Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230255

Exploring New Possibilities for Research Data Exploration Using the Example of the German Core Data Set

Felix MENZEL[a,b], Dagmar WALTEMATH[a,b] and Ron HENKEL[a,1]
[a] *Department of Medical Informatics, Institute for Community Medicine, University Medicine Greifswald, Germany*
[b] *Core Unit Data Integration Center, University Medicine Greifswald, Germany*

Abstract. The German Medical Informatics Initiative (MII) aims to increase the interoperability and reuse of clinical routine data for research purposes. One important result of the MII work is a German-wide common core data set (CDS), which is to be provided by over 31 data integration centers (DIZ) following a strict specification. One standard format for data sharing is HL7/FHIR. Locally, classical data warehouses are often in use for data storage and retrieval. We are interested to investigate the advantages of a graph database in this setting. After having transferred the MII CDS into a graph, storing it in a graph database and subsequently enriching it with accompanying meta-information, we see a great potential for more sophisticated data exploration and analysis. Here we describe the extract-transform-load process which we set up as a proof of concept to achieve the transformation and to make the common set of core data accessible as a graph.

Keywords. Core Data Set, Graph Database, Data Integration, HL7/FHIR, Data Exploration

1. Introduction

The Medical Informatics Initiative (MII) [1] has established data integration centers (DIZ) at all university hospitals in Germany. Together, the large consortium has agreed on a core data set (CDS)[2] to enable data sharing for research purposes. The core data set is divided in six modules (person, case, diagnosis, procedure, laboratory test results and medication). The agreed-upon interface for data exchange across the DIZs is HL7/FHIR [2]. However, the local data stores differ across the university hospitals, e.g., due to pre-existing IT-infrastructures.
As part of the HealthECCO[3] community effort we contributed to the development of CovidGraph [3] and thereby demonstrated the power of graph databases for fast and flexible biomedical data exploration during the COVID-19 pandemic.

[1] Corresponding Author: Ron Henkel, Walther-Rathenau-Str. 48, Department of Medical Informatics, University Medicine Greifswald, D-17475 Germany; E-mail: ron.henkel@uni-greifswald.de.
[2] CDS: https://www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set
[3] HealthECCO: https://healthecco.org/

This work elucidates the potential of transforming and storing the MII CDS into a graph database to facilitate easy exploration using SemSpect [4] and to enable domain-spanning queries [5] in the broad field of medical informatics.

2. Data and Methods

The development of our **e**xtract, **t**ransform, **l**oad (ETL) process started with a detailed analysis of the MII CDS. We then evaluated commonly used ETL tools [6,7]. Based on this analysis, a graph model is developed and accompanying meta-information is identified. The data is accessed via the HL7/FHIR interface, mapped to the graph model, and loaded into the graph database[4]. In a last step, accompanying meta information, related clinical studies and ontology terms, will be loaded, cross-referenced and linked to the CDS.

3. Results and Discussion

Storing and linking medical data in a graph database opens new doors for data exploration. For example, SemSpect [4] supports expansion and filtering, and automatic grouping of similar data items. Taken together, these features enable to easily traverse the graph. Visual representations can be created without detailed knowledge of the underlying data model. In addition, Neo4j Bloom[5], an application for graph exploration, offers semi-natural language queries, rule-based styling and allows to search for phrases and pattern. Results from the different exploration steps can be visualized or accessed in tabular format as well as attribute-value pairs. Future plans include the implementation of sophisticated data analysis procedures using graph algorithms. As the core data set contains medical data, this proof of concept will obey to the same data access restrictions as the original data set.

References

[1] Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med. 2018;57(S 01):e50–6. http://dx.doi.org/10.3414/ME18-03-0003.
[2] Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: Systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform. 2021;9(7):e21929. http://dx.doi.org/10.2196/21929.
[3] Gütebier L, Bleimehl T, Henkel R, Munro J, Müller S, Morgner A, et al. CovidGraph: a graph to fight COVID-19. Bioinformatics. 2022;38(20):4843–5. http://dx.doi.org/10.1093/bioinformatics/btac592.
[4] Liebig T, Vialard V, Opitz M. Connecting the Dots in Million-Nodes Knowledge Graphs with SemSpect. ISWC Posters Demos Ind. Tracks, 2017, p. Vol-1963:paper587.
[5] Henkel R, Wolkenhauer O, Waltemath D. Combining computational models, semantic annotations and simulation experiments in a graph database. Database (Oxford). 2015;2015. http://dx.doi.org/10.1093/database/bau130.
[6] Pall AS, Khaira JS. A comparative review of extraction, transformation and loading tools. Dbjournal.ro. http://dbjournal.ro/archive/12/12_5.pdf.
[7] Cheng KY, Pazmino S, Schreiweis B. ETL processes for integrating healthcare data - tools and architecture patterns. Stud Health Technol Inform. 2022;299:151–6.

[4] Neo4j: https://neo4j.com

[5] Neo4j Bloom: https://neo4j.com/docs/bloom-user-guide/current/