Achieving data FAIRification in a distributed analytics research platform for rare diseases

Anna Bernasconi[1,*], Cinzia Cappiello[1], Stefano Ceri[1] and Pietro Pinoli[1]

[1]*Department of Electronics, Information and Bioengineering – Politecnico di Milano, Milan, Italy*

Abstract

Data-driven medicine is fundamental to improving the accessibility and quality of the healthcare system. The availability of data is crucial for this purpose. In the context of a distributed analytics platform for analyzing healthcare data – employing the Personal Health Train paradigm – we propose to implement a solid data FAIRification infrastructure. This will allow us to achieve findability, accessibility, interoperability, and reusability of data, metadata, and results within a network of several medical centers participating in the BETTER Horizon Europe project, where the study of rare diseases (such as intellectual disability and inherited retinal dystrophies) will be targeted. Impacts will be visible to a large population of healthcare practitioners, prospectively influencing health policymakers.

Keywords

FAIR principles, healthcare, distributed analytics, rare diseases

Data-driven medicine is a crucial research area for the achievement of a more high-quality accessible healthcare system. Typically, the more data available for the intended analysis, the higher the chance to achieve accurate results [1]. However, the amount of available patient data is critical, especially in the context of rare diseases; here, even more predominantly than in other diseases, data sets are available and usable only at single medical centers. Reasons for the lack of data sharing are connected to ethical, legal, and privacy aspects and rules. Data centralization is not a viable option due to privacy concerns, particularly within the European Union, where the General Data Protection Regulation (GDPR) imposes stringent privacy standards.

The Horizon Europe project BETTER (Better rEal-world healTh-daTa distributEd analytics Research platform), started Dec. 1st, 2023, proposes the design and implementation of a decentralized infrastructure that will allow us to exploit the full potential of large sets of multi-source health data. This will be achieved by using customized AI tools to compare, integrate, and analyze datasets in a secure as well as cost-effective fashion. The project will target various use cases involving 7 European medical centers; they will provide sensitive patient data, including, possibly, clinical reports, medical images, genomic data (whole-exome, whole-genome sequences), biological data (cellular and molecular pathways), metabolic, environmental and demographic data, patient interviews, forms, and therapies details. Only the secure information will be made available and analyzed with a GDPR-compliant mechanism via a Distributed

The 15th Int. Semantic Web Applications and Tools for Health Care and Life Science conference (SWAT4HCLS 2024)
*Corresponding author.
anna.bernasconi@polimi.it (A. Bernasconi); cinzia.cappiello@polimi.it (C. Cappiello); stefano.ceri@polimi.it (S. Ceri); pietro.pinoli@polimi.it (P. Pinoli)
0000-0001-8016-5750 (A. Bernasconi); 0000-0001-6062-5174 (C. Cappiello); 0000-0003-0671-2415 (S. Ceri); 0000-0001-9786-2851 (P. Pinoli)

Analytics paradigm called the Personal Health Train (PHT) [2]. PHT can be explained via a railway system analogy that includes *trains*, *stations*, and *train depots*. The trains use the network to visit different stations to transport several goods, which in this analogy correspond to analytical tasks. By adapting this concept to BETTER, the analytical task is brought to the data provider (i.e., a medical center), whereas the data instances remain in their original location (called *station*).

As a technical partner of the project, Politecnico di Milano (i.e., the authors) will particularly focus on BETTER's objective to guide medical centers in collecting patients' data following a common schema in order to promote interoperability and re-use of datasets in scope. This includes legal/ethical data protection authorizations as well as data FAIRification (data documentation, cataloging, and mapping to well-established ontologies [3]). We will design a unified schema repository for medical centers' (meta)data integration, keeping a high abstraction level to encourage maximum interoperability (see [4, 5]).

Importantly, the project will aim at the integration of external sources such as European Health Data Space (EHDS), the 1+Million Genomes initiative (1+MG), and the European Open Science Cloud. Legal and ethical implications will be duly considered, and data access and re-use procedures will be proposed. Data pseudonymization will be performed as a default preprocessing step, mitigating the risk of personal data leaks. A real-world large-scale data integration framework (based on well-established ontologies) will be demonstrated taking into account heterogeneous datasets. The platform will be tested primarily on two rare disease use cases, inherent to pediatric intellectual disability and inherited retinal dystrophies.

In conclusion, the BETTER project relies on "bringing computation to data" via incremental and federated learning, which avoids unnecessary data moving across medical centers while exploiting much of the information encoded in such data. The project will enable EU medical centers and beyond to make full use of the potential offered by a safe and secure exchange, use, and reuse of health data fostered by robust data FAIRification. In the context of intellectual disability and inherited retinal dystrophies – with the potential of expanding the same paradigm to other diseases – the generated analytical tools will help healthcare professionals become more proficient in cutting-edge digital technologies, data-driven decision support, health risk surveillance, and control activities, monitoring and management of healthcare quality levels.

Acknowledgements. The work is supported by BETTER, Grant agreement 101136262.

References

[1] S. Welten, et al., DAMS: A distributed analytics metadata schema, Data Intelligence 3 (2021) 528–547.

[2] O. Beyan, et al., Distributed analytics on sensitive medical data: the personal health train, Data Intelligence 2 (2020) 96–107.

[3] A. Bernasconi, et al., Ontology-driven metadata enrichment for genomic datasets, in: SWAT4HCLS 2018, volume 2275 of *CEUR Workshop Proceedings*, 2018.

[4] A. Bernasconi, et al., Conceptual modeling for genomics: building an integrated repository of open data, in: ER 2017, Springer, 2017, pp. 325–339.

[5] A. Bernasconi, et al., A review on viral data sources and search systems for perspective mitigation of COVID-19, Briefings in Bioinformatics 22 (2021) 664–675.