Merging and Validating Health Ontologies: A Framework for Ontological Integration and Evaluation

Safaa Menad[1,*], Saïd Abdeddaïm[1] and Lina F. Soualmia[1]

[1]Univ Rouen Normandie, Normandie Univ, LITIS UR 4108 F-76000 Rouen, France

Abstract

The integration of health ontologies is essential for addressing semantic heterogeneity across biomedical systems and enabling interoperability. In this study, we propose a structured and systematic framework for merging ontologies in the healthcare domain, with a focus on harmonizing terminologies derived from diverse sources. The resulting ontology was validated using Protégé, leveraging reasoning tools such as ELK to ensure logical consistency and detect potential conflicts. Furthermore, the ontology is now available on the BioPortal platform. This work demonstrates the potential of an integrated approach for enhancing biomedical data interoperability while maintaining high standards of consistency and usability.

Keywords

Biomedical Ontologies, Merging, Health, Language models

1. Introduction

Ontologies are fundamental tools in the biomedical domain, playing a pivotal role in enabling semantic interoperability across heterogeneous systems. They provide a structured and formal representation of domain-specific knowledge, supporting critical tasks such as clinical decision-making, information retrieval, and the integration of diverse biomedical datasets. Despite their immense utility, the proliferation of domain-specific ontologies, such as SNOMED CT and FMA, often leads to semantic inconsistencies, overlapping concepts, and redundancy. These challenges significantly hinder their effective deployment in real-world applications, where seamless and unified representations of knowledge are crucial.

To address these challenges, this paper proposes a structured approach for merging healthcare ontologies, aiming to align concepts, resolve semantic conflicts, and harmonize terminologies. The focus is placed on ontologies representing clinical terminologies, which are integral to many biomedical workflows. By integrating these resources, we aim to provide a cohesive and logically consistent knowledge base that supports interoperability and enhances usability in downstream applications.

The proposed framework involves validating the merged ontology using Protégé's reasoning capabilities, employing tools such as ELK to ensure logical consistency and detect conflicts. To further assess the utility of the merged ontology, it is uploaded in BioPortal. This evaluation ensures that the resulting resource is practically usable within the biomedical domain.

This paper is organized as follows: Section 2 provides an overview of related works, highlighting existing approaches and challenges in ontology merging. Section 3 details the methodology used to merge and validate ontologies. Section 4 presents the results of the evaluation. Finally, Section 5 concludes with key findings and potential directions for future research.

SWAT4HCLS 2025: The 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 24–27, 2025, Barcelona, Spain

*Corresponding author.

safaa.menad1@univ-rouen.fr (S. Menad); said.abdeddaim@univ-rouen.fr (S. Abdeddaïm); fatima.soualmia@univ-rouen.fr (L. F. Soualmia)

0009-0009-2204-7786 (S. Menad); 0000-0002-7521-7955 (S. Abdeddaïm); 0000-0001-7668-2819 (L. F. Soualmia)

2. Related Works

Ontology merging and alignment have been extensively studied in the biomedical domain to address the challenges posed by the increasing number of specialized and overlapping ontologies. Several works have proposed methods for semantic alignment of merged ontologies.

Prominent approaches for ontology alignment rely on techniques such as semantic similarity computation, machine learning, and rule-based reasoning. For instance, the AML (AgreementMakerLight) [1] tool has been widely used for ontology alignment in biomedical contexts, offering features like similarity metrics and filtering heuristics to improve matching accuracy. Similarly, LogMap [2] employs scalable reasoning-based techniques for aligning large biomedical ontologies and addressing inconsistencies during the merging process. These tools highlight the need for both semantic matching and logical consistency in ontology integration.

SAMBO [3] is a system specifically designed for aligning and merging biomedical ontologies in OWL format. It uses strategies such as linguistic matching, structure-based strategies, and machine learning algorithms to generate alignment suggestions. The system also checks for logical consistency and cycles after merging. SAMBO has been enhanced over time to incorporate session-based alignment, which allows user interaction at different stages, including preprocessing and validation, to improve the merging process.

Recent studies have leveraged machine learning and deep learning for ontology alignment. Sentence-BERT [4], a pre-trained transformer model for text similarity, has been adapted for ontology alignment tasks by generating embeddings for terms and concepts, enabling scalable and efficient matching. Our work builds on this approach by employing a domain-specific model, SBio_ClinicalBERT, fine-tuned on biomedical data, to improve alignment accuracy.

3. Methodology

3.1. Ontologies Selection

The selected ontologies were chosen as part of the PreDiBiontoL[1] project, which aims to predict diagnoses and recommend medications for patients. To our knowledge, there has been no study that aimed to enrich these specific ontologies in the same resource :

- The Human Disease Ontology (DOID) [5] serves as a comprehensive resource that classifies diseases and medical terminologies by integrating data from multiple external sources. Currently, it encompasses a total of 13,910 distinct concepts, making it a valuable tool for biomedical research.
- The Drug Ontology (DRON) [6] is a curated dictionary of molecular entities, focusing on "small" chemical components and drugs. It is developed using mappings from ChEBI (Chemical Entities of Biological Interest) and contains 8,282 unique concepts.
- The Symptom Ontology (SYMP)[2] is a standardized framework created to represent symptoms of human diseases. It provides definitions, labels, and synonyms for symptoms and comprises a total of 1,013 concepts.

3.2. Merging Process

The ontology merging process followed a structured pipeline, leveraging our proposed framework, SiMHOMER, developed in our study [7]. The merging workflow consisted of the following steps:

1. Data extraction : In this initial step, we extracted textual information from the concepts in the utilized ontologies. This included definitions, entities, and synonyms for all concepts across:
    - DOID (Disease Ontology): Definitions and entities were extracted to provide a foundation for disease-related concepts.

[1] Predicting Clinical Diagnosis using Biomedical Ontologies and Language Models
[2] https://www.ebi.ac.uk/ols4/ontologies/symp?viewMode=list

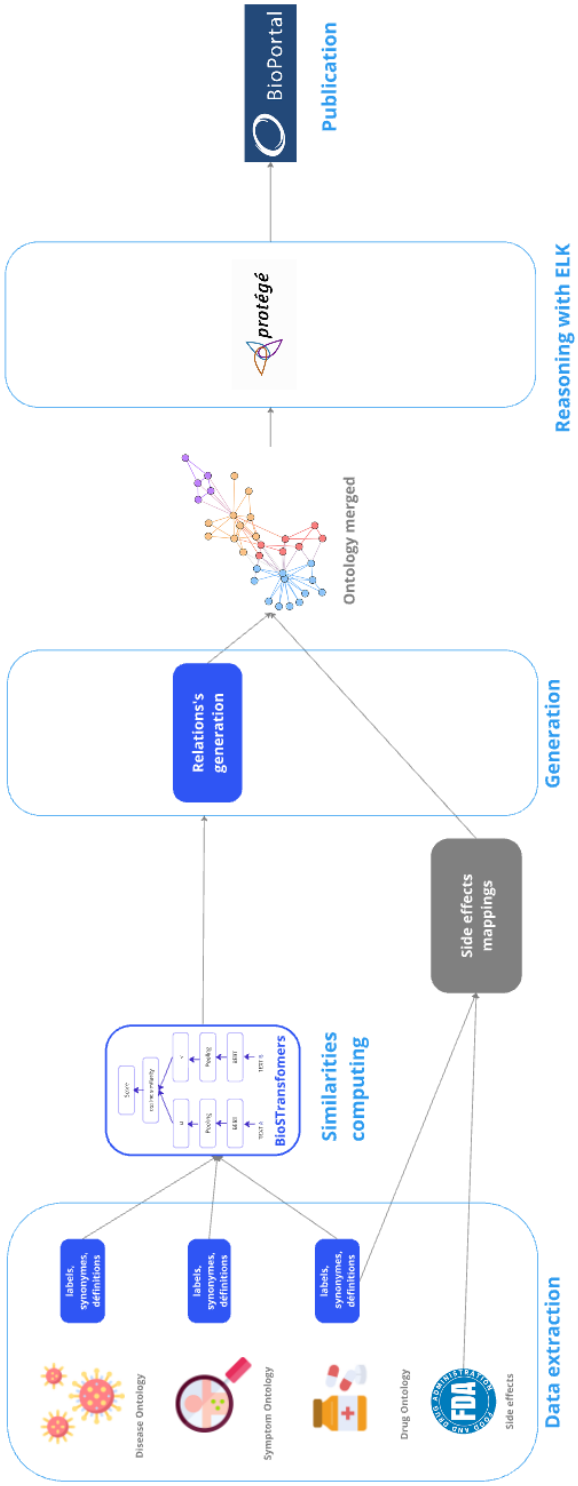

Figure 1: Framework overview

- Symptom Ontology (SYMP): Definitions and synonyms were obtained to describe clinical symptoms.
- Drug Ontology (DRON): We focused on extracting drug definitions and related data.

This information served as input for the subsequent similarity computation step.

2. Similarity computing : To calculate semantic similarities between concepts across ontologies, we employed SBio_ClinicalBERT, a customized sentence transformer model developed in our study

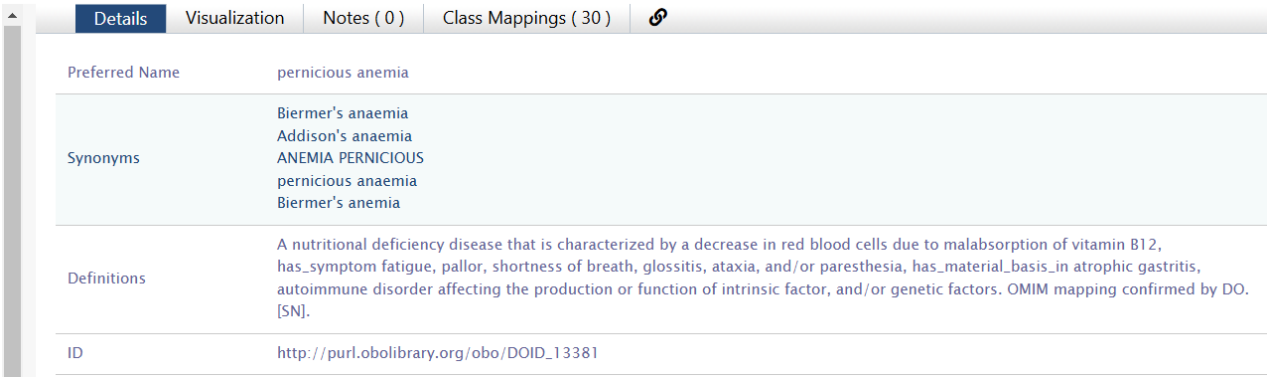

Figure 2: Example of data extracted from DOID ontology

[8].
Our model, inspired by Sentence-BERT, builds upon BERT by incorporating biomedical transformer models pre-trained on the PubMed corpus. This adaptation includes:

- A pooling layer to generate fixed-length embeddings for input text.
- A contrastive learning objective using the Multiple Negative Ranking Loss (MNRL) function[3].

Sentence transformers operate as siamese neural network models, designed to compare sentence pairs by generating embeddings that capture semantic meaning. The cosine similarity of these embeddings A and B reflects their semantic closeness (Equation 1). The training process minimizes the angular distance between embeddings of similar pairs, ensuring that semantically related sentences are mapped closer in the vector space.

$$\text{Sim}_{\cos}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \tag{1}$$

In this study, SBio_ClinicalBERT was pre-trained on pairs of PubMed data, mapping: (Titles, MeSH terms) and (Abstracts, MeSH terms) which is based on the Bio_ClinicalBERT [9] already trained on MIMIC III database.
This model was used to compute similarity scores between pairs of concepts from the three ontologies, enabling the identification of potential matches and relationships. Scores vary between 0 and 1.

3. Relation generation : In this step, we proposed new semantic relationships between concepts from the three ontologies:

   - DOID - SYMP: A new relation, *has_symptom*, was introduced to link diseases to their associated symptoms.
   - DOID - DRON: The relation *has_drug* was introduced to connect diseases to potential drugs for treatment.

   In this study, we retained the relations with a similarity score above 0.5. A more in-depth analysis to assess scores below 0.5 is currently in progress.
   To enrich our merged ontology, we also added new relations from OpenFDA[4] database that we named *has_side_effect* between each concept drug from DRON ontology and a side effect mapped from the database.
   By defining these relationships, we created a unified resource that integrates all three ontologies into a cohesive knowledge base. The new ontology imports the original ontologies and augments

[3]https://sbert.net/docs/package_reference/sentence_transformer/losses.html
[4]https://open.fda.gov/

Table 1
Number of proposed relations

Relation	Number
has_symptom	8726
has_drug	3921
has_side_effect	41118

them with the proposed relations to enhance semantic interoperability. These new relations were already validated with our previous work [8, 7] using external resources like UMLS and OpenFDA and also with soliciting health experts.

3.3. Validation Using Protégé Reasoner

After constructing the merged ontology, the next step is to validate the consistency using a reasoner. For that, we used Protégé[5]. The following steps were performed: **Ontology Loading:** where the newly created ontology, including all imported concepts and proposed relations, was loaded into Protégé. **Reasoning Activation:** where the ELK[6] reasoner was employed to classify the ontology and detect any inconsistencies. It was chosen because it was the fast one compared to Hermit. The validation process confirmed that the merged ontology was logically consistent, with no conflicts or errors identified during the merging process. This step ensured the robustness and quality of the resource.

3.4. BioPortal

To facilitate community use and further evaluation, the final merged ontology was uploaded to BioPortal[7], a platform widely used for hosting and sharing biomedical ontologies via this link : https://bioportal.bioontology.org/ontologies/DDSS. Additionally, this public availability allows researchers and developers to leverage the ontology for their applications, ensuring its practical impact and accessibility.

4. Results

The merged ontology resulted in a comprehensive structure with more than 50,000 new relations and a unified hierarchy integrating clinical and anatomical terms. Table 1 shows the number of the different relations that we proposed.

5. Conclusion

In this study, we proposed a framework for merging biomedical ontologies using our custom SBio_ClinicalBERT model. The framework facilitates the alignment of different ontologies by generating embeddings for ontology terms and leveraging their semantic similarity. The embeddings are used to compute a similarity matrix, which allows for the identification of mappings between concepts in different ontologies. This process was applied to three ontologies related to symptoms, diseases, drugs, and adverse events's database. The resulting merged ontology was further processed using Protégé, a widely used ontology management tool, to perform reasoning and check for inconsistencies. By leveraging Protégé's reasoning capabilities. This step ensured that the merged ontology was logically consistent and semantically robust. Future efforts will focus on enhancing the merging process through deep learning techniques and expanding the scope to include more ontologies, such as UMLS ontologies.

[5]https://protege.stanford.edu/
[6]https://www.cs.ox.ac.uk/isg/tools/ELK/
[7]https://bioportal.bioontology.org/

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, F. M. Couto, The AgreementMakerLight Ontology Matching System, in: R. Meersman, H. Panetto, T. Dillon, J. Eder, Z. Bellahsene, N. Ritter, P. De Leenheer, D. Dou (Eds.), On the Move to Meaningful Internet Systems: OTM 2013 Conferences, Springer, Berlin, Heidelberg, 2013, pp. 527–541. doi:10.1007/978-3-642-41030-7_38.

[2] E. Jiménez-Ruiz, B. Cuenca Grau, LogMap: Logic-Based and Scalable Ontology Matching, in: L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, E. Blomqvist (Eds.), The Semantic Web – ISWC 2011, Springer, Berlin, Heidelberg, 2011, pp. 273–288. doi:10.1007/978-3-642-25073-6_18.

[3] P. Lambrix, H. Tan, Sambo—a system for aligning and merging biomedical ontologies, Journal of Web Semantics 4 (2006) 196–206.

[4] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992.

[5] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, et al., Human disease ontology 2018 update: classification, content and workflow expansion, Nucleic acids research 47 (2019) D955–D962.

[6] J. Hanna, E. Joseph, M. Brochhausen, W. Hogan, Building a drug ontology based on rxnorm and other sources, Journal of biomedical semantics 4 (2013) 44. doi:10.1186/2041-1480-4-44.

[7] 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.

[8] S. Menad, W. Laddada, S. Abdeddaïm, L. F. Soualmia, Biostransformers for biomedical ontologies alignment, in: Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2023, Volume 2: KEOD, SCITEPRESS, 2023, pp. 73–84. doi:10.5220/0012188600003598.

[9] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78.