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

Anna Bernasconi, Cinzia Cappiello, Stefano Ceri, and Pietro Pinoli

Contact: anna.bernasconi@polimi.it

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

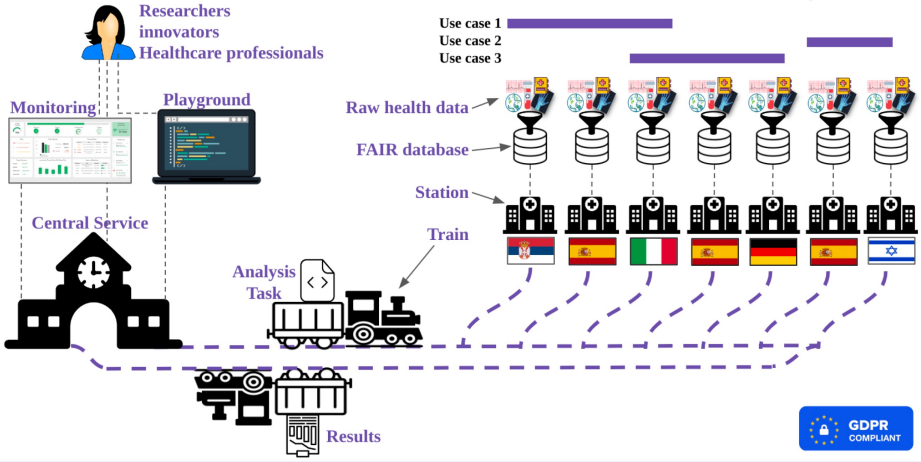
Goal

Achieving findability, accessibility, interoperability, and reusability (**FAIRification**) of data, metadata, and study results within a **network of several medical centers** participating in the BETTER Horizon Europe project, targeting the study of **rare diseases** (such as intellectual disability and inherited retinal dystrophies)

Better rEal-world healTh-daTa distributEd analytics Research platform

- Horizon Europe project started Dec. 1st, 2023 (https://www.better-health-project.eu/)
- Design and implementation of a decentralized infrastructure to exploit the full potential of large sets of multi-source health data.
- Various use cases involving 7 European medical centers providing sensitive patient data (e.g., clinical reports, medical images, genomic data, biological data, metabolic, environmental and demographic data, patient interviews, …)
- Only the secure information made available and analyzed with a GDPR-compliant mechanism via a Distributed Analytics paradigm (Personal Health Train)

Personal Health Train paradigm



- Railway system analogy that includes trains, stations, and train depots [1].
- **Train** transports goods (= **analytical tasks**).
- **Station** (= **data provider**), accessible by the Train. It executes the task, which processes the available data.
- **Depot** (= **Central Service**). It includes procedures for Train orchestration, business and operational logic, data management and discovery [2].
- Further modules for privacy and security enforcement.

Health datasets FAIRification and preprocessing

Our group is involved in overcoming cross-border barriers to health data integration, access, FAIRification, and preprocessing. Practical objectives include:

(1) Discovering and collecting datasets available at each medical centre, anticipating interoperability with external databases.

(2) Designing a unifying repository schema useful for integration (exploiting, e.g., FHIR HL7 standards) [3].

(3) Prepare ETL for processing health datasets (**reusability**) [4].

(4) Harmonizing data by employing standardized terminologies and ontologies (**interoperability**) [5].

(5) Loading aggregated information and metadata into the project's repository (**findability** and **accessibility**).

Clinical Use Cases

- Integration of genomic and phenotypic data from paediatric rare diseases to decipher pathways of intellectual disability
- Accelerate Inherited Retinal Dystrophies Diagnosis using AI
- Predicting the risk of self-harm and suicidal behaviors in patients with Autism Spectrum Disorders

Take a picture to download the full poster paper

Bibliography

[1] O. Beyan, et al., Distributed analytics on sensitive medical data: the personal health train, Data Intelligence 2 (2020) 96–107.
[2] S. Welten, et al., DAMS: A distributed analytics metadata schema, Data Intelligence 3 (2021) 528–547.
[3] A. Bernasconi, et al., Conceptual modeling for genomics: building an integrated repository of open data, in: ER 2017, Springer, 2017, pp. 325–339.
[4] A. Bernasconi, et al., META-BASE: a novel architecture for large-scale genomic metadata integration, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022, 19(1), pp.543-557.
[5] A. Bernasconi, et al., Ontology-driven metadata enrichment for genomic datasets, in: SWAT4HCLS 2018, volume 2275 of CEUR Workshop Proceedings, 2018.

Acknowledgments

The project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101136262. The communication reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.