Ontology-Driven Approaches for Standardizing Rare Disease Terminology

Dr. Divya Pillai[1], and Dr. Sanjay Bhatia[2]

[1]*Amrita Institute of Medical Sciences, Kochi.*

[2]*Amrita Institute of Medical Sciences, Kochi.*

Received: 04/March/2024; Revised: 06/April/2024; Accepted: 06/May/2024; Published: 28/June/2024

Abstract

Rare diseases represent approximately 7000 disease entities that affect almost 3.5 % of the world population. Each rare disease receives inconsistent nomenclature from different medical systems which impedes research, diagnosis and treatment. In this paper, we study the problem of integrating rare disease terminologies using semantic models through an ontology driven approach. Using existing ontological datasets and ontology reasoners, we develop an upper ontology which automates the mapping and alignment of diseases' names and definitions within databases. The data interoperability, improved retrieve and enhanced framework adaptability presented in the results demonstrate the effectiveness of the system. This study presents an integration approach that gives precise and scalable solution for harmonization of rare disease informatics terminologies.

Keywords: Rare Diseases Ontology, Semantic Standardization, Harmonization, Medical Informatics, Data Interoperability, Automated Reasoning, Disease Ontology

1 INTRODUCTION

An ailment is called rare when their prevalence is less than one in two thousands persons. Individually, they are rare but there are over seven thousand such diseases occurring all over the world. While the informatics of rare diseases captures information and facilitates the organizations of rare diseases data, it still grapples with the phenomenon of data silos accessibility. One of which is ontological fragmentation due to unclassified nomenclature systems stems from heterogeneous diagnosis criteria and locale specific public health coding system. As previously mentioned, these overlapping terms lead to suboptimal results. Take, for instance, a single rare disease that could be referred to by different names within different databases, making its identification and analysis exceedingly difficult. As far as the systematization of medical terminology is concerned, there is a growing demand for a fully integrated approach system that focuses on the underlying semantics that is clear in recent years, especially with the advent of efforts directed towards individualized and precision approaches to medicine.

The rapid advancement in life sciences increases the volume and complexity of medical knowledge which makes it necessary to continually update the changes in the underlying concepts of medicine across varying fields and branches of medicine. Ontology driven approaches provide a structure of knowledge in a specific field using a controlled vocabulary and relationships through which domain

concepts can be arranged forming a hierarchy enabling formal representations. This elaborative language permits the synthesis of heterogeneous data sources into one cohesive body referred to as ontology which enables data to be synthesized from different sources. The opus of the Human Phenotype Ontology (HPO), Orphanet Rare Disease Ontology (ORDO), and SNOMED CT can serve as underlying biomedical ontologies.

Alongside these resources, there lies a problem regarding the integration, updating, and enhancing medical concepts and knowledge with ever-evolving medical knowledge terminologies. Not only have these most ontologies been developed in silos, creating gaps in usefulness resulting in poor reliability and validity, but also scope gaps have been created causing overlap, which blurs the distinction between these concepts. By using an ontological approach for unifying rare disease terminology, more effective data mining, seamless patient care through interoperable health records, and better inter and intra country collaborations can be attained.

The paper discusses a framework that uses existing disease ontologies, which are augmented with automated reasoning, to identify and correct semantic mismatches. These frameworks are constructed to be scalable and adaptable for different regions and institutions, thus harmonizing the terminology system. We evaluate its efficacy by conducting performance evaluations and comparing the results with traditional approaches. The following sections provide a thorough description of the research's system design, findings, and their relevance to healthcare.

2 LITERARY REVIEW

An increasing number of recent studies have been documented focusing on the utilization of ontology-based approaches for standardization of healthcare data, especially for the orphanet rare disease namespace. A few notable studies done in 2022 and 2023 demonstrate the progress in this area.

Li et al., (2022) focused on the issue of a semantic gap in multi-source healthcare data and proposed a hybrid ontology alignment model that raised both recall and precision for equivalent medical term identification. Their findings emphasized integrating a diverse range of lexical, structural, and domain-level information.

Wang & Zhou, (2023) looked into the incorporation of Orphanet's ORDO into clinical decision support systems. They noted a significant improvement in diagnosis of rare diseases, with ontology-driven systems increasingly identifying phenotypic overlaps and synonyms for diseases.

Silva et al., (2023) made another noteworthy contribution by developing a framework for cross-linking scarce disease registries in Europe using the HPO and SNOMED CT. This approach standardization of terminology enhanced the searchability of patient cohorts with comparable phenotypes (HPO 2023).

In NLP, Patel & Nguyen, (2022) advanced the study by proposing a transform model to map unstructured clinical text to rare disease ontologies. Their model appears to successfully tackle the

mapping of detailed relationships within complex diseases, but shortcomings of generalizability across language and institution were noted.

Fink et al., (2023) added volatility of ontology as an emergent building block of interoperability to the already existing discourse on the importance of applying the FAIR principles of Findable, Accessible, Interoperable, and Reusable in rare disease research. They promoted concepts of collaborative ontology curation where tools can be designed to maintain perpetual revision and alignment.

As a whole, these works emphasized the value of using ontology-based approaches to enhance the synthesis of fragmented data for precise diagnostics and seamless data interoperability. There are, however, still inconsistencies in the integration of competing ontologies along with constant modular frameworks. Our approach combines these solutions and proposes a dynamic bridge for ontology discrepancy based on reasoning.

3 METHODOLOGY

This work's system framework is built around four main modules: semantic reasoning, validation, conflict resolution, and ontology integration. The framework aims to align the different terminologies used across rare disease databases into a single ontology-based systematized representation.

1. Ontology Integration Module: The module downloads and parses disease ontologies like ORDO, HPO, and SNOMED CT. The system uses the OWL API to reconcile applicable class hierarchies directory merging synonym shared object properties.

2. Conflict Detection and Resolution: Conflicting definitions or labels for any one concept create conflict. A hybrid string similarity (Levenshtein distance), semantic distance (WordNet), and concept context using embedding models (BioBERT) are applied for conflict detection. Conflicts are scored and sorted in accordance with their confidence scores and then resolved based on an established rule-based hierarchy.

3. Reasoning Engine: Ensures all new assertions made in the ontology will not contradict the pre-existing axioms. An OWL reasoner, for example, HermiT, runs a consistency check and derives new equivalences. The module is also responsible for verifying that all assertions made after merging are logically valid and that subsumption relations, such as disease syndrome and symptom, are correctly inferred.

4. Validation and Output: We evaluate the harmonized ontology by applying it to several clinical cases and checking whether there is enhanced resolution of terms (i.e., better recognition of diseases across records). The ontology is output in RDF and JSON-LD formats for system interoperability.

The system was built with Protégé 5.5 and was tested against 500 rare disease entries from Orphanet as well as clinical data from MIMIC-III. Interoperability metrics were calculated through the

harmonized terminology comparing unstructured clinical records against a provided gold standard for annotation.

4 RESULTS AND DISCUSSION

Our system demonstrated significant improvements in terminology consistency and data interoperability.

Table 1: Term Resolution Accuracy Comparison

Method	Precision	Recall	F1-Score
Manual Curation	0.81	0.72	0.76
NLP-only (BERT)	0.84	0.78	0.81
Proposed Ontology System	0.91	0.88	0.89

Both manual and NLP based methods were outperformed by the proposed ontology-driven framework. The reason for this advantage is its capacity to resolve ambiguities of meaning associated with diseases and deduce relations that other methods based on string matching fail to identify.

Table 2: Performance Comparison of Ontology-Driven Methods

Method	Accuracy (%)	Speed (ms/query)	Adaptability (score out of 10)
Manual Curation	81	150	6
NLP-based (BERT)	84	100	7
Proposed Ontology System	91	80	9

These results showcase the reasoning engine's accuracy and tracking performance across different datasets. The resolution success rate was best in the Orphanet dataset and somewhat lower in unstructured clinical notes. Nevertheless, in all noted conditions, the improvement was statistically significant ($p < 0.01$).

Study Ontology-backed synergistic approaches ontologies enhance not only interoperability but also analyses such as patient similarity, phenotype-based diagnostic, and patient model simulations. The outlined issues are ontology drift, constant change mandates, and occasional ambiguity in the synonym mapping leading to ‘tissues’. These can be solved through collaborative spaces and automated updating systems.

5 CONCLUSION

Ontology methodologies formulate rare diseases' standards in a flexible and precise manner. Using formal semantics and automated reasoning allows our system to amalgamate heterogeneous terminologies through and beyond data interoperability and phenomenical analytics accuracy. Results have outperformed traditional methods, as well as NLP-based approaches, demonstrating greater

efficacy. The method provides equal relevance to any changes, additions, or removals of medical terms, thus improving research and clinical value. Work with regard to real-time electronic health census integration, multilingualism, and expansion towards treatment and drug related ontologies under rare disease informatics will be prioritized in precision care infrastructure initiatives.

REFERENCES

[1] Li, J., Zhang, Y., & Hu, W. (2022). Semantic integration in heterogeneous medical databases using hybrid ontologies. *Journal of Biomedical Informatics*, *128*, 104050.

[2] Wang, F., & Zhou, X. (2023). Enhancing rare disease diagnosis with integrated ontology-based decision systems. *Artificial Intelligence in Medicine*, *139*, 102460.

[3] Silva, R. et al., (2023). Harmonization of European rare disease registries through ontology linkage. *Orphanet Journal of Rare Diseases*, *18*(1), 113.

[4] Patel, A., & Nguyen, T. (2022). Mapping unstructured clinical narratives to rare disease concepts with deep learning. *Journal of Biomedical Semantics*, *13*(1), 22.

[5] Fink, D., Thomas, J., & Meyer, P. (2023). FAIR data principles in rare disease research. *Nature Reviews Drug Discovery*, *22*, 182–184.

[6] HPO Consortium. (2023). Human Phenotype Ontology update: 2023 release. *Nucleic Acids Research*, *51*(D1), D964–D969.