Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240689

From Clinical Information Systems to Personalized Health Knowledge Graphs

Sareh AGHAEI[a], Remzi CELEBI[b], Markus KREUZTHALER[a,1] and Stefan SCHULZ[a]
[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[b] *Institute of Data Science, Faculty of Science and Engineering, Maastricht University, The Netherlands*

Abstract. This paper presents a versatile solution to formally represent the contents of electronic health records. It is based on the knowledge graph paradigm, and semantic web standards RDF and OWL. It employs the established semantic standards SNOMED CT and FHIR, which warrant international interoperability. A graph-based form is not only useful to feed different target visualizations, but it can also be subject to AI-powered services such as (fuzzy) retrieval and summarization.

Keywords. SNOMED CT, RDF Knowledge Graphs, EHR, Interoperability

1. Introduction

Electronic health records (EHRs) contain a full range of patient data, both structured and unstructured, which is crucial for improving patient care, clinical decision-making, and healthcare service delivery. To support EHR interoperability, multiple ontologies have been introduced over the decades. SNOMED CT stands out as the most comprehensive clinical healthcare ontology, focusing on logical consistency and the representation of about 350,000 concepts within multiple hierarchies.

Knowledge graphs (KGs) in conjunction with semantic standard data models based on RDF (Resource Description Framework) and ontologies such as SNOMED CT using OWL (Ontology Web Language) bear the promise of ensuring interoperability and compatibility across different healthcare systems. This paper describes RDF KGs as a canonical form for representing EHR content, with a particular focus on tabular data as part of the structured data within EHRs, utilizing SNOMED CT as the underlying ontology.

The main research questions in response to this study's objective are summarized as follows: (1) How to annotate column headers in tabular data, which often lack informative documentation in real clinical settings? (2) How to extract a smaller portion of SNOMED CT for a given use case to facilitate tasks such as navigation, reasoning, sharing, and integration?

[1] Corresponding Author: Markus Kreuzthaler; E-mail: markus.kreuzthaler@medunigraz.at.

2. Materials and Methods

Data: To showcase this work, we choose a use case from the AIDAVA project [1], focusing on breast cancer disease. De-identified data collected for this use case comes from an Austrian hospital information system, rendered in tabular format encompassing medical history, progress notes, prescribed medications, pathology reports, surgical procedure descriptions, multidisciplinary meeting reports, and TNM staging.
Tabular Data Annotation: Semantic annotation of a column in tabular data involves identifying SNOMED CT concepts that capture the semantics of the data. In real-world scenarios, the documentation of database columns often lacks detail, which raises both syntactic and semantic challenges. Our approach to this problem is based on prompting ChatGPT to obtain contextual hints and collaborating with domain experts and data owners. Moreover, lexically searching terms within the SNOMED CT browser is performed, following guidelines established in our previous work [2].
***Sub-ontology Extraction*:** To extract a small and relevant portion of SNOMED CT (referred to as the module within the SNOMED CT community) given a use case, we employ the personalized page rank (PPR) technique for obtaining the most relevant concepts as a graph, taking into account the concepts mapped to the column headers in the tabular data annotation step. The PPR computes scores of node importance in the graph based on a personalized vector biasing toward the query nodes (i.e., the mapped concepts or the signature within the SNOMED CT community). We select the nodes with the highest scores to shape the sub-ontology for the given use case.
Evaluation: The evaluation of KG quality is multidimensional, and we use two main dimensions, including coherency and coverage. For the coherency dimension, consistency is measured to demonstrate that the constructed KGs are free of logical contradictions using SHACL shapes. For the coverage dimension, completeness is evaluated through queries along with their expected results provided by domain experts and data owners.

3. Results and Outlook

Each patient has their own graph known as personalized health KG (PHKG). We store RDF triples of patients in distinct named graphs in GraphDB, resulting in a set of triples with a unique name (i.e., a named graph) associated with a PHKG. This study addresses the imperative demand for interoperable representation of EHRs through the construction of RDF KGs, utilizing SNOMED CT. In line with the terminology box (TBox) and the assertions box (ABox), the extracted sub-ontology corresponds to the TBox, while representing tabular data in terms of the TBox pertains to ABox. The future direction is to integrate unstructured portions of EHRs into PHKGs. This research has received funding from the European Union's Horizon Research and Innovation Programme under grant agreement No 101057062 (AIDAVA, https://aidava.eu/).

References

[1] AI-powered Data Curation & Publishing Virtual Assistant, https://www.aidava.eu
[2] Schulz S, et al. Towards principles of ontology-based annotation of clinical narratives. International Conference on Biomedical Ontologies. 2023