MeDaX project plan: How to sustainably improve biomedical (research) data management

Judith A.H. Wodke[1,*], Ron Henkel[1] and Dagmar Waltemath[1]

[1]*University Medicine Greifswald, Institute for Community Medicine, Medical Informatics Laboratory, Walther-Rathenau-Str. 48, 17475 Greifswald, Germany*

Abstract

Several requirements, from technical (e.g. data interoperability) to juridical (e.g. data privacy), of high relevance and in some cases of contrary characteristics render digitalisation of health care in Germany at least challenging. With MeDaX (bio**Me**dical **Da**ta e**X**ploration) we aim at designing and implementing innovative and efficient methods for biomedical data storage, combination, enrichment, and for data retrieval and analysis based on graph technology. Following a federated approach we will provide an open source tool for building local knowledge graph instances of by default classified data. Declassification of publishable information will allow to export non-sensitive sub-graphs that can be combined into a global graph. A user interface for visualising and querying the incorporated data complete the open source and open access MeDaX information and research platform.

Keywords

biomedical data, data reusability, federated knowledge graph, information and research portal, medax

When dealing with digital health care data a Pareto solution for two contradicting principles has to be found. On the one hand, digital biomedical data collection is required for providing high-quality health care service to patients. On the other hand, the only safe measure to prevent data fraud and data theft is to not at all collect sensitive personal digital data. In consequence, at least in Germany, data reuse is hampered as digital biomedical data is largely locked away with (nearly) nobody knowing about its existence. But, particularly for biomedical data, reuse is highly desirable to prevent duplicate costs, time, effort, and possibly inconveniences associated with data collection or storage.

With MeDaX, we aim at building a public integrated information and research platform based on graph technology. The three main goals are i) to enable clinicians and researchers by rendering sensitive or otherwise classified information findable and thus possibly accessible for reuse, ii) to inform the public by providing aggregated and anonymised biomedical data open access, and iii) to follow highest technical and security standards as well as ethical principles, such as FAIR [1] and privacy by design.

To achieve those goals, we will build on the federated data structure implemented by the Medical Informatics Initiative[1] at the German university clinics and provide to their data

[1]https://www.medizininformatik-initiative.de/en/start, Accessed: 2022-12-12

integration centres a knowledge graph tool for local application. This tool integrates health care data [2] and population studies [3, 4] with information from public databases [5] into a graph database. It semantically enriches incorporated data with ontological [6] and provenance information. And it scores the contained knowledge according to data quality, to measures for similarity analysis, and to structure-based querying.

Graph databases have been shown specifically suitable for storage of complex heterogeneous data [7, 8]. The MeDaX approach maximises benefits i) for authorised medical personnel and researchers by rendering biomedical (research) data findable, interoperable, and possibly accessible, ii) for the public by providing transparent information about biomedical data usage, and iii) ultimately for patients by considering their biomedical data both, highly interesting for exploration but also worth maximum protection.

Acknowledgments

References

[1] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016).

[2] T. Ganslandt, M. Boeker, M. Löbe, F. Prasser, J. Schepers, S. Semler, S. Thun, U. Sax, Der kerndatensatz der medizininformatik-initiative: Ein schritt zur sekundärnutzung von versorgungsdaten auf nationaler ebene, in: Forum der Medizin-Dokumentation und Medizin-Informatik, volume 20, 2018, p. 17.

[3] U. John, E. Hensel, J. Lüdemann, M. Piek, S. Sauer, C. Adam, G. Born, D. Alte, E. Greiser, U. Haertel, et al., Study of health in pomerania (ship): a health examination survey in an east german region: objectives and design, Sozial-und Präventivmedizin 46 (2001) 186–194.

[4] H. J. Grabe, H. Assel, T. Bahls, M. Dörr, K. Endlich, N. Endlich, P. Erdmann, R. Ewert, S. B. Felix, B. Fiene, et al., Cohort profile: Greifswald approach to individualized medicine (gani_med), Journal of translational medicine 12 (2014) 1–15.

[5] S. Thun, J. Sass, A. Bartschke, M. Lehne, A. Essenwanger, E. Rinaldi, S. Rudolph, K. U. Heitmann, J. J. Vehreschild, C. von Kalle, The german corona consensus dataset (gecco): A standardized dataset for covid-19 research, BMC Med Inform Decis Mak 20 (2020).

[6] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, et al., Bioportal: ontologies and integrated data resources at the click of a mouse, Nucleic acids research 37 (2009) W170–W173.

[7] C. T. Have, L. J. Jensen, Are graph databases ready for bioinformatics?, Bioinformatics 29 (2013) 3107.

[8] A. Fabregat, F. Korninger, G. Viteri, K. Sidiropoulos, P. Marin-Garcia, P. Ping, G. Wu, L. Stein, P. D'Eustachio, H. Hermjakob, Reactome graph database: Efficient access to complex pathway data, PLoS computational biology 14 (2018) e1005968.