008.09.017