Transitioning ANANKE to OMOP2OBO for more robust NLP extraction and knowledge graph data representation leveraging the OHDSI vocabulary

Juan M. Banda, Tiffany J. Callahan

Background

Over the years, there have been several attempts to bring the OHDSI vocabulary into the semantic data world(1) while linking different clinical vocabularies and terminologies (i.e. UMLS) via Ananke(2). These efforts have been sporadic and not properly funded, so they have been mostly slow and out of the necessity of workgroups like the Natural Language Processing (NLP) workgroup to be able to link the outputs of NLP tools like CLAMP(3) from UMLS CUI's to OHDSI concept identfiiers. Not until 2020, Callahan et al.(4) a more comprehensive, rigorous, and use-cased based utility was developed. In this work we outline the inclusion of ANANKE into OMOP2OBO in order to provide a single, comprehensive utility for these tasks.

Methods

OMOP2OBO is a Python 3 library designed to align standard OMOP terminology concepts to terms in Open Biomedical Ontology (OBO) Foundry ontologies. To make these alignments, OMOP2OBO first utilizes exact alignment between database cross-references and strings (i.e., labels, synonyms, and definitions). If no alignment can be found at the concept-level, the algorithm then looks for alignment between the OMOP concept and ancestor concepts in the OBO ontology hierarchy. Finally, OMOP2OBO also includes modules to enable mappings via concept similarity. All mappings are annotated with evidence and provenance that explains how a mapping was created. ANANKE functionality was added to the existing OMOP2OBO as a new function, allowing it to seamlessly be incorporated with existing functionality. Figure 1 shows an overview of the OMOP2OBO process.

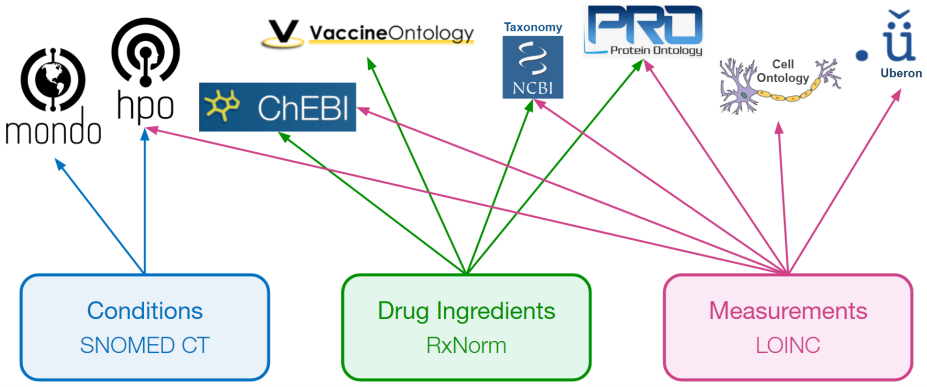

Figure 1. Sample vocabularies/ontologies that OMOP2OBO links to OMOP (5).

Results

Prior to incorporating ANANKE into OMOP2OBO, a brief analysis was performed to align SNOMED to the Human Phenotype Ontology as a means of understanding similarities and differences between each method. Figure 2 showcases the original differences and the utility of incorporating ANANKE into OMOP2OBO.

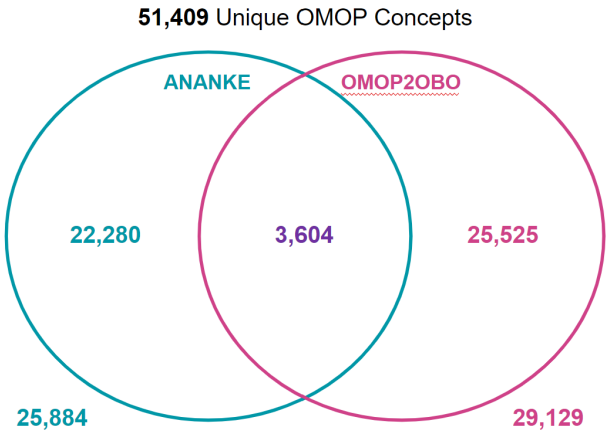

Figure 2. Total number of OMOP concepts that cover HPO terms, and their mappings via ANANKE (left) and OMOP2OBO (right).

Here we can see that while there is some overlap (3,604 concepts), there are very valuable OMOP concept mappings on both resources, and joining them provides the best available resource for such mappings. Similar analysis could be performed for other vocabularies (SNOMED_CT, RxNorm, etc.) but that is out of the scope of this small submission and will be part of a research paper in the future.

Conclusion

By merging the functionality of Ananke with OMOP2OBO, we are not only able to link UMLS CUI's to OHDSI concept identifiers, but also leverage OBO ontologies, and expert curation to enrich concept sets, and provide the semantic integration of additional standardized clinical terminologies. Keeping all this functionality under a single tool allows the community to easily find our resource and and improve its discoverability and maintenance.

References/Citations

1. Banda JM. Fully connecting the Observational Health Data Science and Informatics (OHDSI) initiative with the world of linked open data. Genomics Inform [Internet]. 2019 Jun;17(2):e13. Available from: http://dx.doi.org/10.5808/GI.2019.17.2.e13

2. Banda JM. Introducing Ananke, A Tool for Mapping Between OHDSI Concept Identifiers to Unified Medical Language System (UMLS) identifiers [Internet]. 2018. Available from: https://www.ohdsi.org/wp-content/uploads/2018/10/2018-Poster-Session-FINAL-200-black-and-white-copies-double-sided-stapled.pdf

3. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc [Internet]. 2018 Mar 1;25(3):331–6. Available from: http://dx.doi.org/10.1093/jamia/ocx132

4. OMOP2OBO: Semantic Integration of Standardized Clinical Terminologies to Power Translational Digital Medicine Across Health Systems [Internet]. [cited 2022 Jun 11]. Available from: https://www.ohdsi.org/2020-global-symposium-showcase-23/

5. Callahan TJ. Ontologizing Health Systems Data at Scale: Making Translational Discovery a Reality [Internet]. 2021. Available from: https://zenodo.org/record/5895112