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


*doi:10.3233/SHTI230089*

# The MeDaX Knowledge Graph Prototype

Judith A.H. WODKE[a,1], Lea MICHAELIS[b] and Ron HENKEL[a]
[a] *Department of Medical Informatics, University Medicine Greifswald, Germany*
[b] *Core Unit Data Integration Center, University Medicine Greifswald, Germany*
ORCiD ID: Lea Michaelis https://orcid.org/0000-0001-9691-2677

**Abstract.** Data sharing is sustainable for several reasons, including minimising economical and human costs or maximising knowledge gain. Still, reuse of biomedical (research) data is often hampered by the diverse technical, juridical, and scientific requirements for biomedical data handling and specifically sharing. We are building a toolbox for automated generation of knowledge graphs (KGs) from diverse sources, for data enrichment, and for data analysis. Into the MeDaX KG prototype, we integrated data from the core data set of the German Medical Informatics Initiative (MII) with ontological and provenance information. This prototype is currently used for internal concept and method testing only. In subsequent versions it will be expanded by including more meta-data and relevant data sources as well as further tools, including a user interface.

**Keywords.** Knowledge graphs, biomedical data, data enrichment, data reuse, open source, MeDaX

## 1. Introduction

The MII [1], aiming at digitisation of health care in Germany, follows a federated storage approach: Every German university clinic has set up a data integration center (DIZ) that, based on a common core data set (CDS) [2], provides digital solutions for biomedical data management. And while the CDS itself is standardised in HL7/FHIR [3] format, technical solutions provided for data storage, management, and analysis often are not.

KGs have proven suitable for representation of complex heterogeneous data [4,5]. Within the MeDaX project, we are building a toolbox for bio**Me**dical **Da**ta e**X**ploration. This includes innovative and efficient methods for data harmonisation and storage in KGs, for data enrichment, analysis, and retrieval. The presented first prototype serves internal testing purposes. With the entire project, apart from data FAIRification [7] and improving reuse opportunities for biomedical data, we aim at informing and empowering several stakeholders at the same time: medical personnel, scientists, and the public, including the patients the data might originate from.

## 2. Methods

Aligning with the federated storage approach of the MII, the MeDaX toolbox will be applied locally to enrich and integrate available health care data with data from other sources, including biomedical ontologies [6] and public databases. For the current

[1] Corresponding Author: Judith Wodke, E-mail: judith.wodke@uni-greifswald.de.

prototype, dummy FHIR resources are transferred into a Neo4J KG using BioCypher [8] in strict mode. This first implementation serves as an internal proof of concepts and allows us to start developing analysis and querying tools.

## 3. Results

The MeDaX KG prototype is work in progress and includes dummy FHIR resources representing data from the MII CDS basic modules [2] plus semi-automatically added ontological [7] and provenance information. In addition, converting different data sources into RDF* format is currently under investigation. Prototype testing is accomplished in cooperation with the DIZ at University Medicine Greifswald and aimed at evaluating our ETL-process, at testing features that can be included into the BioCypher input adapter, and at deciding for a visualisation approach.

## 4. Outlook

Upon approval of prototype functionality, more data sources will be integrated and routines for data quality and similarity scoring will be added. Also, a feature for data declassification (default: classified) and standardised publication of KG structure will be implemented. Published local KG subsets can be combined into a global public MeDaX KG, providing information about the availability of non-public research. For the first beta release, a graphical user interface for querying the MeDaX KG clinic-internally is planned. In summary, MeDaX will provide a combined data resource to clinicians, scientists, and the interested public. The public MeDaX platform will foster biomedical data reuse by improving findability, accessibility, and interoperability of biomedical data gathered and stored at German hospital clinics and other health care providers.

## 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. vol. 20; 2018. p. 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. p. 326-31.
[4] Finlayson SG, LePendu P, Shah NH. Building the graph of medicine from millions of clinical narratives. Scientific data. 2014;1(1):1-9.
[5] Balaur I, Saqi M, Barat A, Lysenko A, Mazein A, Rawlings CJ, et al. EpiGeNet: a graph database of interdependencies between genetic and epigenetic events in colorectal cancer. Journal of Computational Biology. 2017;24(10):969-80.
[6] Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research. 2009;37(suppl 2):W170-3.
[7] 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.
[8] Lobentanzer S, Aloy P, Baumbach J, Bohar B, Danhauser K, Dogan T, et al. Democratising Knowledge Representation with BioCypher. ArXiv. 2022:doi: https://arxiv.org/abs/2212.13543.