# A Toxin Vocabulary for the OMOP CDM

**Maksym Trofymenko, MD[1], Polina Talapova, MD, PhD,[1,2,5], Tetiana Nesmiian MD, LLM [1,3], Andrew Williams, PhD[5], Denys Kaduk, MD[1, 4], Max Ved[1], Inna Ageeva[1]**

**[1]IT company SciForce, Kharkiv, Ukraine**

**[2]Kharkiv National Medical University, Kharkiv, Ukraine**

**[3]Kharkiv National Pedagogical University named after H. S. Skovoroda, Kharkiv, Ukraine**

**[4]V. N. Karazin Kharkiv National University, Kharkiv, Ukraine**

**[5]Tufts Clinical and Translational Science Institute, Boston, MA, United States**

## Background

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is a widely adopted open community data standard that facilitates large-scale evidence generation by harmonizing the structure and content of observational data. Adoption of the OMOP CDM allows researchers and healthcare organizations around the world to perform drug safety monitoring, conduct comparative effectiveness research, design and assess clinical trals, and predict healthcare outcomes.

On the other hand, environmental epidemiology places significant emphasis on investigating the impacts of exposure to toxic substances on human health, both in the short and long term perspective. To support these studies, Geographic Information Systems (GIS) are utilized.[1] While recent efforts aim to integrate GIS data with the OMOP CDM,[2] the lack of sufficient standards impedes the comprehensive evaluation of environmental exposures.

Here we introduce our hierarchical Toxin Vocabulary model as a solution to address this gap in representing toxic substances. This model has the potential to be seamlessly integrated into the OMOP Standardized Vocabularies.

## Methods

In this study, we conducted a systematic review of existing toxicological literature, analyzed open-source toxin databases, and leveraged domain expertise on regulatory documents to synthesize the anatomy of the Toxin Vocabulary. Our goal was to identify pertinent terms and classifications related to various exposomes and their influence on human health. The Toxin and Toxin Target Database (T3DB)[3] stood out as the most comprehensive and reliable resource for toxin terminology.

T3DB currently houses over 3,000 toxins described by 41,602 synonyms, including pollutants, pesticides, drugs, and food toxins, with corresponding toxin target records. Each toxin record (ToxCard) contains comprehensive information such as chemical properties, toxicity values, molecular and cellular interactions, and medical details across more than 90 metadata fields.

The integration process involved automatically uploading the source data to the PostgreSQL database using Python. We then extracted essential metadata, established cross-term connections, and performed a semi-automated mapping of selected terms to the OMOP Vocabulary standards

To establish the identification of concepts using concept_codes, we employed CAS codes due to their alignment with GIS data and the CAS Registry[4], one of the largest registries encompassing around 204 million organic substances. Toxins without CAS codes were assigned unique T3DB codes, while manually injected Classification concepts were assigned auto-generated codes.

To ensure the seamless integration of the Toxin Vocabulary into the OMOP CDM, staging tables such as vocabulary_stage, concept_stage, concept_relationship_stage, concept_ancestor_stage and concept_synonym_stage were utilized to incorporate both semantics and syntactics.

**Results**

The structure of the developed Toxin Vocabulary, depicted in Figure 1, encompasses over 79,377 internal relationships among toxins, target cellular and tissue structures, proteins, related medical conditions, biological processes, and toxin categories. Additionally, there are 7,822 external "Maps to" relationships that implement the mapping of the Toxin Vocabulary to OMOP CDM Standardized Vocabularies, including SNOMED CT, RxNorm, and RxNorm Extension. Exposomes without direct equivalents are maintained as standard concepts.

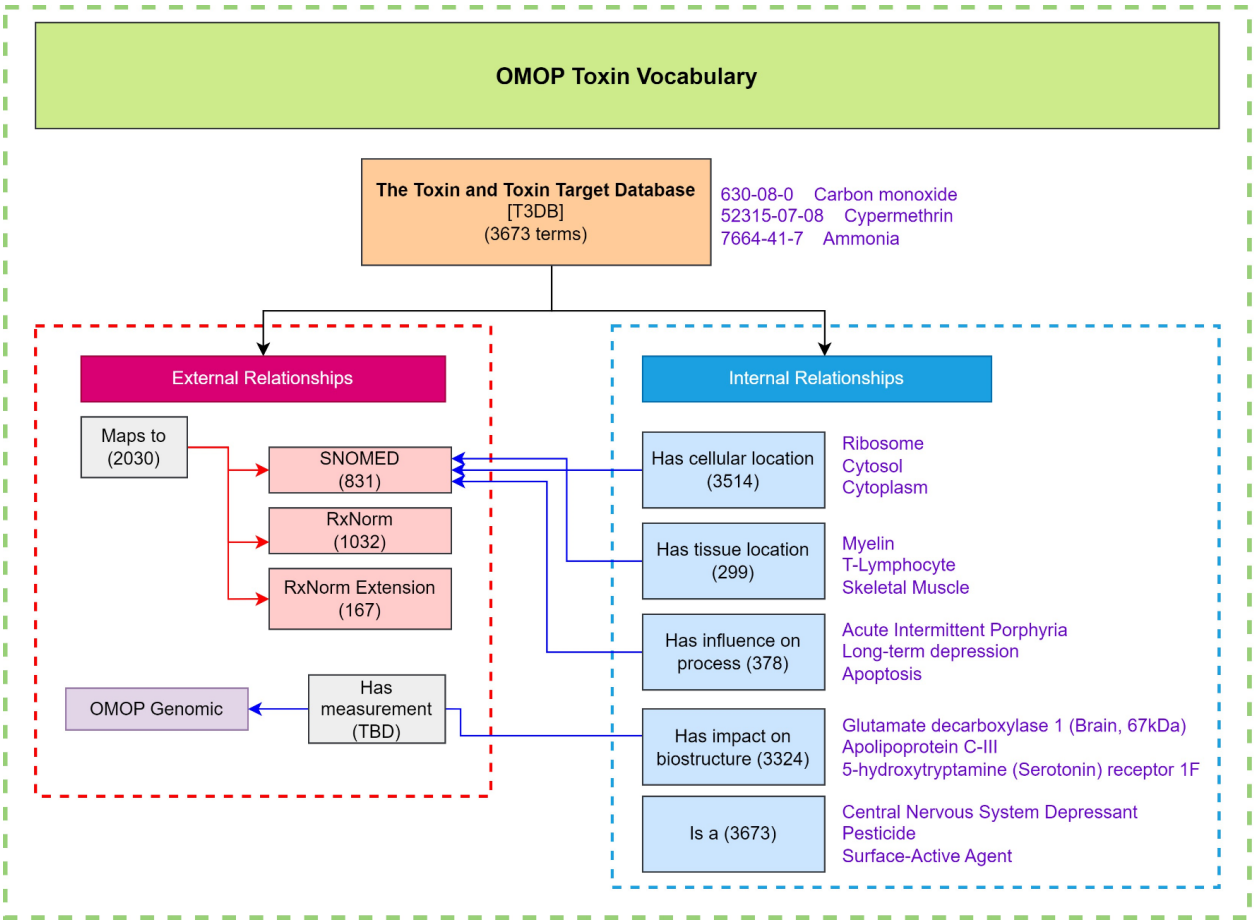


**Figure 1. Architecture of the Toxin Vocabulary** (violet text indicates source data examples, red arrows represent newly developed mappings, blue arrows denote future cross-links)

During the toxin term mappings, 239 instances of standard duplicates shared by SNOMED and RxNorm/RxNorm Extension were identified. To resolve this issue, we utilized replacement mapping, where SNOMED concepts were mapped to RxNorm/RxNorm Extension concepts.

**Conclusion**

The next steps involve expanding the toxin mapping network, conducting thorough quality assurance, presenting the Toxin Vocabulary to the OHDSI community, gathering user feedback, and submitting the Toxin Vocabulary for inclusion in the OMOP Standardized Vocabularies using the OHDSI Vocabulary team's template #4 “adding_vocabulary”. In the case of acceptance, the newly added ontology of toxins for the OMOP CDM will be a valuable addition to the impressive family of OMOP Standardized Vocabularies. The implementation of such terminology will open the window to unlimited opportunities for GIS-related and toxicoepidemiological studies.

**References**


1. Rosenkrantz L. Leveraging geographic information systems (GIS) for environmental public health practice. Environmental Health Review. 2022;65(2):31-36. DOI: 10.5864/d2022-013.
2. Cho J, You SC, Lee S, Park D, Park B, Hripcsak G, Park RW. Application of Epidemiological Geographic Information System: An Open-Source Spatial Analysis Tool Based on the OMOP Common Data Model. International Journal of Environmental Research and Public Health. 2020; 17(21):7824. DOI: 10.3390/ijerph17217824.
3. The Toxin and Toxin Target Database (T3DB) [Internet]. Available from: http://www.t3db.ca/
4. CAS. CAS Registry [Internet]. Available from: https://www.cas.org/cas-data/cas-registry.
5. Wishart D, Arndt D, Pon A, et al. T3DB: the toxic exposome database. Nucleic Acids Res. 2015;43(Database issue):D928-D934. DOI: 10.1093/nar/gku1004.