Declaration on Plagiarism

Assignment Submission Form

This form must be filled in and completed by the student(s) submitting an assignment

Name(s):	Vinit Saini
Programme:	MSc in Computing - Blockchain
Module Code:	CA640I
Assignment Title:	Securing Personal Data with Blockchain Technology: A Review of Literature
Submission Date:	21 October 2022
Module Coordinator: Renaat Verbruggen	

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. I/We have read and understood the Assignment Regulations. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

I/We have read and understood the referencing guidelines found at
http://www.dcu.ie/info/regulations/plagiarism.shtml , https://www4.dcu.ie/students/az/plagiarism
and/or recommended in the assignment guidelines.

Name(s): _____Vinit Saini______________ Date: ___21 October 2022_______

Securing Personal Data with Blockchain Technology: A Review of Literature

Introduction

The past decades have seen phenomenal growth in the area of the internet and related technology. According to Internet World Stats (2022), the total number of active internet users globally has crossed the five billion mark, constituting approximately 69% of the world's population. Further, there is a huge amount of information that is generated every day, which usually consists of different personal attributes, such as names, addresses, phone numbers and financial details. Such data, which are generally regarded as personally identifiable information (PII), are crucial and can help identify individuals. Easily accessible PII is vulnerable to fraud, unwanted promotions, etc., often resulting in financial damages. PII could be used to access someone's private life by gaining access to photographs, financial information or business details. In extreme cases, a leak of personal information could lead to targeted attacks, fraud or slander. In a data breach recently, an adult cam site was compromised and the personal data of over 10 billion users was exposed, which left the users with a lot of anxiety and worry (Barrett, 2022). Statistics show an alarming trend in these attacks with ease of access to the internet and a spurt in online activities (Statista, 2022). Managing PII online is a growing concern, which must be challenged in light of the new technology and development. Blockchain technology and self-sovereign identity management systems promise to address these concerns. Hence, I evaluated several research papers to identify how Blockchain technology could be used to secure PII and what is the current state of the art of self-sovereign identity systems.

This literature review highlights four peer-reviewed papers, retrieved from reputed publishers after carefully evaluating a bunch of related papers for their industry challenges and solutions. Each selected paper cites a specific set of problems from the relevant industry and proposes a viable solution. Interestingly, all papers suggest the use of blockchain technology to solve their discrete problems. However, there are different strategies and contrasts in how they approach solutions. Table 1 summarises a quick comparison.

Although several papers contributed to the development of concepts, not all have been included in this review. For instance, the paper of Hariharasudan and Quraishi (2022) lacks cohesion, while the work of Pöhn, Grabatin and Hommel (2021) and Mukta et al. (2020) describes their specific use cases exhaustively. Therefore, have been excluded in the favour of a concise and focused review.

Technology Overview

Nakamoto (2008) coined the term blockchain through the Bitcoin cryptocurrency technology whitepaper. Nowadays, Almost every industry recognises some valid use cases where blockchain technology can be utilised to simplify the business processes like Trading, Finance, Shipping, Transport, IoT, Healthcare, Supply-chain and many more. Blockchain is a distributed data structure comprising several connected nodes called blocks. Each node has a parent node except the genesis node. Thus, it can be seen as an ever-growing long chain of connected nodes. Many characteristics make the blockchain technology ideal for business use cases that needs immutability, traceability, verifiability, transparency and decentralisation. Every node in the blockchain contains a cryptographically secured hash value which further contains the reference of its parent node hash and so on, which makes it immutable. The tiniest change in any block impacts the whole chain connected afterwards and hence it is easily detectable. Since all of the nodes are connected to their parent, this data structure provides complete traceability to the very first node of the transaction. Every transaction is timestamped and cryptographically signed through a private key that provides easy verifiability through the public key infrastructure methods (Haber and Stornetta, 1991, pp. 437-455). Likewise, Transparency can be achieved as all the transactions are on a public ledger and can be viewed and verified by anyone. Furthermore, Decentralisation is one of the biggest advantages of blockchain technology as there is no single point of authority to authorise any transaction on the network. Decentralisation of a typically centralised identity ecosystem can be achieved through blockchain technology. Such decentralised identity systems are known as self-sovereign identity (SSI) systems which provide more control, privacy, and security for the end users. SSI systems use the concept of decentralized identifiers (DIDs) and verifiable credentials (VCs) as building blocks. DID can be seen as a globally unique identifier to locate a DID document on the blockchain, which contains identity information such as cryptographic signatures, verification methods and service endpoints. Structure of DID is `did:<DID method>:<method-specific identifier>`. DID refer to the identity information whereas VC is bound to a DID through its DID document, which means to provide trust to the identity. VCs are typically obtained through a trusted entity like a university, passport department etc. which are cryptographically verifiable. VCs also provides a means to selective disclosure through a verifiable presentation which allows user to disclose the subset of attributes present in a VC.

Challenges

According to Schlatt *et al.* (2021), methods to obtain PII are not secure and prone to information leakage. KYC (know your customer) is a well-known process in the finance industry which is commonly used by banks, insurance and utility providers to obtain personal information from prospective clients. It is not only a tedious and time-consuming process but also demands strict verification of sources and data validation during collection. Further, the KYC process may include multiple stages to verify and synthesise the data which makes it highly vulnerable to fraud and data breaches. Nevertheless, this process is very important for verifying the customer before they can access resources and services. Institutions must trust their customers before providing any services like mortgages, credits and other financial services. Similarly, Customers also need assurance from the institutions for the safety of their data. Schlatt *et al.* (2021) outlined a few typical methods used by institutions to capture KYC information including taking photographs, identity documents obtained from government bodies, other verifiable documents and sometimes in-person verification. However, These methods have potential problems like data cloning and loss of documents, are time-consuming and sometimes in-person visits are just not feasible. Furthermore, such processes are error-prone, cost-intensive and have a high probability of fraud if not done with caution. Similar concerns have also been raised by Xu *et al*. (2020) in the telecom industry, where users undergo a mandatory registration process with their local network operator and furnish their personal information to receive a unique International Mobile Subscriber Identification (IMSI) number and corresponding symmetric keys to get the network access. Network providers store a large amount of customers' PII to authenticate the user's requests, and to facilitate a seamless experience for their users while they use various network services. Similar to Schlatt *et al.* (2021), Xu *et al*. (2020) also questioned the current authentication methods which are dependent on symmetric encryption, the operator needs to authenticate every request against the user's symmetric keys stored in their databases, which makes the authentication service or database a single point of failure. Such databases store millions of users' data, which makes them highly vulnerable and susceptible to attacks, any security compromise could result in a catastrophic impact. Xu *et al*. (2020) highlighted authentication and access management of users are mandatory to safeguard the network and to watch unfair usage. However, the traditional authentication method appears sub-optimal. Comparably, Shuaib *et al.* (2021) also raised concerns about the methods and limitations of healthcare information systems to obtain data from the latest information sources. Modern data generators like smart medical devices, wearable devices and mobile devices tend to generate a lot of low-level data like health metrics, routine sensory data etc. in a combination of patient's identity. Such low-level personal data poses many risks regarding information sharing with trusted partners, security compliance and regulatory compliance. The current state of the healthcare information systems does not provide any way to securely correlate this data, which causes patients to use third-party applications that are often untrusted and might lead to information leakage. Shuaib *et al.* (2021) also pointed out that most identity management systems deployed in healthcare institutions are based on traditional client-server architecture, which is often centralised and a single point of failure. Again, such systems are vulnerable to various attacks or data breaches, which puts the privacy of patients' PII under serious threat. The risk of a single point of failure is also highlighted by Liu *et al.* (2021) in media management systems. They argue, In recent years the use of online media has seen tremendous growth. Online media content is largely unstructured and spread in various forms i.e. text, audio, video, graphics and animation. It has been noted that most of such online data is being managed through media platforms, which are generally based on traditional client-server technologies often managed centrally. Single point of failure is one of the biggest concerns of such centrally managed architectures.

Security and control of data in such centralised or centrally-managed systems are also questionable, if the multimedia service provider is compromised or if someone tempers the original content, there is no way for the owner to restore the media content. In such scenarios, the integrity of the content can be challenged and it is tough to track the changes (Liu *et al.*, 2021). Once the media content is transferred to the multimedia providers, the owner of the content is left with very little control over it. Further, there is no easy method to track the potential alterations to the original content which leads to copyright infringement. The author also brought up the point of stats manipulation on the media like the number of likes/views used to calculate the royalties and earnings on the content. The process of tracking all these attributes is completely at the will of the multimedia providers, the data owner has very little control over such methods. Xu *et al*. (2020) also echo a similar concern, traditional identity management systems used by network operators provide minimal data control to users after their registration process is over. The network providers hold on to the user's PII, further this information resides in the data centres owned by the network providers only. Any modification to existing information needs to be gone through a tedious process laid out by the operator. Further submission of PII documents is often needed for re-verification purposes. On a similar note, Shuaib *et al.* (2021) also raised the limitation of classic healthcare systems, where patients have no control over their information. The author argues that current technology and health information systems are inefficient in processing patients' digital information. Also, the privacy of patients' data is in the hands of healthcare service providers only. There is no easy way for a user to know how healthcare service providers process their data and how they share it with other parties. Patients have no other choice than to trust health providers. Shuaib *et al.* (2021) have summarised the limitations of the current healthcare system very well. However, one important point has been overlooked in the paper. Traditional healthcare information systems tend to perform poorly in terms of efficiency while retrieving existing information especially. Patient information retrieval is a complex and time-consuming process which often requires the involvement of healthcare staff, once the information is obtained it must be sent to the patient securely. There is no straightforward way for the patients to go and retrieve their information on their own. Apart from this, data-in-transit is highly susceptible to illicit alterations and data breaches as it travels in the human-readable form mostly. Furthermore, the system provides no way to generate an alarm or revert those changes in case of any suspicion.

Too much information disclosure is another challenge identified by Xu *et al.* (2020), if users want to use the services of a foreign network, the local operator is required to share the user's PII with the foreign network operator. Such dependencies often lead to unnecessary information disclosure. The user has limited visibility on what information has been exposed by the local operator. PII is very crucial and could be easily misused. Xu *et al.* (2020) also claim that operator portability is another challenge as the process involves the re-submission of PII in other applications which were earlier associated with previous identity or IMSI. However, I'm not fully convinced of this thought as this requirement is eradicated to a great extent nowadays. Nevertheless, How much information interchanges during the portability process is still a valid concern. Besides, Schlatt *et al.* (2021) also mentioned the repetitive nature of the KYC process as there is no common trustable entity and defined format in the finance industry that can provide personal data to all institutions securely and safely. Thus, users of multiple institutions often end up providing more information than expected to institutions. Furthermore, sometimes institutions outsourced KYC compliance to third parties specialised in collecting data which might lead to obtaining unnecessary information. However, best practices in data collection and handling of collected data are still a big concern. Similarly, healthcare systems are required to share patients' PII with many stakeholders like insurance companies to settle claims, pharmacies for providing drugs etc. which might involve the risk of unnecessary information disclosure due to the lack of methods which could provide selective disclosure of information.

Towards solution

Emerging blockchain-based self-sovereign solutions (SSI) is a novel idea aiming to revolutionise data ownership. SSI aim to provide complete control to individuals to manage their online identities, which leverages users to add, edit, and delete their PII without trusting any central authority. Advanced controls like where data is being stored, who can access data and what data is accessible are also available to the users. A prototype of a blockchain-based SSI solution has been presented by Xu *et al.* (2020), which allows users to self-generate and control their self-sovereign identities which can be verified later through trusted entities. Xu *et al.* (2020) prototype demonstrates how operators can authenticate users based on SSIs and verifiable claims. On successful authentication, the user's SSI and corresponding public keys are added to the blockchain by network operators, which can also be queried and verified by others as they are publicly available. The chameleon hash technique has been suggested for revoking users from services instead of maintaining the revocation lists of users which is a storage overhead (Xu *et al.* 2020). Additionally, it is quite efficient for an operator to mark a user's SSI as revoked. Also, blockchain is a distributed and immutable data structure, which helps in securely spreading this change to all interested stakeholders. Likewise, Schlatt *et al.* (2021) advocate blockchain-based solutions to simplify the KYC process. Transparency, immutability, auditability and easy verification of data are key characteristics of blockchain technology, making it very lucrative to build KYC solutions. However, a contradiction in respect of GDPR compliance has also been identified by Schlatt *et al.* (2021). 'Right to be forgotten states, information should be deleted once its intended purpose is over. For example, once the contract with a bank is over user may ask the bank to remove all his personal information from their records. As blockchain is an immutable data structure, it doesn't allow information to be removed once written. Therefore, Schlatt *et al.* (2021) consider the self-sovereign identity (SSI) concept as a more suitable alternative to implementing KYC solutions which provide a neutral platform, distributed governance and easy data verification. SSI solutions are based on blockchain technology, which provides essential benefits like decentralisation, security and privacy. Furthermore, self-sovereign identity provides easily revocable access rights, allowing users to revoke existing access permissions anytime. The unavailability of generic design principles (DPs) for further development of self-sovereign identity solutions has been highlighted by Schlatt *et al.* (2021), which motivated them to introduce three design principles through their study, that can be summarised as,

1. Minimal involvement of the blockchain suggests not storing personal information on the blockchain in the shape of DIDs or VCs in compliance with privacy regulations.
2. Interoperability of different blockchains, there could be many different blockchains that exist therefore consideration must be given to their interoperability during the design phase.
3. Flexibility to use centralised services, there could be scenarios where centralised services e.g. cloud service are required to perform some specialised operations.

DPs presented by Schlatt *et al.* (2021) are one of the major contributions of their study and can be used across industries to design SSI solutions, specific to their use cases. Moreover, Shuaib *et al*. (2021) focus on trust, transparency, security and compliance which are absolute requirements for any identity management system in healthcare. Therefore, they strongly favour SSI management systems that can enable patients to have full control of their digital identity and data. SSI management systems provide a quick and efficient way for identity verification, which ensures easy and trustworthy verification by different stakeholders like doctors, pharmacies and insurance companies in the healthcare domain. Further, such systems can also aid in the efficient retrieval of existing information quickly, which results in minimum data verification effort from healthcare staff. Similarly, Liu *et al*. (2021) also recommend the use of SSI management systems to overcome the challenges related to media management. SSI solutions are decentralised and provide autonomous control to the author of data. Characteristics of blockchain like Traceability, Auditability, Transparency and Immutability all support implementing multimedia service systems that can protect the integrity of content and provide sophisticated access controls.

Additionally, Liu *et al*. (2021) suggest deploying smart contracts which are user-defined programs containing business logic, these smart contracts run on a blockchain network and can execute on triggers and conditions. However, the author overlooked mentioning that the smart contracts also provide repudiation property which can be used to enforce legal accountability, enforce fulfilment of obligations of contract etc. From a multimedia provider's point of view, smart contracts can also be used to automate the maintenance of content for example a smart contract can be written to remove or disable content after a certain date or number of views.

Liu *et al*. (2021) present a prototype of multimedia data management system, having a three-layer architecture consisting of a service layer, an off-chain data layer and an on-chain data layer. Their prototype keeps the copyright data off-chain but the hash of those copyright data is stored on-chain to facilitate the integrity of sensitive data. Further, the prototype suggests keeping sensitive and complex information on-chain. Similarly, Xu *et al.* (2020) also favour keeping users' SSI and public keys on the blockchain to enable other operators to quickly authenticate the users by validating the blockchain. However, both arguments contradict the design principle of minimal involvement of blockchain defined by Schlatt *et al.* (2021). Moreover, it's an arguable question, which should be seen as a possible gap for further research in the context of use cases.

Conclusion

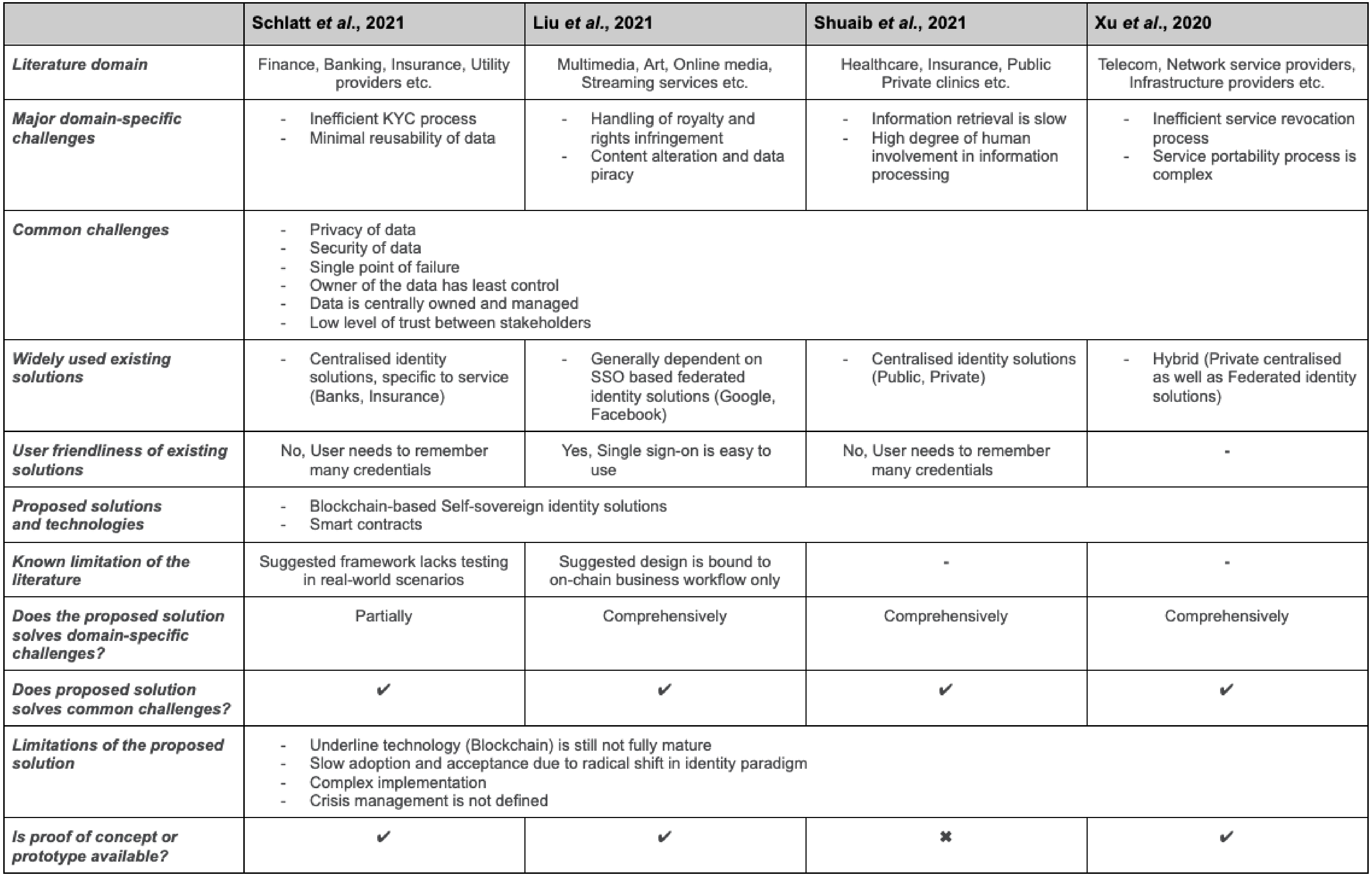
Despite domain-specific challenges, there are a few challenges that are common across industries. Privacy and security of personally identifiable information (PII), selective disclosure of information, data ownership, information accessibility, information verifiability, and revocation of rights are a few of them. Analysis of related literature reveals that centrally managed PII is prone to inefficiencies and highly vulnerable to data leakage, loss and attacks.

Blockchain-based self-sovereign identity (SSI) management systems seem very promising in resolving challenges related to the privacy and security of digital entities and personal data. Decentralisation, immutability, transparency and traceability are the biggest stakes of blockchain technology and its applications. Nevertheless, Blockchain technology in itself is not fully matured at the moment and needs further research in the application context. Further, apart from promising benefits, technology adoption also needs regulatory compliance and general acceptance, which is still distant. The literature review identifies a few unexplored areas where further research is needed, including

- Scalability aspects of SSI solutions in terms of performance and synchronisation of new entities, especially when blockchain reaches a considerable size.
- Interoperability of SSI solutions built on different blockchains (blockchain compatibility)
- Migration of SSI solutions built on one blockchain to another (blockchain portability)
- Crisis management in such solutions

More and more pilot projects and blockchain-based deployments in real-world applications will uncover new questions and potential solutions to the problems.

Table 1: Comparison summary



	Schlatt *et al.*, 2021	**Liu *et al.*, 2021**	**Shuaib *et al.*, 2021**	**Xu *et al.*, 2020**
Literature domain	Finance, Banking, Insurance, Utility providers etc.	Multimedia, Art, Online media, Streaming services etc.	Healthcare, Insurance, Public Private clinics etc.	Telecom, Network service providers, Infrastructure providers etc.
Major domain-specific challenges	- Inefficient KYC process - Minimal reusability of data	- Handling of royalty and rights infringement - Content alteration and data piracy	- Information retrieval is slow - High degree of human involvement in information processing	- Inefficient service revocation process - Service portability process is complex
Common challenges	- Privacy of data - Security of data - Single point of failure - Owner of the data has least control - Data is centrally owned and managed - Low level of trust between stakeholders			
Widely used existing solutions	- Centralised identity solutions, specific to service (Banks, Insurance)	- Generally dependent on SSO based federated identity solutions (Google, Facebook)	- Centralised identity solutions (Public, Private)	- Hybrid (Private centralised as well as Federated identity solutions)
User friendliness of existing solutions	No, User needs to remember many credentials	Yes, Single sign-on is easy to use	No, User needs to remember many credentials	-
Proposed solutions and technologies	- Blockchain-based Self-sovereign identity solutions - Smart contracts			
Known limitation of the literature	Suggested framework lacks testing in real-world scenarios	Suggested design is bound to on-chain business workflow only	-	-
Does the proposed solution solves domain-specific challenges?	Partially	Comprehensively	Comprehensively	Comprehensively
Does proposed solution solves common challenges?	✔	✔	✔	✔
Limitations of the proposed solution	- Underline technology (Blockchain) is still not fully mature - Slow adoption and acceptance due to radical shift in identity paradigm - Complex implementation - Crisis management is not defined			
Is proof of concept or prototype available?	✔	✔	✖	✔

Bibliography

Barrett, B. (2020) *Hack Brief: An Adult Cam Site Exposed 10.88 Billion Records.* Available at: https://www.wired.com/story/cam4-adult-cam-data-leak-7tb (Accessed: 21 October 2022).

Haber, S. and Stornetta, W.S. (1991) 'How to Time-Stamp a Digital Document', *Advances in Cryptology-CRYPTO' 90. CRYPTO 1990. Lecture Notes in Computer Science,* 537, pp. 437-455. Available at: https://doi-org.dcu.idm.oclc.org/10.1007/3-540-38424-3_32

Hariharasudan, V. and Quraishi, S.J. (2022) 'A Review on Blockchain Based Identity Management System', *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 735–740. Available at: https://doi.org/10.1109/ICIEM54221.2022.9853110

Internet World Stats (2022) *World Internet Users Statistics and 2022 World Population Stats*. Available at: https://www.internetworldstats.com/stats.htm (Accessed: 21 October 2022).

Liu, Y., Lu, Q., Zhu, C. and Yu, Q. (2021) 'A blockchain-based platform architecture for multimedia data management', *Multimedia Tools and Applications*, 80(20), pp. 30707–30723. Available at: https://doi.org/10.1007/s11042-021-10558-z

Mukta, R., Martebs, J., Paik, H., Lu, Q. and Kanhere, S.S. (2020) 'Blockchain-Based Verifiable Credential Sharing with Selective Disclosure', *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 959–966. Available at: https://doi.org/10.1109/TrustCom50675.2020.00128

Nakamoto, S. (2008) 'Bitcoin: A peer-to-peer electronic cash system', *Decentralized Business Review,* p. 21260. Available at: https://www.debr.io/article/21260.pdf (Accessed: 21 October 2022).

Pöhn, D., Grabatin, M. and Hommel, W. (2021) 'eID and Self-Sovereign Identity Usage: An Overview', *Electronics*, 10(22), p. 2811. Available at: https://doi.org/10.3390/electronics10222811.

Schlatt, V., Sedlmeir, J., Feulner, S. and Urbach, N. (2022) 'Designing a Framework for Digital KYC Processes Built on Blockchain-Based Self-Sovereign Identity', *Information & Management*, 59(7), p. 103553. Available at: https://doi.org/10.1016/j.im.2021.103553

Shuaib, M., Alam, S., Alam, M.S. and Nasir, M.S. (2021) 'Self-sovereign identity for healthcare using blockchain', *Materials Today: Proceedings.* Available at: https://doi.org/10.1016/j.matpr.2021.03.083

Statista Research Department (2022) *Cyber crime: all-time biggest online data breaches 2022.* Available at: https://www.statista.com/statistics/290525/cyber-crime-biggest-online-data-breaches-worldwide (Accessed: 21 October 2022).

Xu, J., Xue, K., Tian, H., Hong, J., Wei, D.S.L. and Hong, P. (2020) 'An Identity Management and Authentication Scheme Based on Redactable Blockchain for Mobile Networks', *IEEE Transactions on Vehicular Technology*, 69(6), pp. 6688–6698. Available at: https://doi.org/10.1109/TVT.2020.2986041