Transformation and integration of heterogeneous health data in a privacy-preserving distributed learning infrastructure

Chang Sun[1], Vincent Emonet[1], Johan van Soest[1,2], Annemarie Koster[3], Andre Dekker[2], and Michel Dumontier[1]

[1] Institute of Data Science, Maastricht University, Maastricht, the Netherlands
[2] Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre+, the Netherlands
[3] Department of Social Medicine, CAPHRI Care and Public Health Research Institute, Maastricht University, The Netherlands

Problem statement: A growing volume and variety of personal health data are being collected by different entities, such as healthcare providers, insurance companies, and wearable device manufacturers. Combining heterogeneous health data offers unprecedented opportunities to augment our understanding of human health and disease. However, a major challenge to research lies in the difficulty of accessing and analyzing health data that are dispersed in their format (e.g. CSV, XML), sources (e.g., medical records, laboratory data), representation (unstructured, structured), and governance (e.g., data collection and maintenance)[2]. Such considerations are crucial when we link and use personal health data across multiple legal entities with different data governance and privacy concerns.

Proposed approach: Our approach to tackling this challenge is extending the Personal Health Train Architecture [1, 2] to analyze health data from multiple sources in a privacy-preserving manner[5]. Instead of centralizing the data for the analysis, researchers send data-processing algorithms (application trains) to each data source. To be able to run the application train, each source deploys a data station which stores the data required for the analysis. The data station only returns the results of the analysis rather than any of the original data. Data stations should be able to provide data in certain standard formats and structures in line with the FAIR (Findable, Accessible, Interoperable, Reusable) principles[3]. In our method, the FAIR data stations are based on Semantic Web technologies. The data is stored as a knowledge graph using the Resource Description Framework (RDF), while its representation complies with the ontologies accepted by the community. We apply the Data2Service framework[4], first described at the 2018 SWAT4LS, to now semi-automatically transform and integrate heterogeneous health data into RDF data. This knowledge graph complies to a target set of ontologies, and is subsequently made available as FAIR data stations through a set of interfaces and services (e.g., SPARQL, API). The framework defines a set of scalable and sustainable transformation workflows to convert any structured data sources to a target data model. To help the researchers with the mapping process, Data2Services generates SPARQL mappings files based on the input data structure. Then, researchers create the analysis as an application train to query the RDF data available at FAIR data stations using SPARQL, pre-process data, and execute machine learning models[6]. The FAIR data stations are always in the governance of the data entities. Instead of duplicating the original data and sending to the researchers, our method ensures the original data will not leave the data entities and be exposed to any others including the researchers. The researchers only receive the results of applications.

Use case: The goal is to study annual healthcare costs in relation to the incidence of Type 2 Diabetes Mellitus (T2D) without revealing any original data. We used patients' health data from De Maastricht Studie, which is a population study on T2D, and their healthcare cost data from Statistics Netherlands. The original data files (SAV, CSV) were automatically transformed to RDF by defining SPARQL construct queries, and subsequently loaded into a triple store and made available as per the FAIR data station specification. Researchers can retrieve RDF data by SPARQL queries from FAIR data station and generate to CSV, JSON, XML formats for the data analysis. The variable names, labels, keywords were used to find the most relevant ontologies using the NCBO Ontology Recommender[7]. RDF data was structured to the relevant ontologies - Ontology of Consumer Health Vocabulary and SNOMED Clinical Terms. The structured RDF data is stored at FAIR data stations, with metadata compliant to the HCLS Dataset Description Profile. Analysis models such as correlation matrix, linear regression were successfully tested in the infrastructure.

References

1. Personal Health Train, Dutch Techcentre for Life Sciences. (n.d.). https://www.dtls.nl/fair-data/personal-health-train/
2. van Soest, J., Sun, C., Mussmann, O., Puts, M., van den Berg, B., Malic, A., van Oppen, C., Towend, D., Dekker, A. and Dumontier, M., 2018. Using the Personal Health Train for Automated and Privacy-Preserving Analytics on Vertically Partitioned Data. Studies in health technology and informatics, 247, p.581.
3. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3.
4. Emonet, V., Malic, A., Zaveri, A., Grigoriu, A.; Dumontier, M. (2018): Data2Services: enabling automated conversion of data to services. figshare. Journal contribution.
5. Sun, C., Ippel, L., Wouters, B., Malic, A., Adekunle, O., Mussmann, O., Koster, A., Townend, D., Dekker, A. and Dumontier, M., 2019. A Privacy-Preserving Infrastructure for Analyzing Personal Health Data in a Vertically Partitioned Scenario. Studies in health technology and informatics, 264, pp.373-377.
6. Deist, T.M., Jochems, A., van Soest, J., Nalbantov, G., Oberije, C., Walsh, S., Eble, M., Bulens, P., Coucke, P., Dries, W. and Dekker, A., 2017. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clinical and translational radiation oncology, 4, pp.24-31.
7. Martínez-Romero, M., Jonquet, C., O'Connor, M. J., Graybeal, J., Pazos, A., Musen, M. A. (2017). NCBO Ontology Recommender 2.0: An Enhanced Approach For Biomedical Ontology Recommendation. Journal of Biomedical Semantics,8(21)