MeDaX Prototype v0.2

Ilya Mazein[1], Sarah Braun[1,2], Tom Gebhardt[1], Ron Henkel[1], Lea Michaelis[1], Dagmar Waltemath[1,2] and Judith A.H. Wodke[1]

[1]*Medical Informatics Department, Institute of Community Medicine, University Medicine Greifswald*
[2]*Core Unit Data Integration Center, University Medicine Greifswald*

Abstract
The MeDaX project aims to develop and implement concepts and tools for bioMedical Data eXploration using graph technologies. Here, we present v0.2 of our prototype, representing FHIR formatted clinical data in a graph format. We build on the pre-existing CyFHIR tool for generic conversion, optimise the resulting graph structure to lessen complexity, and incorporate the BioCypher framework, integrating the clinical data with ontology information. This makes the data more accessible and convenient for querying, information retrieval and analysis.

Keywords
Clinical data, graph database, knowledge graph, FHIR, Neo4j

1. Introduction

Accessibility of clinical data and its consequent analysis are widely discussed among researchers and clinicians. The Medical Informatics Initiative (MII) Germany aims to digitally provide reusable clinical data for research purposes (https://www.medizininformatik-initiative.de/en/start). The MeDaX project focuses on developing a resource for biomedical data exploration [1]. Here, we present a prototype that represents patient data in a graph format, beneficial for querying and analysis of complex heterogeneous data [2, 3].

2. Data formats

As part of the MII, healthcare data of university hospitals are collected, harmonised, semantically enriched, and then made available for research purposes by Data Integration Centers. To overcome the diversity of source formats and individual data processing pipelines, data is provided in a FHIR (Fast Healthcare Interoperability Resources) format. Hence, we base our prototype on synthetic data of the same format. For our graph we use Neo4j (https://github.com/neo4j/neo4j), one of the most popular and continuously maintained graph database systems currently available. We opt for a labelled property graph representation in place of the RDF format as it is less complex when it comes to storing large quantities of patient data and provides a more intuitive visualisation for clinicians.

SWAT4HCLS conference, February 26–29, 2024, Leiden, The Netherlands
ilya.mazein@uni-greifswald.de (I. Mazein)
0009-0000-1130-8332 (I. Mazein); 0000-0001-6211-2719 (R. Henkel); 0000-0001-9691-2677 (L. Michaelis); 0000-0002-5886-5563 (D. Waltemath); 0009-0009-9712-060X (J. A.H. Wodke)

3. Challenges and solutions

For generic conversion from FHIR to Neo4j we applied an existing project named CyFHIR, "a native Neo4j plugin that acts as the bridge between FHIR and Neo4j" (https://github.com/Optum/CyFHIR). It parses the tree-like structure of a FHIR resource JSON file, creating a corresponding Neo4j graph structure, regardless of the type of FHIR resource being used as input.

Due to the fact that CyFHIR builds a graph according to the raw internal structure of a FHIR resource, the resulting order of magnitude for nodes and relationships can be equal to 4 and higher. While most of the graph is relevant in its utility, we found some parts to be redundant. We optimised the graph structure, making it less complex and storage-intensive, while keeping relevant information intact.

1) References between various FHIR resources represent high-level connections between FHIR entities - a Patient and a Diagnosis they have, for example. By simplifying the structure of these connections we decreased the test patient graph size by ~25%.

2) In its raw graph representation the structure of a FHIR resource property can consist of multiple nodes and edges collectively describing the same feature. Condensation of property structures so far resulted in up to ~40% less elements in the graph. However, the optimal ratio of complexity to searchability remains to be determined.

3) To further organise entities after conversion and remove redundancies within the graph, we apply the BioCypher framework [4], integrating clinical data with ontological information. An overarching ontology will provide a unifying framework for incorporating various clinical data formats in addition to FHIR.

4. Conclusion

Our graph-based solution and planned clinical user interface will help to provide a unified way of access for researchers and direct interaction with enriched data for clinicians. Intuitive visualisation alongside structured graph database queries allows to organise information and answer treatment and research related questions more efficiently.

References

[1] J. A.H. Wodke, L. Michaelis, R. Henkel, The MeDaX knowledge graph prototype, Studies in Health Technology and Informatics (2023). doi:10.3233/SHTI230089.

[2] A. Lysenko, I. A. Roznovăţ, M. Saqi, A. Mazein, C. J. Rawlings, C. Auffray, Representing and querying disease networks using graph databases, BioData Mining (2016). doi:10.1186/s13040-016-0102-8.

[3] B.-H. Yoon, S.-K. Kim, S.-Y. Kim, Use of graph database for the integration of heterogeneous biological data, Genomics Informatics (2017). doi:10.5808/GI.2017.15.1.19.

[4] S. Lobentanzer, P. Aloy, J. Baumbach, et al, Democratizing knowledge representation with BioCypher, Nature Biotechnology (2023). doi:10.1038/s41587-023-01848-y.