A Knowledge Graph for Enhanced Study Discovery and Semantic Exploration

Lea Gütebier[1,*], Dagmar Waltemath[1,2] and Ron Henkel[1]

[1]Medical Informatics Laboratory, Institute for Community Medicine, University Medicine Greifswald, Walther-Rathenau-Straße 48, 17475 Greifswald, Germany
[2]Data Integration Center, University Medicine Greifswald, Walther-Rathenau-Straße 49a, 17475 Greifswald, Germany

Abstract

A knowledge graph provides a powerful framework for integrating and linking diverse study resources, offering enhanced capabilities for study discovery and exploration. We created a semantically enriched graph that facilitates the retrieval and understanding of clinical studies by consolidating data from ClinicalTrials.gov, a German portal for medical data models, and ontologies such as UMLS and MeSH. Built on a Neo4j graph database and domain-specific ETL processes, the knowledge graph allows for efficient targeted searches and semantic exploration of clinical studies. By leveraging ontological annotations, it connects studies across domains, revealing relationships and providing a comprehensive context for research. This approach significantly enhances the accessibility, findability, and usability of clinical studies for patients, researchers, and clinicians alike.

Keywords

Knowledge graph, clinical studies, FAIR Principles, study discovery, semantic integration

1. Introduction

The ability to efficiently locate and select relevant clinical studies is essential for advancing both medical research and patient care. Researchers and clinicians rely on access to suitable studies for designing new protocols, identifying participants, and situating their findings within the state of the art.

However, study information is often distributed across a wide array of resources, including registries for clinical studies, medical metadata repositories, and ontological frameworks. This poses significant challenges for targeted searching, making the process of finding, retrieving, and exploring studies labor-intensive and inefficient.

To address these issues, we introduce a comprehensive framework for integrating complementary study resources into a single, unified knowledge graph. This approach not only consolidates data from diverse resources but also employs advanced graph representations for data and metadata to enable efficient exploration and retrieval. Here, we provide a detailed explanation of the meta-graph design and the graph representation of the integrated resources, laying the foundation for an accessible and interoperable study discovery platform.

2. Knowledge graph design

The knowledge graph is implemented using a Neo4j graph database and is constructed through an ETL (Extract, Transform, Load) process that ensures seamless data integration. Data extraction involves retrieving information from source repositories, followed by transformation according to a labeled property graph. This model specifies how data is represented: nodes depict entities such as studies or medical terms, while edges represent the relationships between them. Once transformed, the data is loaded into the graph database, enabling efficient storage and retrieval.

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
lea.guetebier@uni-greifswald.de (L. Gütebier); dagmar.waltemath@uni-greifswald.de (D. Waltemath); ron.henkel@uni-greifswald.de (R. Henkel)
0000-0001-5504-5108 (L. Gütebier); 0000-0002-5886-5563 (D. Waltemath); 0000-0001-6211-2719 (R. Henkel)

Clinical studies integrated into the knowledge graph originate from two key sources. Data from ClinicalTrials.gov [1] provides detailed information on study protocols. Additional data is incorporated from medical data models available via the MDM Portal [2], which include case report forms, structured eligibility criteria, data dictionaries, and base datasets. These resources are semantically annotated using Medical Subject Headings (MeSH) [3], respectively the Unified Medical Language System (UMLS) [4], ensuring consistency and interoperability.

By linking studies from different resources together through ontological information, the knowledge graph provides a more comprehensive understanding of the relationships between studies. The integration of ontologies such as the UMLS and MeSH enriches the graph's semantic context, allowing studies to be explored through deeper connections. This linking enables us to connect studies from ClinicalTrials.gov and medical data models from the MDM Portal by leveraging their respective annotations. The UMLS, as a comprehensive thesaurus of biomedical ontologies and vocabularies, includes MeSH and facilitates cross-referencing between these resources. This integration provides a cohesive framework for connecting and exploring study data across multiple resources.

By consolidating these resources, the knowledge graph establishes an integrated platform for exploring and understanding study data. It simplifies the process of finding studies while offering insights into their semantic relationships, thus improving accessibility and usability for patients, researchers, and clinicians alike.

The poster will depict the structure of this knowledge graph, emphasizing the integration and linking of diverse data sources. It will detail the ETL process, focusing on how data from ClinicalTrials.gov is transformed into a graph structure and how study annotations are mapped to ontologies. Additionally, it will showcase the graph schema of medical data models from the MDM Portal. For example, the representation of structured eligibility criteria and their connections to respective studies from ClinicalTrials.gov will be elucidated.

3. Discussion and Conclusions

Our approach contributes to the findability of studies, aligning with the FAIR guiding principles for data stewardship [5]. By facilitating easy access to and efficient traversal of the integrated data, the knowledge graph enables targeted searches through its ontological annotations. This capability enhances the semantic exploration and retrieval of studies, creating a more effective platform for study discovery. Overall, the knowledge graph promotes efficient search, exploration, and integration of study data, addressing challenges in accessibility and usability.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, N. C. Ide, The clinicaltrials. gov results database—update and key issues, New England Journal of Medicine 364 (2011) 852–860.

[2] S. Riepenhausen, M. Blumenstock, C. Niklas, S. Hegselmann, P. Neuhaus, A. Meidt, C. Püttmann, M. Storck, M. Ganzinger, J. Varghese, et al., Europe's largest research infrastructure for curated medical data models with semantic annotations, Methods of Information in Medicine (2024).

[3] C. E. Lipscomb, Medical subject headings (mesh), Bulletin of the Medical Library Association 88 (2000) 265.

[4] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.

[5] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.