July 2020

Integrating Healthcare Data for Enhanced Citizen-centred Care and Analytics

Juliana K. F. BOWLES [a,1], Juan MENDOZA-SANTANA [a],
Andreas F. VERMEULEN [a], Thais WEBBER [a], and Euan BLACKLEDGE [b],
[a] *School of Computer Science, University of St Andrews, St Andrews KY16 9SX, UK*
[b] *Sopra Steria, Orchard Brae House, 30 Queensferry Rd, Edinburgh EH4 2HS, UK*

Abstract.
The potential of healthcare systems worldwide is expanding as new medical devices and data sources are regularly presented to healthcare providers which could be used to personalise, improve and revise treatments further. However, there is presently a large gap between the data collected, the systems that store the data, and any ability to perform big data analytics to combinations of such data. This paper suggests a novel approach to integrate data from multiple sources and formats, by providing a uniform structure to the data in a healthcare data lake with multiple zones reflecting how refined the data is: from raw to curated when ready to be consumed or used for analysis. The integration further requires solutions that can be proven to be secure, such as patient-centric data sharing agreements (smart contracts) on a blockchain, and novel privacy-preserving methods for extracting metadata from data sources, originally derived from partially-structured or from completely unstructured data. Work presented here is being developed as part of an EU project with the ultimate aim to develop solutions for integrating healthcare data for enhanced citizen-centred care and analytics across Europe.

Keywords. Healthcare, Data Lake, Integration, Blockchain, Data Analytics

Introduction

The EU project Serums[2] addresses a recently exacerbated need - in the presence of a global pandemic - of improving the coordination of healthcare provision across Europe and beyond. As citizens move between countries, their newly produced medical data, including data from personal devices, must be continuously integrated to complement medical records across the countries where they have lived or where they need to be treated. This is essential to guarantee that all required information on a patient is available, and can thus be used to improve the quality of the treatment they receive. This vision requires novel mechanisms to exchange confidential medical records to personalise clinical advice and enhance treatment plans, whilst enabling trust in data security and privacy at all times. In order to be able to integrate personal medical data from multiple sources

[1]Corresponding Author: Juliana K. F. Bowles, University of St Andrews, UK; E-mail: jkfb@st-andrews.ac.uk.
[2]For more information please see `www.serums-h2020.org`.

such as personal healthcare devices, primary, secondary and/or tertiary care, we need a coherent and unified notion of a *smart patient health record* (SPHR). The integration further requires solutions that can be proven to be secure, such as patient-centric data sharing agreements (smart contracts) on a blockchain, and novel privacy-preserving methods for extracting metadata from data sources, originally derived from partially-structured or from completely unstructured data. Some aspects of our work within Serums are described next.

Methods

A data lake is a universal data storage space which in our context is used for any healthcare data gathered from various healthcare providers and devices [6]. One advantage of a data lake is that it scales with ease, and its usage can range from simple storage, to a base from which to run analytics or big data processing, and machine learning (ML) at scale. A data lake consists of different zones (*workspace*, *raw*, *structured*, *curated*, *consumer*, *analytic* and *trash*) depending on the pre-processed state of the data it contains, and is responsible for carrying out data processing activities such as Retrieve, Assess, Process, Transform, Organise and Report (R·A·P·T·O·R). The R·A·P·T·O·R processing pipeline autocoder is scalable and a very efficient way of processing large amounts of data.

The pipeline transforms the data according to a standard structure where data is classified into five groups: Time-Person-Object-Location-Event (T·P·O·L·E), forming what is known as a *data vault* model within the *curated* data lake zone. This model enables the standardisation of all data into an expandable hyper-scalable structure that can load any kind of health or social care related data. This makes the process of combining varied data sources easier as well as the ability to gain new insights from considerably more data through data analytics and machine learning (in the respective *analytic* zone).

To address security concerns, the Serums tool-chain [3] makes use of a blockchain to control data access through well defined rules. Rules can, for instance, limit *what* patient data can be seen by *who* and *when*, and logs are kept on every attempt to access patient records. Access is controlled under the General Data Protection Regulation (GDPR[3]) using smart contracts. In addition, access logs can be explored to detect attempts at security breaches over medical systems if the patterns of logged attempts are unusual. For authenticated users, the blockchain controls the data that can be shown to the user, and the extraction is obtained from the T·P·O·L·E data lake. Figure 1 shows a general overview of the Serums project components [1]. Patients and healthcare providers interact with the system through the front end (Serums Web Interface) which communicates with a back-end (Serums API) responsible for managing the integration of all components including authentication modules (refer to [2]), blockchain and the data lake.

[3]Information on GDPR can be found at `https://gdpr-info.eu/`

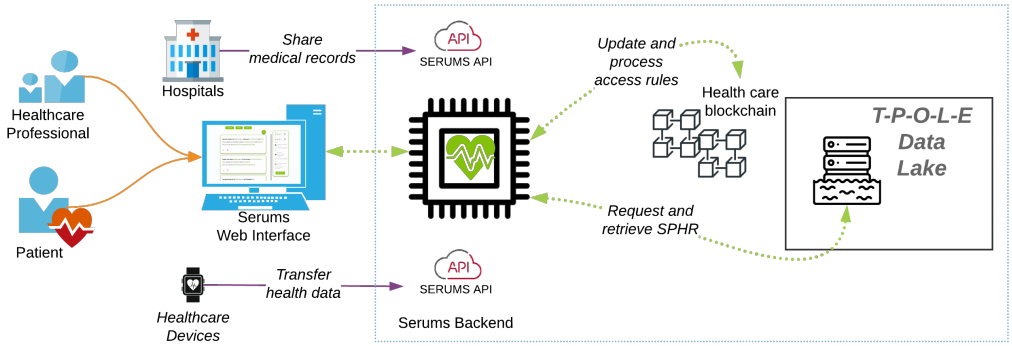

Figure 1. Serums Overview

Figure 2 shows how the T·P·O·L·E data lake expands to hubs, links and satellites to enable the effective and efficient storage of the health and social care data into a globally universal data storage.

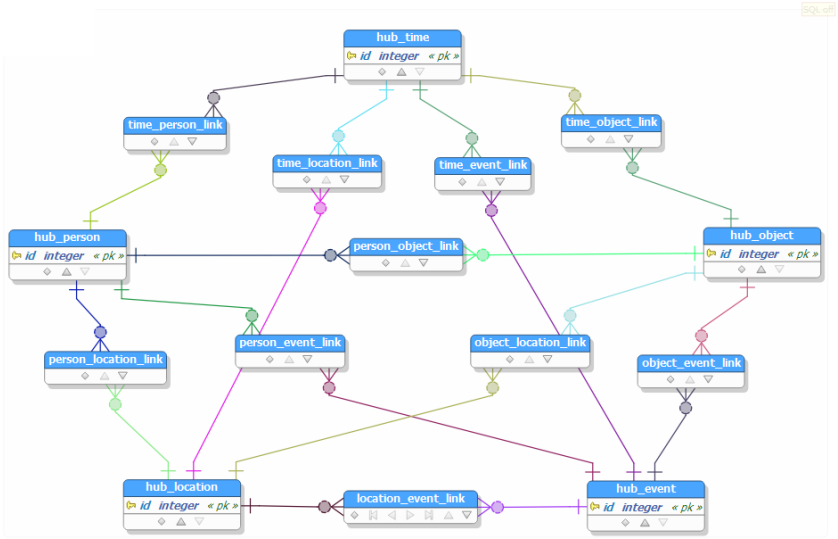

Figure 2. The T·P·O·L·E structure within a data lake

The T·P·O·L·E model has the potential to resolve many challenges including the one identified in [4] on bringing together multiple sources of information on medications to provide a so-called *My Medication Passport (MMP)* for patients. Studies have shown that MMPs help patients understand their medications and promote adherence [5] contributing to an improved quality of life. The flexibility of the R·A·P·T·O·R processing on healthcare data lakes and the T·P·O·L·E data vault means that we can combine varied data for a single patient more easily, and we can process more data through analytics leading to the faster discovery of novel insights than what is currently possible. A further benefit is that we are able to bring new sources of data into the data lake at all times

without interference with already processed data, as the data within the lake is split in zones as mentioned earlier.

Results and discussion

The integration result is a data lake with advance analytic capabilities that can handle the complexities of new global healthcare requirements. Data from various data sources enter the R·A·P·T·O·R processing ecosystem and is structured following the T·P·O·L·E model. Within Serums we explore three uses cases provided by hospitals in the Netherlands (sensor information on patient mobility for patients that have received a hip replacement), Catalonia (device information to monitor elderly patients with diabetes and cardiovascular disease from home) and Scotland (cancer patients that report daily on their symptoms in between chemotherapy treatments) [3].

The data is stored in a Linux shared file system within the proof-of-concept. We have already production grade solutions that are using big data technologies like AWS Data Lake Formation, Azure Data Lake v2.0 and Google Cloud Storage. The data processing in the proof-of-concept is done using custom Python code. In the production versions we use AWS Sagemaker plus Deep Learning, Azure Data Factory plus Data Lake Analytics and Google Cloud Datalab plus ML Engine. The data lake can grow and further adapt to new health and social care data that is added to it enhancing the information we may have on individual patients, on general cohorts of patients (e.g., cancer patients) and on novel treatments, further improving knowledge we can gain through ML and data analytics. Figure 3 shows the steps in which the data lake interacts with the Serums API to connect and process the data into a SPHR.

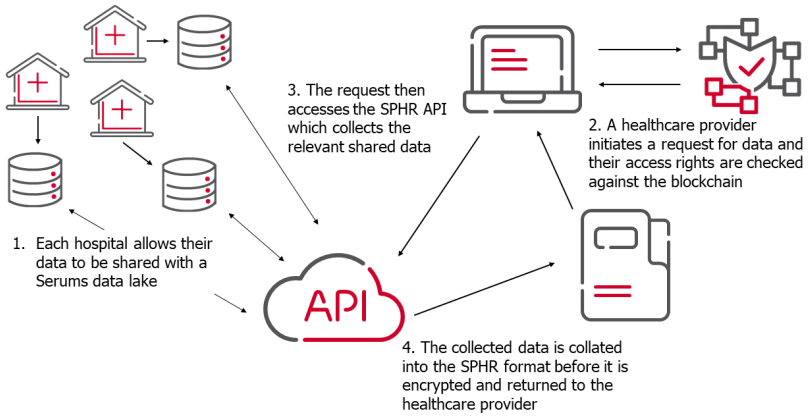

Figure 3. T·P·O·L·E connections

Health and social care providers, in our case three hospitals, share their data with the data lake via a Serums API gateway that was custom build for the proof-of-concept. The providers use a Web Interface (cf. Figure 1) to request the data in accordance with an underlying agreed smart contract data request from the health care blockchain. The

T·P·O·L·E data factory then prepares the data and places an encrypted version in the consumer and analytics zone of the data lake ready for the SPHR API gateway to process. The health and social care providers collect the data via the API gateway after it has been decrypted internally with the keys they receive from the blockchain.

Most healthcare systems today consist of distributed heterogeneous systems that do not necessarily communicate with each other making it very challenging, if not impossible, to readily integrate data from medical practices, hospitals, medical devices, and so on, in real-time and in a straightforward manner. The approach followed in Serums with a healthcare data lake allows us to combine different data sources because they are pre-processed in the same way through the T·P·O·L·E model. The data lake concept thus removes the complexities of healthcare systems while opening novel and unprecedented capabilities to deploy any T·P·O·L·E compliant analytics and ML algorithms to process the data lake at scale.

Conclusion

Serums comes with a methodology that can easily be expanded into a global health and social care data model to address current and future requirements to support near-real-time analytics on all citizens. Serums will supply a base model for a selected set of healthcare providers initially (cf. [3] for further details), however, it is not limited to this selection. The vision of Serums is to provide flexible structures which can be expanded to a European-wide solution for integrated medical records accessible anywhere in Europe.

Acknowledgements

This research is funded by the EU H2020 project SERUMS: Securing Medical Data in Smart Patient-Centric Healthcare Systems (grant code 826278).

References

[1] Bowles, J., Mendoza-Santana, J., Webber, T.: Interacting with next-generation smart patient-centric healthcare systems. In: Adaptive and Personalized Privacy and Security Workshop (APPS 2020), UMAP (Adjunct Publication), in press (2020)

[2] Constantinides, A., Belk, M., Fidas, C., Pitsillides, A.: Design and development of the serums patient-centric user authentication system. In: Adaptive and Personalized Privacy and Security Workshop (APPS 2020), UMAP (Adjunct Publication), in press (2020)

[3] Janjic, V., et al.: The SERUMS tool-chain: Ensuring Security and Privacy of Medical Data in Smart Patient-Centric Healthcare Systems. In: Proceedings of the 2019 IEEE International Conference on Big Data, Big Data 2019. pp. 2726–2735. IEEE, IEEE (Dec 2019)

[4] Jubraj, B., Blair, M.: Use of a medication passport in a disabled child seen across many care settings. BMJ Case Reports (2015). https://doi.org/10.1136/bcr-2014-208033

[5] Leavey, G., Abbott, A., Watson, M., Todd, S., Coates, V., McIlfactrick, S., McCormack, B., Waterhouse-Bradley, B., Curran, E.: The evaluation of a healthcare passport to improve quality of care and communication for people living with dementia (EQuIP): A protocol paper for a qualitative, longitudinal study. BMC Health Services Research (2016). https://doi.org/10.1186/s12913-016-1617-x

[6] Vermeulen, A.F.: Practical Data Science: A Guide to Building the Technology Stack for Turning Data Lakes into Business Assets. Apress (2018)