LLM-Based Ontology Mapping for Privacy-Preserving Healthcare Data Management

Maria Papoutsoglou[1,*], Apostolos Mavridis[1], Stergios Tegos[1], Christos Anastasiou[1] and Georgios Meditskos[1]

[1]School of Informatics, Faculty of Sciences, Aristotle University of Thessaloniki, Greece

Abstract

The rapid growth of healthcare data requires efficient privacy-preserving management and analysis methods. Traditional relational databases often lack the necessary contextual understanding for advanced analytics and regulatory compliance. Using techniques such as large language models (LLMs), we can enhance relational data with semantic metadata, creating knowledge graphs. These graphs support applications such as automated data de-identification, clinical decision support, and research analytics. This paper presents a framework combining LLMs, ontologies and vector databases to improve healthcare data understanding and standardization.

Keywords

Large Language models, LLM, ontology, knowledge graph, SNOMED

1. Introduction

As data grows exponentially, the processing of sensitive information such as medical records while preserving privacy becomes urgent. Relational databases (RDBs) are essential in healthcare but often lack the context for effective privacy strategies due to missing legal and technical terms and multiple stakeholders. Semantic Web technologies link data with semantic metadata, uncovering implicit meanings and relationships. Leveraging existing ontologies to enrich RDBs creates knowledge graphs that improve data understanding and privacy management. Advanced methods such as large language models (LLMs) outperform traditional algorithms such as BERT in processing complex data relationships.

The combination of SNOMED CT, vector databases, and LLMs offers a robust framework for semantic privacy analysis. Vector databases efficiently manage high-dimensional healthcare data, while SNOMED CT provides extensive medical terminology. Enhanced with LLMs, this system analyzes unstructured data (e.g., clinical notes), to identify sensitive information and propose de-identification strategies. This study presents an empirical ontology mapping framework that integrates LLMs and vector databases to semantically map clinical terms to SNOMED CT, addressing challenges in healthcare data standardization and scalability.

SWAT4HCLS 2025: The 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences

*Corresponding author.

mpapo@csd.auth.gr (M. Papoutsoglou); apostolos@enchatted.com (A. Mavridis); stergios@enchatted.com (S. Tegos); christos@enchatted.com (C. Anastasiou); gmeditsk@csd.auth.gr (G. Meditskos)

2. Related Work

LLMs and ontologies are key to advancing biomedical data analysis. Integrating ontologies such as SNOMED CT with LLMs improves natural language understanding and data standardization, enhancing tasks like medical concept normalization and entity extraction. However, challenges in standardization and evaluation remain [1]. Efforts like SiMHOMer use domain-specific models and LLMs to merge and enrich healthcare ontologies, creating new semantic relationships, and paving the way for improved ontology-driven systems [2]. LLM pipelines have automated clinical condition detection and reimbursement coding in EHRs, achieving high sensitivity and specificity for conditions such as gastrointestinal bleeding [3]. Transformer-based NLP models outperform traditional methods in mapping EHR data to standardized concepts, improving interoperability and enabling automated knowledge graph creation [4]. LLMs have also been used for out-of-knowledge-base discovery and concept placement in SNOMED CT [5]. Our approach integrates SNOMED CT with context-aware relationships from LLMs, creating dynamic semantic mappings for improved healthcare data standardization and an adaptable ontology for evolving datasets.

3. Methodology

We developed an ontology mapping system that leverages large language models (LLMs) and vector database technology to automatically generate semantic mappings between clinical terms and standardized medical ontologies. The approach is depicted in Figure 1.

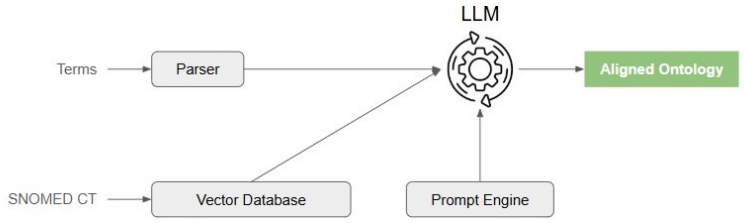

Figure 1: Proposed framework

More_than_55yo	Age	Sex	Eco_pre_surg	Procedure_presurgery_therapy	Surgery
0	48	1	1	2	1
1	78	1	0	2	2
0	27	1	0	1	3
1	78	1	1	0	2
1	70	1	0	4	2
0	52	1	0	4	2
0	48	1	1	0	1

Table 1
Example subset of the initial health dataset.

The system consists of two components: an LLM-based NLP module with knowledge graph generation and a vector database containing SNOMED CT ontology embeddings. It processes

tabular data with medical terms (see Table 1), which undergoes pre-processing to standardize and remove inconsistencies. The preprocessed data are used for ontological mapping. We use Claude 3.5 Sonnet v2 as the LLM, which processes medical terms to generate semantic understanding. Using Chain of Thought (CoT) and template-based prompting, the system interprets clinical concepts, even when expressed in a non-standard way.

A vector database was created with SNOMED CT concept embeddings, chosen for its comprehensive clinical terminology and widespread use in healthcare systems [6]. The database enables efficient semantic similarity searches. The mapping process uses a hybrid approach, combining vector and keyword-based retrieval. Input terms are processed by the LLM to generate semantic representations, which are queried in the vector database and through full-text searches of SNOMED CT. Cohere's rerank-english-v3.0 model with a TopK value of 10 re-ranks results for better semantic relevance. This hybrid method addresses the limitations of pure vector search, ensuring more accurate matches. Final mappings are selected based on confidence scores above a set threshold.

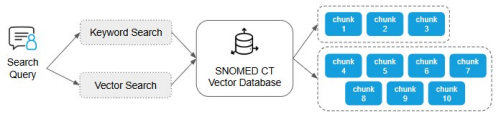

Figure 2: Word processing pipeline

```
:procedure_presurgery_therapy rdf:type owl:DatatypeProperty ;
    rdfs:domain :Patient ;
    rdfs:range xsd:string ;
    rdfs:comment "The type of therapeutic intervention or treatment administered to the patient before surgery" ;
    skos:related <http://snomed.info/id/443425001> .

:PTC rdf:type owl:DatatypeProperty ;
    rdfs:domain :Patient ;
    rdfs:range xsd:string ;
    rdfs:comment "Post-tetanic count measurement, which is a neuromuscular monitoring technique used to assess
    the degree of neuromuscular blockade" ;
    skos:related <http://snomed.info/id/258067009> .
```

Figure 3: Example of ttl file

The system outputs mappings as a knowledge graph in RDF triple format, serialized in Turtle (TTL) syntax, following Semantic Web standards. Each mapped concept is represented as a node with relationships defined using standard ontologies like RDF, RDFS, OWL, and SKOS. The knowledge graph includes class definitions for clinical entities, data properties describing attributes, semantic relationships to SNOMED CT using `skos:related` predicates, labels, comments for readability, and domain and range specifications for properties.

The resulting knowledge graph is encoded in TTL format with standard namespace prefixes (rdf, rdfs, owl, xsd, skos). Each mapped concept retains its semantic link to the corresponding SNOMED CT concept via unique identifiers (e.g., SNOMED). The system implements quality control measures, including RDF syntax validation, semantic consistency checks, SNOMED CT concept reference verification, assessment of mapping confidence scores, and expert manual review. This approach enables scalable, automated mapping of clinical terms to standardized ontologies while ensuring semantic accuracy and adherence to Semantic Web standards. An example of the process and the generated TTL file can be found in Figure 2 and Figure 3, respectively.

4. Evaluation

To evaluate the quality and practical utility of our ontology mapping system, we conducted a comprehensive assessment with four domain experts. The evaluation focused on a knowledge graph generated from 48 medical terms in our dataset, and was structured into three areas: Overall system performance, technical accuracy, and usability. Experts rated the system on a 5-point Likert scale (1=lowest, 5=highest) across three dimensions: (A) usefulness of the generated

graph for their work, (B) required post-processing effort, and (C) trust for production use. The technical evaluation assessed each of the 48 properties individually. Experts rated the accuracy of SNOMED CT concept mappings on a 5-point scale and evaluated the appropriateness of selected datatypes using categorical responses (Yes/No/Partially). They also identified any missing or incorrect relationships. The usability evaluation focused on the clarity and usefulness of property comments, and the suitability for clinical applications, with experts rating these aspects on a 5-point scale. Additionally, they provided qualitative feedback on potential improvements for each mapping.

5. Results

The evaluation was conducted with four domain experts who assessed both overall system performance and specific property mappings. All results are clearly depicted in Figure 4. The system's overall usefulness received high ratings (M=4.75, SD=0.43 out of 5), with three experts giving maximum scores (5/5) and one expert rating it 4/5. The required post-processing effort was consistently rated as minimal to moderate (M=1.75, SD=0.43 out of 5), indicating efficient initial mappings. Trust in the system for production use was also highly rated (M=4.25, SD=0.43 out of 5).

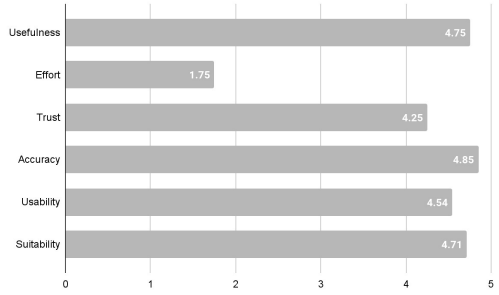

Figure 4: Results of metrics

For the technical evaluation of individual properties (n=48), the SNOMED CT concept mapping accuracy showed strong consistency across evaluators, with a mean score of 4.85 (SD=0.36). All of the properties received appropriate datatype assignments. The usability assessment revealed high scores for comment clarity and usefulness (Q4: M=4.54, SD= 0.58) and clinical application suitability (Q5: M=4.71, SD=0.46). Core clinical properties such as ID, Birth_date, Age, Sex, and Surgery consistently received maximum scores (5/5) from all evaluators across all assessment dimensions. Inter-rater reliability analysis revealed satisfactory agreement among evaluators, with Fleiss' Kappa values exceeding 0.80 for all metrics (system usefulness: $\kappa = 0.80$, post-processing needs: $\kappa = 0.85$, trustworthiness: $\kappa = 0.82$, mapping accuracy: $\kappa = 0.83$, comment clarity: $\kappa = 0.81$, and clinical suitability: $\kappa = 0.82$).

The open-ended feedback from evaluators was largely positive. Experts confirmed that core clinical relationships and hierarchical structures were accurately captured, with proper representation of direct and temporal relationships between medical concepts. Suggested modifications

were focused on potential enhancements rather than corrections, such as adding cross-references to other medical terminologies (e.g., ICD-10), incorporating temporal constraints, and expanding capabilities to handle unstructured clinical notes. This qualitative feedback complemented the strong quantitative results, indicating that the system met its primary mapping objectives while leaving room for future feature expansions.

6. Conclusions and Future challenges

This paper presents a framework integrating SNOMED CT with LLMs and vector databases to improve healthcare data standardization. By creating a flexible, adaptable ontology, our approach provides a scalable solution for clinical data mapping. However, challenges remain in aligning diverse data structures and incorporating complex terminologies, while data privacy concerns limit full schema-based mappings. Future research will focus on refining the mapping process, integrating relational structures within SNOMED CT, and enhancing privacy measures. The framework also holds potential for other sectors requiring semantic interoperability and privacy-compliant computation, enabling knowledge sharing and decision-making across domains.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101070670 (ENCRYPT).

Declaration of Generative AI Use

During the preparation of this work, the author(s) used ChatGPT for grammar and spelling checking, and for paraphrasing and rewording text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] E. Chang, S. Sung, et al., Use of snomed ct in large language models: Scoping review, JMIR Medical Informatics 12 (2024) e62924.

[2] S. Menad, S. Abdeddaïm, L. F. Soualmia, Simhomer: Siamese models for health ontologies merging and validation through large language models, in: International Work-Conference on Bioinformatics and Biomedical Engineering, Springer, 2024, pp. 117–129.

[3] N. S. Zheng, V. K. Keloth, K. You, D. Kats, D. K. Li, O. Deshpande, H. Sachar, H. Xu, L. Laine, D. L. Shung, Detection of gastrointestinal bleeding with large language models to aid quality improvement and appropriate reimbursement, Gastroenterology (2024).

[4] X. Zhou, L. S. Dhingra, A. Aminorroaya, P. Adejumo, R. Khera, A novel sentence transformer-based natural language processing approach for schema mapping of electronic health records to the omop common data model, medRxiv (2024) 2024–03.

[5] H. Dong, J. Chen, Y. He, I. Horrocks, Ontology enrichment from texts: A biomedical dataset for concept discovery and placement, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 5316–5320.
[6] S. El-Sappagh, F. Franda, F. Ali, K.-S. Kwak, Snomed ct standard ontology based on the ontology for general medical science, BMC medical informatics and decision making 18 (2018) 1–19.