Design of a Trustworthy and Resilient Data Sharing Platform for Healthcare Provision [⋆]

Matthew Banton[0000−0001−8170−3899], Juliana Bowles[0000−0002−5918−9114], Agastya Silvina[0000−0002−0012−9256], and Thais Webber[0000−0002−8091−6021]

School of Computer Science, University of St Andrews St Andrews KY16 9SX, UK
jkfb@st-andrews.ac.uk

Abstract. Healthcare data sharing platforms have been gaining prominence over the last decade, especially with the emergence of technologies dedicated to increase system security and users' privacy. Moreover, these platforms are becoming less centralised as time progresses, with need for more data from a variety of locations and settings to be transferred between authorised parties. These requirements also include legal and ethical concerns when creating such solution. Through data sharing, organisations can gain access to previously unknown information or higher quality data, share research findings, and make decisions based on larger (and hopefully more representative) datasets. Such platform should be resilient to attack or loss of data and be able to recover quickly and efficiently from unexpected events. This paper focuses on the blend of emerging technologies (data lake and blockchain) in a design to provide secure and resilient data sharing to only those patients and healthcare professionals authorised to access it across multiple European countries.

Keywords: Healthcare · Data Sharing Systems · Security · Resilience.

1 Introduction

Healthcare provision has been the target of different technological advances over the decades, ranging from processes improvement within organisations to data and devices integration for real-time access. [3, 9, 13]. Medical data sources have become less centralised with more data being used in a variety of settings and for different purposes. Several challenges arise from the distributed nature of medical data, including confidentiality issues such as how to securely and reliably share the data with different healthcare providers whilst preserving patients' privacy [9, 13, 20]. This has led to the inclusion of security and resilience aspects early in the development of such systems [7, 13, 14, 19].

Currently, a focal point in the European healthcare domain is the development of flexible and secure data sharing platforms for healthcare provision including emerging technologies [11, 12, 18]. This comes with a variety of legal,

[⋆] The research in this paper was supported by the EU H2020 project SERUMS: Securing Medical Data in Smart Patient-Centric Healthcare Systems (grant code 826278).

ethical and technical challenges such as protecting sensitive health information under different legislation [16] together with the ability to control and audit the access to the confidential medical data [12,14].

An important aspect when designing data sharing systems is organising the architectural key assets so that they can be protected at different layers against external events and malicious users that could cause, for instance, breaches in both data integrity and confidentiality; and to have strategies to resist, detect and to restore the system normal functioning after an outage occurs [22]. Resilience reflects the ability of a system to continue to deliver essential functions and services to legitimate users while it is under occurrence of unexpected events as well as refers to its ability to recover from those events [22]. Literature also defines resilience based on failure probabilities aiming to reduce the impact of disruptive situations by minimising the probability of failures in the first instance, and then reducing the consequences of disruptive events, thus improving system recovery time [10]. In data sharing systems design for healthcare, the approaches applied to ensure system resilience may contribute to the difference between life and death for patients, especially when professionals could have access to crucial health information provided by the system and the data is unavailable, missing or incorrect [13,17].

The EU Horizon 2020 research project Serums[1] [15,23] proposes an architectural model to share health information data, acquired from a variety of sources and formats, only to authorised individuals and healthcare organisations in a secure and efficient way. The ultimate platform goal is to be resilient to data corruption and breaches and allow a secure user-friendly access to only authorised users. Serums core design is a combination of data lake and blockchain technologies [6,23,24] that synergies to provide an indelible record of access requests and changes to data, while also allowing health centres to continue accessing their data as they see fit. Even with availability assured, it is critical that the system can be trusted to not disclose data to untrusted parties, and to not allow data breaches from unauthorised users, given the sensitive and confidential nature of medical data [21].

This paper aims to showcase the resilience characteristics on Serums platform design, and how these features can be beneficial to both withstand and recover from external unexpected events. Thus, we structure the paper as follows. Section 2 provide the details of the Serums architecture and how the various systems interlink and operate. Section 3 discusses the particular resilience aspects related to Serums' core technologies, blockchain and data lake, highlighting important security and performance aspects when introducing resilient properties in data sharing systems like Serums. Section 4 we conclude the paper summarising Serums platform design aspects towards achieving resilient characteristics.

[1] For more information refer to `www.serums-h2020.org`

2 Data Sharing Platform Design

The Serums platform is a tool-chain [4,15] built to demonstrate the need for and the advantages of an integrated data sharing platform for healthcare provision in Europe, ensuring privacy to users and security when accessing medical records [2,8]. The architectural design workflow [23] describes the overall system design process and detail its core functionalities [5,6].

Fig. 1 illustrates Serums platform components and their connections, as well as its potential end-users (individuals and organisations), and the expected interconnection with other external medical data sources. Serums Smart Health Centre System (SHCS) integrates essential technologies for building layered security attributes since design. The core layers to securely process users requests in the Serums platform are, respectively, the authentication component [8] (user login feature), the blockchain component [5] (user authorisation to access data and personalised access rules creation) and the data lake component [6] (fine-grained medical data retrieval).

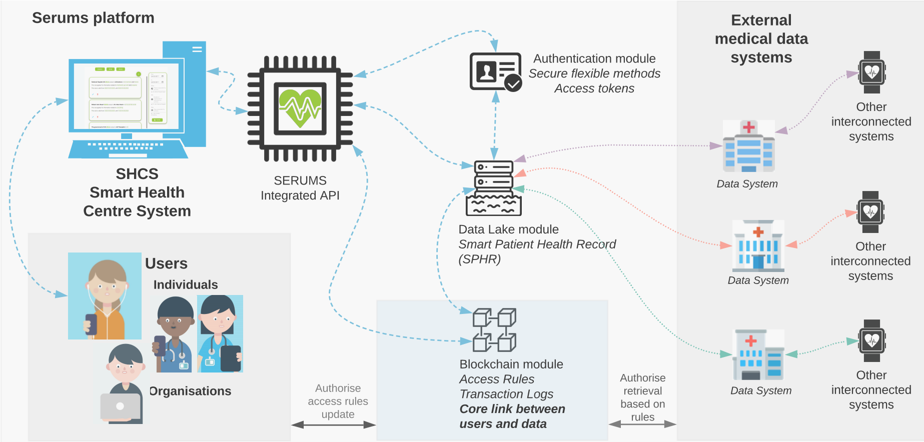

Fig. 1. Serums platform overview: core components, users and interactions

Authentication is the first layer of security in the system. Serums proposes new flexible techniques to the user create passwords and authenticate themselves in such a way it is less likely the occurrence of security breaches [2, 8]. This component acts as a certifying agent for the authenticity of users and confirms to other modules that they can meet the requests originated from these users.

A core module in Serums architecture is the Blockchain component, which is the core link between individuals and medical data, since it authorises the users to access medical records and other data sharing features within the platform like personalised access rules creation. This component stores the access rules as customised transparent fine-grained access permissions to individuals, and it is responsible for keeping immutable logs of all requested transactions.

Access rules formally are defined as strict tuples identifying the 'who' (patient), 'to whom' (professionals), 'what' (parts of medical data) and 'when' (validity of the rule) the data can be shared in the platform front-end after successful authentication into the system. These rules state always up-to-date information on access privileges and provide on-demand check for retrieval eligibility [5]. Exploiting blockchain technology in this way allows for greater resilience than a standard database with access lists [12], as the indelible nature of the blockchain ensures that all transactions are logged, and the distributed aspect ensures that there is not a single entity that has control over the access rules. For more detail on the blockchain component setup and behaviour, as well as access rules creation formal approach please refer to [5, 6].

Finally, the data lake component manages the medical data, as well as provides an area for medical centres to securely upload that data for processing. Tags (or labels) are added to the data by their providers, to define the data purpose and subsequent retrieval granularity [6]. This feature, when combined with filters and rules within the platform, ensures that the authorised users allowed to access only a subset of a patient data, only retrieve that specific subset of data from the medical centres systems [5]. Storing only metadata and access rules on-chain, using a data lake to manage the data, like Serums does, integrating records in unified way (SPHR - Smart Patient Health Record [6]), allows the scale of the data sharing platform as the European medical healthcare provision system requires [6, 9, 15].

It is worth mentioning that all of Serums core technologies have weaknesses that have been mentioned in previous literature [1, 11, 24]. However, Serums attempts to account for these vulnerabilities by employing an holistic system, which allows the components strengths to synergise with one another [2, 23].

3 Blockchain and Data Lake Components Resilience

This section describes how Serums can potentially reduce the occurrence of disruptive events and how it would react when they occur. Following, we highlight the characteristics that increase Serums platform's ability to minimise the impacts of these events and restore its normal operation afterwards.

- **Potential resilience of a distributed database:** Blockchain as a distributed database can allow alternative nodes to take over if one goes down. The advantage of using blockchain in the way Serums platform does is that it ensures availability of the rules to control the access over medical data through the Serums Data Lake. If the local node were to go down, or even the data lake component of the system, then patients could still update rules relating to their data. These rules would then be updated on the local node when it is returned to service. Another important aspect is medical centres are integrated within Serums, each one has its own node on the blockchain, with each node replicating the other nodes. They decide individually on what the data should be, and through consensus it increases the probability that data is correct, with each medical centre contributing to the whole chain, there is no single entity controlling the data.

- **Advantages of immutability with regards to resilience:** sharing the access rules in the way Serums does, makes tampering a difficult task [2]. A malicious user would need to find a way to gain access to accounts that have legitimate access to change the required rules (e.g., phishing), rather than use injection type attacks to alter the state of rules. Even in the case of an attacker being able to gain access to a legitimate account and alter rules, the scope will be limited by the access to medical records of that account [2]. Logs will be recorded onto the chain containing the transaction undertaken, the grantor and grantee, as well as the date and time of its occurrence. The indelible nature of the blockchain means that these logs are correct, and free from any malicious influence (something malicious actors do when they seek to infiltrate a system covertly). In case the access rules are overwritten by users, it is a trivial task to track the origin of these changes by analysing the immutable blockchain logs and return the rules in place to their previous state. Activity logs captured by the blockchain component are essential assets for determining suspicious or failed transactions. They provide useful information for defining the actions to reduce the impact of disruptive events as how to safely restore system to a reliable state of operation.
- **Only essential data storage in the blockchain:** the medical data is not on-chain itself in Serums. The reason is that when data that must be stored on-chain is larger than the blocks, it can impact performance negatively, leading to system disruption and "brown-outs" (i.e., restriction on the availability of particular features). Data lakes, however, are perfectly suited for big data applications and storing large volumes of data. As such, to increase resilience we use both a data lake to store medical data, and the blockchain to manage access and log requests.
- **Coordinated data processing on the data lake:** a common issue for data lakes is becoming a dumping ground for data, and transforming into a "data swamp" [11]. This naturally affects the availability of data, as if useful data cannot be found, then it might as well not be there. Serums combats this particular issue through using a four-stage process:

  – Stage 1: the authorised medical data is selected from data sources according to authorised tags and uploaded to the raw zone of the data lake.

  – Stage 2: structure is added to the medical data, through the use of scripts. The data itself will stay in the raw zone, however the metadata added to it (SPHR) will be added to a structured zone, which allows the data to become efficiently searchable.

  – Stage 3: this stage deals with SPHR access specifically. Relevant data is searched from the structured zone and uploaded from the unstructured zone to the curated zone. This zone is encrypted to ensure integrity. After the health record is sent, the contents of the curated zone are deleted.

  – Stage 4: finally, the completed SPHR is moved to the consumer zone, where it can be accessed. Again, this is a temporary zone, and is encrypted

to ensure integrity. The consumer zone is the only zone where completed SPHR can be accessed from.
Focusing on resilience, we highlight that the data lake also holds two other areas that extract data from the structured (and unstructured) areas: the Workspace and Analytics zones. These zones are for developers to work without risk of any data being destroyed.

- **Extensive use of metadata:** The data lake metadata is used extensively, and actual medical data are accessed only sparingly when absolutely required by authorised users. This allows for greater availability of data on the platform, as the metadata of a medical record can be updated in nearly real-time. For example, if a file is missing from the unstructured zone, the metadata will allow the system to determine where the data came from, and re-upload a copy for use. Additionally, medical centres' own systems will not rely on Serums platform being online, but they could potentially benefit from having an additional copy of the data available from Serums (e.g., for backup purposes) as long as it respects the data protection legislation and recommended standards [16].

4 Conclusion

We have discussed how Serums attempts to ensure that the platform is resilient to data breaches, leaks, or even corruption, and how parts of the Serums system could go down without impacting other areas. However, these features, while aiding resilience overall, would not ensure a resilient system themselves. Typical backups would still be required, as well as well-defined processes to determine how and when to initiate and use those backup procedures.

A contributing factor to Serums resilience, is that it only manages copies of medical data, meaning that should the Serums system go down, local medical records and systems are not affected. This means that should Serums go down, local systems and processes can continue as normal, and then changes and updates be uploaded once Serums comes back online. When combined with the modular engineering approach, which means that aspects of Serums could fail without impacting other areas, we believe this could result in a system that is unlikely to suffer large scale failures and be more easily recovered in case of system disruption.

There are effectively only two components that could stop most of the Serums functionalities should they fail. The first component is the web portal, which would logically block users from accessing the system if it were to go down. However, even this would not cause all functionalities to be unavailable, as data is uploaded to the data lake, and any analysis can be performed independently of the web portal as a measure to increase resilience. The second component is the blockchain, which is essential for authorising users' access to medical data, and if it were to fail then users would therefore not be able to gain the permissions necessary to access the data. It can be argued that the blockchain is the most resilient aspect of the system, however, it being distributed between multiple medical centres. If one medical centre version were to go offline, then users would be directed to another node. The nodes are not user-centric, and all store a copy of each other's data. When the offline node comes back online, it can copy any changes made while it was offline. In the meantime, the worst scenario is that access is delayed slightly due to larger hops to ensure access permissions.

The different technologies involved in Serums also have different trade-offs, but their advantages can function together to build a holistic system that is resistant to errors or failures. For instance, disadvantages of the blockchain when it comes to big data are mitigated by the data lake, while its advantages include immutable logs and records of system access and respective transactions, which can aid in system recovery.

The Serums platform allows the components to communicate and effectively operate together, depending on functionality required, with no tighter integration that could cause a cascading failure in the event of one component failing. We believe the blend of these technologies brings key attributes to enable a secure and resilient data sharing platform for healthcare provision, combining blockchain to log any activity and allow easy restoration, with a data lake which is perfectly suited to this kind of large-scale data platform.

Acknowledgment

We thank all our partners and universities who contribute to the Serums project design and development.

References

1. Abu-elezz, I., Hassan, A., Nazeemudeen, A., Househ, M., Abd-alrazaq, A.: The benefits and threats of blockchain technology in healthcare: A scoping review. International Journal of Medical Informatics **142**, 104246 (October 2020)
2. Banton, M., Bowles, J., Silvina, A., Webber, T.: On the benefits and security risks of a user-centric data sharing platform for healthcare provision. In: UMAP'21 Adjunct: Publication of the 29th ACM Conf. on User Modeling, Adaptation and Personalization. pp. 351–356. ACM, New York, NY, USA (2021)
3. Bardhan, I.R., Thouin, M.F.: Health information technology and its impact on the quality and cost of healthcare delivery. Decision Support Systems **55**(2), 438–449 (May 2013)
4. Bowles, J., Mendoza-Santana, J., Webber, T.: Interacting with next-generation smart patient-centric healthcare systems. In: UMAP'20 Adjunct: Publication of the 28th ACM Conf. on User Modeling, Adaptation and Personalization. pp. 192–193. ACM, New York, NY, USA (2020)
5. Bowles, J., Webber, T., Blackledge, E., Vermeulen, A.: A blockchain-based healthcare platform for secure personalised data sharing. Studies in Health Technology and Informatics, Public Health and Informatics **281**, 208–212 (May 2021)
6. Bowles, J.K.F., Mendoza-Santana, J., Vermeulen, A.F., Webber, T., Blackledge, E.: Integrating healthcare data for enhanced citizen-centred care and analytics. Studies in Health Technology & Informatics **275**, 17–21 (2020)

7. Chen, J., Lv, Z., Song, H.: Design of personnel big data management system based on blockchain. Future Generation Computer Systems **101**, 1122–1129 (Dec 2019)
8. Constantinides, A., Belk, M., Fidas, C., Pitsillides, A.: Design and development of the serums patient-centric user authentication system. In: UMAP'20 Adjunct: Publication of the 28th ACM Conf. on User Modeling, Adaptation and Personalization. pp. 201–203. ACM, New York, NY, USA (July 2020)
9. Dhayne, H., Haque, R., Kilany, R., Taher, Y.: In search of big medical data integration solutions - a comprehensive survey. IEEE Access **7**, 91265–91290 (2019)
10. Dinh, L.T., Pasman, H., Gao, X., Mannan, M.S.: Resilience engineering of industrial processes: Principles and contributing factors. Journal of Loss Prevention in the Process Industries **25**(2), 233–241 (March 2012)
11. Gavrilov, G., Vlahu-Gjorgievska, E., Trajkovik, V.: Healthcare data warehouse system supporting cross-border interoperability. Health informatics journal **26**(2), 1321–1332 (October 2020)
12. Guo, H., Li, W., Nejad, M., Shen, C.C.: Access control for electronic health records with hybrid blockchain-edge architecture. In: 2019 IEEE International Conference on Blockchain (Blockchain). pp. 44–51. IEEE (2019)
13. Hathaliya, J.J., Tanwar, S.: An exhaustive survey on security and privacy issues in healthcare 4.0. Computer Communications **153**, 311–335 (March 2020)
14. Hölbl, M., Kompara, M., Kamišalić, A., Nemec Zlatolas, L.: A systematic review of the use of blockchain in healthcare. Symmetry **10**(10), 470 (October 2018)
15. Janjic, V., Bowles, J., et al.: The serums tool-chain: Ensuring security and privacy of medical data in smart patient-centric healthcare systems. In: 2019 IEEE Int. Conf. on Big Data. pp. 2726–2735. IEEE, New York, NY, USA (2019)
16. Larrucea, X., Moffie, M., Asaf, S., Santamaria, I.: Towards a gdpr compliant way to secure european cross border healthcare industry 4.0. Computer Standards & Interfaces **69**, 103408 (March 2020)
17. Meingast, M., Roosta, T., Sastry, S.: Security and privacy issues with health care information technology. In: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5453–5458. IEEE (2006)
18. Mettler, M.: Blockchain technology in healthcare: The revolution starts here. In: 2016 IEEE 18th international conference on e-health networking, applications and services (Healthcom). pp. 1–3. IEEE (2016)
19. Miyachi, K., Mackey, T.K.: hOCBS: A privacy-preserving blockchain framework for healthcare data leveraging an on-chain and off-chain system design. Information Processing & Management **58**(3), 102535 (May 2021)
20. Rhahla, M., Allegue, S., Abdellatif, T.: Guidelines for gdpr compliance in big data systems. Journal of Information Security and Applications **61**, 102896 (Sept 2021)
21. Seh, A.H., Zarour, M., Alenezi, M., Sarkar, A.K., Agrawal, A., Kumar, R., Khan, R.A.: Healthcare data breaches: Insights and implications. Healthcare **8**(2), 133 (May 2020)
22. Trivedi, K.S., Kim, D.S., Ghosh, R.: Resilience in computer systems and networks. In: 2009 IEEE/ACM International Conference on Computer-Aided Design-Digest of Technical Papers. pp. 74–77. IEEE (2009)
23. Webber, T., Mendoza-Santana, J., Vermeulen, A.F., Bowles, J.K.F.: Designing a patient-centric system for secure exchanges of medical data. In: Int. Conf. on Computational Science and Applications (ICCSA 2020). LNCS, vol. 12254, pp. 598–614. Springer, Cham (2020)
24. Yue, X., Wang, H., Jin, D., Li, M., Jiang, W.: Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control. Journal of medical systems **40**(10), 1–8 (August 2016)