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

doi:10.3233/SHTI240423

MeDaX: A Knowledge Graph on FHIR

Ilya MAZEIN[a,1], Tom GEBHARDT[a,1], Felix ZINKEWITZ[b], Lea MICHAELIS[a], Sarah BRAUN[a], Dagmar WALTEMATH[a,b], Ron HENKEL[a] and Judith A.H. WODKE[a,2]
[a] *Department of Medical Informatics, University Medicine Greifswald, Germany*
[b] *Core Unit Data Integration Center, University Medicine Greifswald, Germany*
ORCiD:I. Mazein https://orcid.org/0009-0000-1130-8332, F Zinkewitz https://orcid.org/0009-0007-4377-8778, L. Michaelis https://orcid.org/0000-0001-9691-2677, S. Braun https://orcid.org/0009-0004-7699-0554 , D. Waltemath https://orcid.org/0000-0002-5886-5563, R. Henkel https://orcid.org/0000-0001-6211-2719, J. Wodke https://orcid.org/0009-0009-9712-060X

Abstract. In Germany, the standard format for exchange of clinical care data for research is HL7 FHIR. Graph databases (GDBs), well suited for integrating complex and heterogeneous data from diverse sources, are currently gaining traction in the medical field. They provide a versatile framework for data analysis which is generally challenging for raw FHIR-formatted data. For generation of a knowledge graph (KG) for clinical research data, we tested different extract-transform-load (ETL) approaches to convert FHIR into graph format. We designed a generalised ETL process and implemented a prototypic pipeline for automated KG creation and ontological structuring. The MeDaX-KG prototype is built from synthetic patient data and currently serves internal testing purposes. The presented approach is easy to customise to expand to other data types and formats.

Keywords. Clinical research data, FHIR, knowledge graph, ETL, MeDaX

1. Introduction

The Medical Informatics Initiative (MII) aims to improve and promote healthcare research in Germany [1]. Data integration centres (DIZ) were set up at university clinics to provide clinical data for secondary use in research according to specified standards. A core data set (CDS) describes data records from routine clinical practice to be captured at the DIZ [1,2]. Data are shared with external researchers upon successful usage applications, according to standard specifications for HL7 FHIR [3]. Challenges in handling clinical healthcare data are, e.g., the large number of heterogeneous source systems and the lack of standardised technical solutions for data analysis, management, and storage [4]. GDBs are highly suited to handle complex interconnected and diverse data and querying is often more efficient than for relational databases [5, 6].
At MeDaX, we develop innovative methods and tools for bio**Me**dical **Da**ta e**X**ploration with graph technologies. Here, we present a modular method for automated generation of GDBs from FHIR-formatted clinical research data. We implement a proof-of-concept

[1] Authors contributed equally to this work
[2] Corresponding Author: Judith Wodke; E-mail: judith.wodke@uni-greifswald.de.

pipeline comprised of a generic ETL process and a harmonisation module. This pipeline was applied to synthetic patient data to generate a prototypic MeDaX-KG.

2. Methods

Synthetic data generation. Synthetic patient data for development and testing was generated using Synthea [7], a synthetic patient population simulator. Using its default settings, we generated a test population of *Patients* with multiple clinical entities such as *Observations*, *Conditions*, and *Encounters* describing them. Each patient bundle contains a *Patient* resource and hundreds of FHIR resources connected to it.

Resource-specific ETL pipeline: from FHIR to KG. We use a data schema for a GDB derived from a careful manual inspection of available FHIR-formatted research data at the DIZ of the University Medicine Greifswald (UMG) [8]. A central, resource-agnostic download module was created using the "fhirclient" Python module from Boston Children's Hospital[3]. Resource-specific transformation modules have been implemented in Python for 7 of the 157 FHIR resource types, namely Condition, Diagnosis, Report, Encounter, Observation, Organisation, Patient, and Procedure, selected based on use frequency in the analysed input data set. Transformation modules, which include the initialization function to prepare the database and the main transformation function, also incorporate a utility module to handle data types, improving maintainability and allowing for easy tool adaptation. Transformed data is stored in a Neo4j graph database via a storage module, using the Neo4j Python driver. Source code is available at Github[4].

Generic ETL pipeline: from FHIR to KG. For the generic ETL process, we apply CyFHIR[5], "a native Neo4j plugin that acts as the bridge between FHIR and Neo4j". 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. The automatically generated graph is post-processed to remove redundant intermediate nodes and relations between FHIR resources linked through FHIR-internal References. Nodes and relations collectively describing the same data feature are condensed into a single node. Code is written in Python.

Harmonisation module. BioCypher (BC), a harmonising framework for standardised KG creation, is applied to transform provided data into an ontology-based structure [9]. By default, it utilises BioLink, a high-level, open-source data model designed to standardise types and relationships in biological KGs [10]. Its modular approach facilitates integration of diverse input data and customisation of the data model. Reusing and adjusting the input adapter for the clinical knowledge graph [11], we defined a BC input adapter for a running Neo4j GDB. In addition, we implemented a feature to automate incorporation of new node and relationship types into the schema YAML file based on the given input data.

[3] https://github.com/smart-on-fhir/client-py

[4] https://github.com/fznkw/fhir2neo4j

[5] https://github.com/Optum/CyFHIR

3. Results

We developed a flexible and low maintenance method for creating an integrated KG for clinical research data and implemented a respective prototype. The workflow of our novel method consists of independent modules that, when executed in consecutive order, generically build a standardised structured KG for clinical research data (Figure 1, top half). CyFHIR transforms the synthetic patient data into a graph instance. This initial graph is vast, spanning up to tens of thousands of nodes, and reveals two obvious undesired side-effects, i) non-informative intermediate graph entities and ii) redundant information. We implemented a module to correct for those two side-effects. Removing intermediate nodes from *FHIR Reference* connections reduces the complexity by approximately 20%. Roughly another 50% are reduced by condensing multiple entities into a single node if they collectively describe the same feature of a resource, e.g., by encoding the same information using two different terminologies. To further structure and enrich our post-processed KG, we joined the BC project [9]. Reusing and adjusting the input adapter of the Clinical Knowledge Graph [11], the MeDaX-KG is structured according to the BioLink data model [10]. We hard-coded frequently used node types and relation types as a base schema.

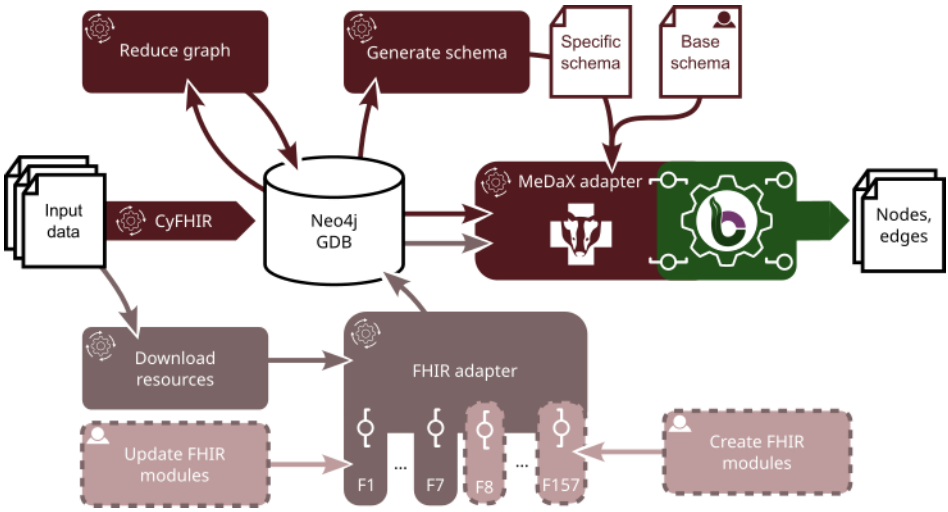

Figure 1. Workflow schema of the ETL processes for conversion of FHIR-formatted clinical research data into an ontology-enriched KG. Top half (dark red): **MeDaX pipeline:** Input data is loaded into a Neo4j GDB with CyFHIR. Afterwards, the *reduce graph* module optimises the graph by removing redundancies and non-informative entities. A data specific schema is generated, combined with our base schema, which covers commonly used nodes and edge types, and utilised in the MeDaX adapter; lower half (brown): **FHIR resource type-specific pipeline:** The input data is downloaded and used as an input for the FHIR adapter. The adapter is extended with FHIR [3] type-specific transformation modules. One module has to be created for each FHIR resource type. Currently seven modules out of 157 are covered. Modules have to be updated if a FHIR resource type changes; The MeDaX adapter connects the GDB with the BioCypher framework [9] (green) to generate an integrated and structured KG.

To assure integration of the complete specified input data we implemented a script that generates an additional resource-specific schema. Both schemas are combined and users are informed about added entity types and how to specify them further. To evaluate the feasibility of KGs for clinical research data, we compared our novel approach with the

resource-specific ETL (Figure 1, lower half). The resource-specific ETL process results in a clear and subject-focused representation of the graph structure, while our generic approach seamlessly integrates the complete scope of possible input FHIR data without any additional effort.

In summary, we designed and prototypically implemented a sustainable generic ETL process for the representation of FHIR-formatted clinical research data in graph format.

Even before incorporating multiple data sources, the MeDaX pipeline prototype effectively validates the efficacy of our design principles, which aim to i) enhance standardization and reproducibility of the produced knowledge graphs, ii) reduce maintenance expenses, and iii) boost feature reusability via a modular approach, affirming the robustness and strategic foresight of our design methodology.

4. Discussion and Outlook

We present the first generalised concept and prototype for an automated transformation of FHIR-formatted clinical research data, and thus of the MII CDS, into a GDB and its subsequent automated semantic enrichment. The graph format allows for an intuitive perception of data and their connectivity. The chosen approach to combine and reuse existing tools, such as CyFHIR and BC, with our own customisation and automation features, assures standardisation and reproducibility of the generated KGs and allows to account for specific constraints when reusing clinical care data in research.

Clinical data is heterogeneous in terms of types and content. Data collections of different MII DIZ and other clinical data holding institutions differ in size and scope. Due to the sensitivity of clinical data, data holders do not usually grant easy access to their data collections. Accordingly, to not interfere with data protection, MeDaX-KGs have to be set up and maintained locally by the data holders. Therefore, minimisation of maintenance costs and maximisation of user support are key aspects of the project and a fully automated transformation of FHIR resources is an essential requirement.

Our approach maximises sustainability of the tool with respect to future changes and expansion of FHIR resource types. Applicability to different FHIR data collections with full input data coverage is assured by the automatic schema generation. Nodes and edges required by the input data but not yet included in the BC base schema provided in the MeDaX pipeline, are generically included by updating the base schema file. This combination of hard-coded graph entity types in the base schema with automated generation of new ones on demand allows to account for the most widely used FHIR resource types with utmost care, while preventing loss of information due to unexpected input data. To optimise user experience, logs inform the user about the added graph entities and where and how to customise them if needed.

Initial analysis of the MeDaX-KG prototype highlights considerable opportunities for enhancing its graph structure and knowledge representation through optimization. In an ongoing effort, the lessons learnt serve to improve the different pipeline modules. The currently used BioLink data model [10] focuses on biological not medical terms. We will overcome this limitation by integrating the biomedical resource ontology (BRO) [12] into our data schema. A graphical user interface will provide simple features for data visualisation and exploration. The two post-processing steps of the MeDaX-KG reduce graph complexity by more than 50%. We anticipate that further investigation will reveal additional potential for structure optimisation and we will iteratively improve our ETL process, working towards an optimal subject-focused representation. To this end, we will

compare the resulting graph structures to the gold standard schema from the resource type-specific ETL process. Since a major limitation of our prototype is the usage of synthetic data, in a next step, the pipeline will undergo rigorous testing within the UMG DIZ productive environment to ensure its robustness and reliability in real-world settings. In this context, we will closely cooperate with the actual users of our tool to set up an appropriate user access control, supporting the responsible handling of sensitive data and ensuring patient safety and trust.

The interest in reusing clinical care data for research is. Critical insights into clinical data management often emerge through the implementation of foundational infrastructures. FHIR, an established standard format for the exchange of electronic health data, is not optimal for data visualisation and exploration. The proposed GDB model suits complex, interconnected, heterogeneous data while remaining accessible and intuitively understandable to various stakeholders and domain experts in biomedicine. Standardisation and semantic enrichment will improve FAIRness of provided research data and lead to better reusability and reproducibility of scientific results [13]. Based on user-friendliness, scalability and broad applicability of the MeDaX pipeline, this tool has the potential to substantially advance clinical research data representation and analysis.

Acknowledgements: This work has been funded by the BMBF, FKZ 01ZZ2019.

References

[1] Semler SC, et al. German Medical Informatics Initiative – A National Approach to Integrating Health Data from Patient Care and Medical Research. Methods Inf Med. 2018;57(S 01):e50-6.

[2] Ganslandt T, Boeker M, Lobe M, Prasser F, Schepers J, Semler S, et al. Der Kerndatensatz der Medizininformatik-Initiative: Ein Schritt zur Sekundarnutzung von Versorgungsdaten auf nationaler Ebene. In: Forum der Medizin-Dokumentation und Medizin-Informatik 2018;20:17.

[3] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. IEEE 2013;326-31.

[4] Park Y, Shankar M, Park BH and Ghosh J. Graph databases for large-scale healthcare systems: A framework for efficient data management and data services. IEEE 30th International Conference on Data Engineering Workshops, Chicago, IL, USA 2014;12-19, doi: 10.1109/ICDEW.2014.6818295.

[5] Timón-Reina S, Rincón M, Martínez-Tomás R. An overview of graph databases and their applications in the biomedical domain. Database (Oxford) 2021;2021:baab026.

[6] Walke D, et al. The importance of graph databases and graph learning for clinical applications. Database (Oxford) 2024;2024:baad045, doi: 10.1093/database/baad045.

[7] Walonoski J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230-238, doi: 10.1093/jamia/ocx079. Erratum in: J Am Med Inform Assoc. 2018;25(7):921.

[8] Menzel F, Waltemath D, Henkel R. Exploring New Possibilities for Research Data Exploration Using the Example of the German Core Data Set. Stud Health Technol Inform. 2023 May 18;302:749-750. doi: 10.3233/SHTI230255. PMID: 37203485.

[9] Lobentanzer S, Aloy P, Baumbach J, Bohar B, Danhauser K, Dogan T, et al. Democratizing Knowledge Representation with BioCypher. Nat Biotech 2023;41:1056-1059.

[10] Unni, D., et al. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. Clinical And Translational Science 2022; 1848–1855, doi: 10.1111/cts.13302.

[11] Santos, A., Colaço, A.R., Nielsen, A.B. et al. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol 2022;40:692–70, doi: 10.1038/s41587-021-01145-6.

[12] Tenenbaum, J. D., Whetzel, P. L., Anderson, K., Borromeo, C., Dinov, I. D., Gabriel, D., et al. The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. Journal Of Biomedical Informatics, 2011; 44(1):137–145, doi: 10.1016/j.jbi.2010.10.003.

[13] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3.