Decentralized verification infrastructure for documents anchored to blockchain

Mirko Stanić[1], Matija Pužar[2]

[1] Agency for Science and Higher Education, Donje Svetice 38, Zagreb, Croatia, mirko.stanic@azvo.hr
[2] Unit – The Norwegian Directorate for ICT and Joint Services in Higher Education and Research, Fridtjof Nansens vei 19, 0369 Oslo, Norway, matija.puzar@unit.no

Keywords
digital documents, digital credentials, blockchain, education, distributed ledger

1. ABSTRACT

Digitizing student credentials presents several unique problems. Traditionally, the issuers must provide infrastructure for hosting digital documents or choose to outsource it to a third party. In this solution, dangers of potential data breach can never be fully mitigated and the validity of these documents is automatically tied to the existence of the institution that issued them. With the advent of blockchain technology it has become possible to store proofs of existence on a permissionless distributed ledger, i.e. blockchain, thus eliminating the need for hosting complex infrastructure as well as rendering data breaches impossible and enabling ownership of the documents to be efficiently managed in the digital space.

2. INTRODUCTION

In order for any digital data exchange format and/or protocol to gain traction, it needs something which can be best described as “display infrastructure”, i.e. software that can read and/or process received files. In use cases which go beyond record exchange between institutions, such as admissions, diplomas etc., it is unreasonable to expect that every stakeholder who wants to verify validity of a document will have the ability or even the desire to host a dedicated verification software. Such an endeavor would require time and resources and also expose the stakeholder to potential data theft. Some countries solve this by means of centralized web portals through which a user with an account can give access to their records to another user. Ignoring the potential data breach of such a solution, there are many other costs included, be they of financial or resource based nature. One also needs to take into account the need for user authentication. This is again sometimes solved through centralized eID systems that are in some cases run by for profit corporations, thus introducing further issues of user tracking and data harvesting. Even in countries where both the eID system and the central data repository are government owned, there is the problem of sharing data with someone in another country.

The concept of issuing educational credentials on the blockchain is based on publishing digitally signed hashes of XML, PDF, JSON or other files, containing information about the credential. The published hash has a twofold purpose, to provide an immutable timestamp of when was the document issued and to ensure that the digital file issued to the user has not been tampered with. In this use case, a person is issued a file whose digital fingerprint is recorded in a transaction on the blockchain. It is important to note that the system being presented here is completely independent of the actual software implementation of the blockchain and that it would be wrong to cater towards one specific blockchain architecture since the only property of a blockchain that is of actual benefit is its immutability. We state that full vendor independence must be observed when implementing long term projects in the digital area to mitigate against potential obsolescence of certain implementations.

3. STATE OF THE ART

In this chapter, we present a few examples of state of the art related to our work.

- **Blockcerts standard**, developed by MIT and Learning Machine, is the most talked about example of credentials on the blockchain, which currently anchors credentials on the Bitcoin blockchain and plans to support Ethereum in the future as well. A Bitcoin transaction basically consists of an input address or addresses and output address or addresses. At some point users started utilizing the immutability of the chain to store information other than transactions. In the cases where proof of existence was needed, this was done by hashing the document whose existence and validity they wanted to prove and using that as an output address in a transaction. The problem of verifying the authenticity of the issuer is solved by having the issuer host their public key on their website, which introduces a centralized point in a decentralized system. MIT and Learning Machine are aware of these issues and they are working on their solution which they see in the form of smart contracts supported by Ethereum and decentralized IDs (DID).
- **Gradbase** is a UK based company, which issues credentials on the Bitcoin blockchain and charges potential employers for verifying their candidates. Their solution is proprietary. It uses internal formats for record storage and the credentials are not owned by the receiver but by Gradbase, thus making it highly vendor dependent.
- **Sony Global Education** The offers a proprietary solution based on IBM Fabric ledger technology. Completely closed source, closed format and vendor dependent.
- **Attores Solutions** is a Singapore based company specializing in custom solutions tailored to individual institutions utilizing Etherium blockchain. Vendor dependent, and closed format.
- **Accredible** is a U.S. company offering institutions credential issuance on a monthly subscription basis. They support Mozilla Open Badges, but are also utilizing closed format and practice vendor dependency.

4. DATA AND VERIFICATION

When talking about digital credentials we see two problems which are being discussed as one, when in reality they should not be. The first problem is data representation. Data should be written in a standardized format both from a technical and ontological perspective, in order to facilitate easier processing. The second problem is data verification. Everyone should be able to independently verify the origin and integrity of a digital document. This means establishing that the document was written or issued by the institution or individual that is stated as the issuer, and that the contents of that document have not been altered after its issuance.

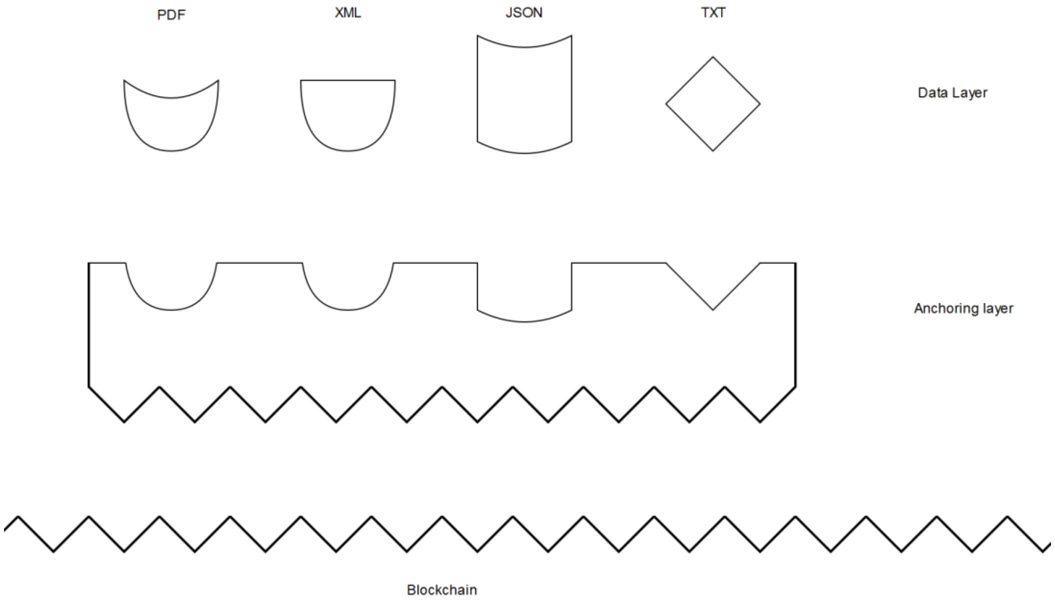

Figure 1. Anchoring layer serving as a middleman

Before discussing any technical details, we must make a clear separation between data representation and data verification. There should be no limitations imposed on the structure of data being carried (the payload) by the process of verifying said data. To achieve this, the data necessary for anchoring the credentials themselves must be kept to a minimum. This anchoring layer acts as a container for the actual data formats which contain useful data. A graphical representation of this setup is shown in Figure 1.

5. NOVELTY AND PERMANENCE

We have to take into consideration the expected lifespan of digital credentials, which should at minimum be the expected working age of the individual holding the credential, about 50 years. This is where we are faced with a decision on whether to build for the future or for the present.

It is very easy to fall prey to the latest technological trends, but the expected lifespan of these trends must also be taken into consideration. At this moment, we are living in the social era of the Internet, where everything is being viewed and graded through a perspective of "*shareability*" and self-exposure. There is a lot of talk about technologies such as Open Badges and integration into popular social networks such as LinkedIn and Facebook. This is a legitimate use case and one worth pursuing, however it is fundamentally wrong to tie the structure of a digital credential to a single data structure only to facilitate a single use case. It must be understood that these social networks are privately owned enterprises which will continue to exists only as long as they are profitable. Even in this early stage, there are already examples of outdated social networks such as Myspace, and it is hard to imagine that the industry landscape will remain fixed for over five decades.

The underlying data must be made accessible in a number of different formats, in order to enable graceful degradation of the user experience should certain standards become obsolete. At the very bottom we have a simple statement, something that a credential in reality is, encoded in UTF-8 as a basic text file. It is difficult to imagine a more basic format in existence today. On top of that, we stack JSON, XML, PDF, Badge, Verifiable Credentials, ProgressiveApp or any other form of data presentation in order to facilitate use cases for the present. If and when a format becomes obsolete, we can fall back to a lower level presentation.

Another noticeable trend, one with regards to blockchain itself, is the unnecessary push towards "wallets" as storage and distribution hubs for digital credentials. To understand this, we need to look back to the first use cases for blockchain technology, used in the financial sector, where the knowledge of private keys tied to a certain transaction output is the only way to prove the ownership of an output of a financial transaction. However, once we start to use the immutability of the blockchain to record hashes, we no longer need this wallet model because verification of a credential does not necessitate spending any transaction outputs – only the issuer has the need for recording transactions and therefore private keys. Also, unlike currency, credentials are issued in name and are not fungible, therefore there is no need to provide proof of ownership through a private key. If ownership of a document is tied to the knowledge of the secret key, that also implies that if a person were to acquire someone else's secret key, they would also take ownership of their credential. By tying a digital credential to an application like a wallet, we are introducing a centralized point of storage, like an app, which is developed by a small group of people and has no guarantees for long term support.

6. PUBLIC VS. PRIVATE

Blockchain is a protocol for synchronizing content of a distributed database between parties who do not know each other and or do not trust each other. From this definition it stands clear that there can be no private blockchain, because in order for it to be private users identities have to be known, and consequently malicious activity can be traced to a specific user. With this in mind, it is worth noting that this does not mean that a private distributed ledger, which a private blockchain really is, has no use cases. It is possible to forego public blockchain in favour of a private distributed ledger although this option would require more development and maintenance resources.

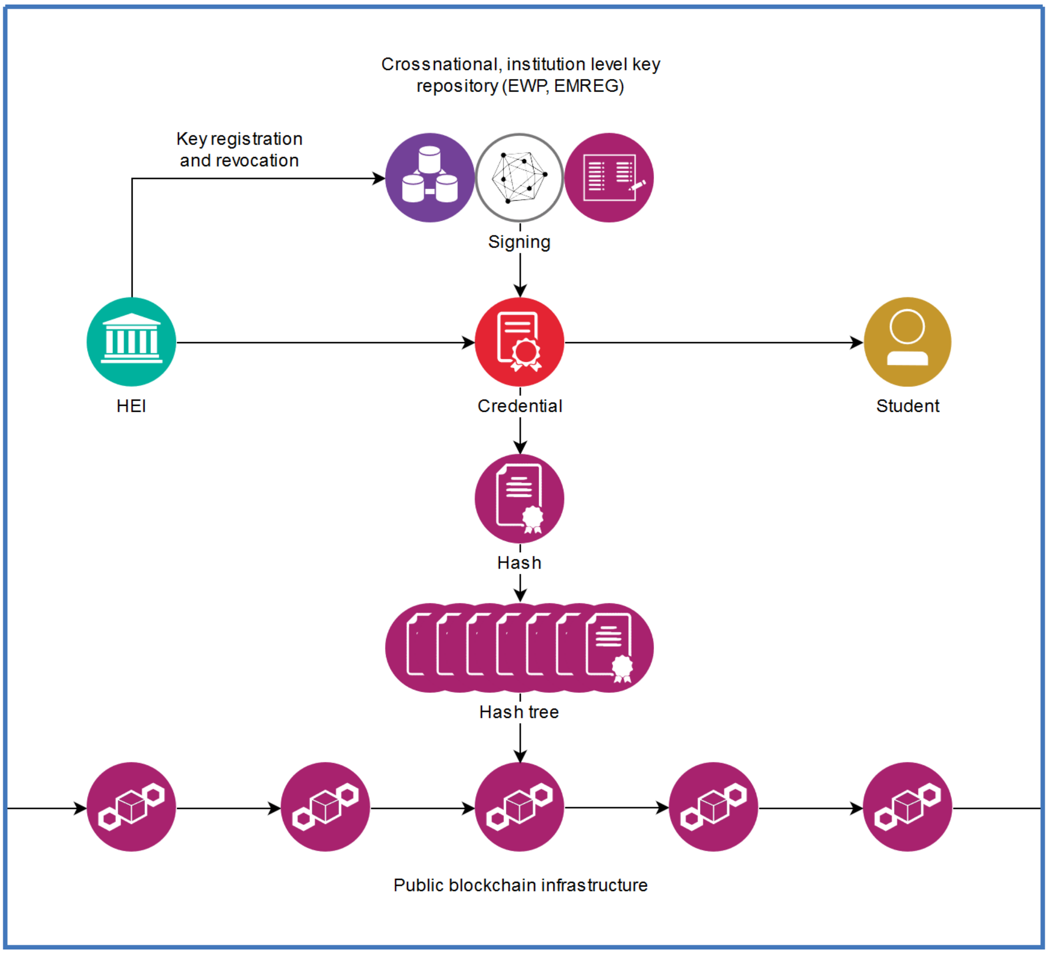
Crossnational, institution level key repository (EWP, EMREG)
Key registration and revocation
Signing
HEI
Credential
Student
Hash
Hash tree
Public blockchain infrastructure

Figure 2. Document issuing

7. DOCUMENT ISSUING

We recognize that the ownership of any document is shared between the issuing institution and the receiving individual, and that in some cases, such as credentials, the issuing institution reserves the right to revoke the document. With this in mind, documents would be hashed and the hash would be signed with the institution's private key to verify their origin. The institution would use a different private-public key pair for each document derived from a single master seed key in what is known as a hierarchical deterministic key pool. The resulting data would then be broadcast to the rest of the network and added to the blockchain. This scheme allows the issuing institution to revoke a document without the need for accessing it. The result owner, on the other hand, controls to whom and when he or she will show his or her document. The storage of public keys is delegated to a distributed database which is mirrored across multiple institutions, this way the disappearance of a single institution does not invalidate the credentials issued by that institution. We hope to be able to utilize the existing institution registries developed as part of other EU level projects such as the EMREG registry developed as a part of the EMREX project for electronic student data exchange between institutions or the Erasmus Without Paper registry. Also, the newly formed EBSI (European Blockchain Services Infrastructure), under the European Blockchain partnership consortium could be used as a storage layer for institutional keys as well. This part of the specification is envisioned as open with the goal being to provide as many alternative storage methods as possible. The entire process is outlined in Figure 2. The relation between the credential and hash is not on a 1 to 1 basis but 1 to N. By utilizing hash trees, we can store only the root of that tree on the blockchain. This tree can also contain dummy documents in order to further obfuscate the actual number of graduating students.

8. DOCUMENT VIEWING

The display of document is done by checking the hash of the individual file against the one stored on the blockchain through a verification app which can be made available as a module hosted on a web site, CMS plugin, mobile app or something else. The actual physical display of the document represents the most difficult problem in the process. The verification process is shown in Figure 3. Current solutions rely almost exclusively on smartphone apps which bring a whole set of problems with them, from modification of viewer apps and display of false documents to the theft of users' secret keys by malicious apps or by transmitting touch inputs. The loss of phone or simple hardware failure is also a problem.

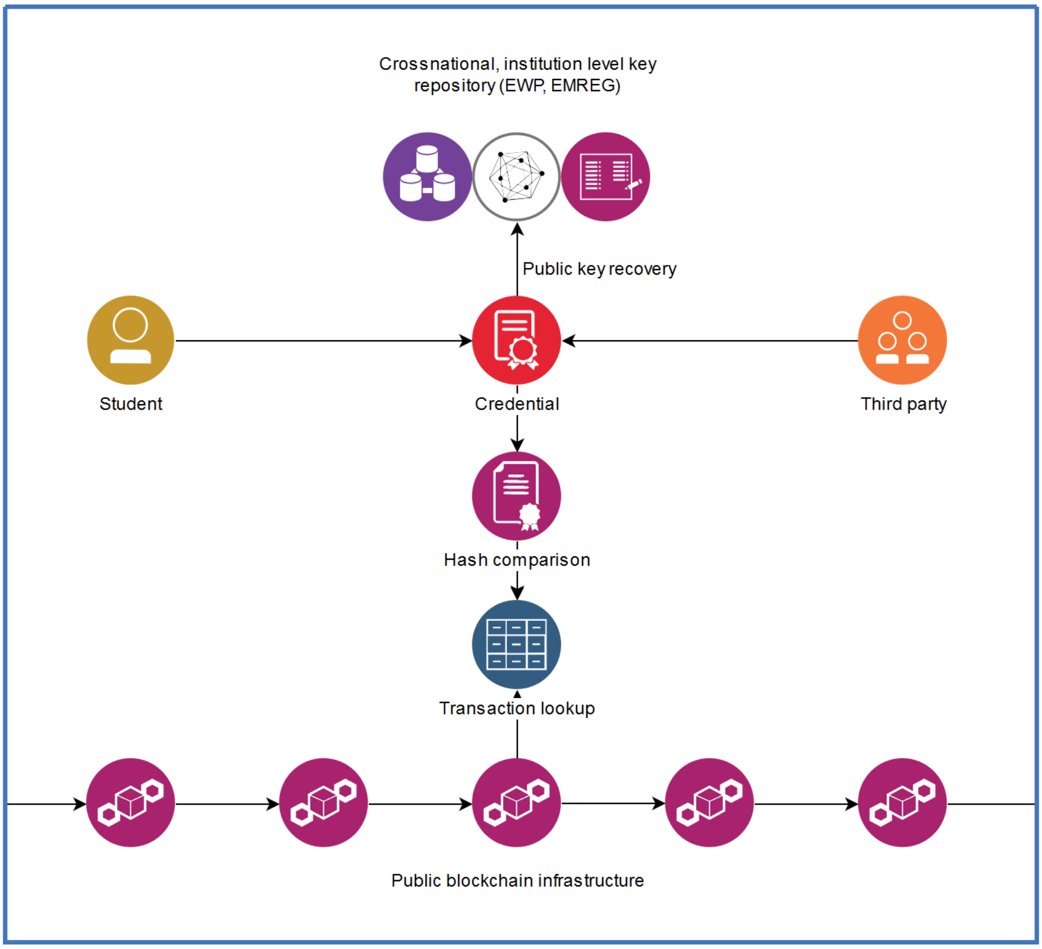
Crossnational, institution level key repository (EWP, EMREG)
Public key recovery
Student
Credential
Third party
Hash comparison
Transaction lookup
Public blockchain infrastructure

Figure 3. Document verification

The computer file system and concept of folders and files is already an abstraction of data representation and we see no need for introducing a middle layer of specialized applications into this process. The software developed here utilizes the ELMO format which is based on the CEN standard EN 15981-2011 EuroLMAI for data formatting and storage. We chose ELMO as it is a mature platform already being utilized in the EMREX and Erasmus Without Paper projects, and also for its support for diploma supplement, since we believe that this further exemplifies our vision for long term storage of records and is aligned with the overall vision of self-sovereign identity use cases. The other important thing to keep in mind is that in order for any digital format to become widely accepted, the cost of viewing/consuming/verifying content has to be minimal. Moreover, if someone wants to verify the validity of a digital document, they should not be required to host and maintain their own verification software. As mentioned before, verification is handled through a web based display and verification software, although offline verification is also supported for remote locations and to further decentralize the process. This software can be hosted by anyone and it is only important for the verifier to trust the institution hosting the software. For example, if a university hosts it on its public website that can be considered a trustworthy host, software can also be hosted by other trustworthy stakeholders such as ministries, accreditation agencies, recruitment agencies, companies etc.

9. DOCUMENT REVOCATION

The revocation of issued documents is achieved through protocol specification. We specify that the blockchain is read from the newest block to the last. If a credential needs to be revoked, the transaction pointing to that document is written to the blockchain. Since the reading is done last to first, the revocation notice will be read before the credential proof and therefore the validator knows that the credential in question is invalid.

10. DIVERSIFICATION

The point of having multiple overlapping functionalities in a decentralized system is not only to guarantee availability during disruptions, but also to safeguard against technological obsolescence. For this reason, we envision the verification process as not being directly tied to the Internet infrastructure. In areas where Internet is unavailable or in case of outages, verification can be performed via satellite receiving dishes. There is already a built network of satellites which enable offline crypto currency transactions, and by utilizing this existing infrastructure we can further decentralize the verification process. We understand that such a network is highly centralized in its nature purely from a financial standpoint, but for us it provides an alternative means of distribution which helps us decouple the process of verification from the Internet.

The storage of issuer keys is also diversified among various centralized repositories which already exist to facilitate student exchange. Another possible avenue of storing them is the upcoming EU level EBSI (European Blockchain Services Infrastructure) distributed ledger. We can even envision a scenario, where through legislative acts or public announcements a transaction id used as the seed could be publicly announced at the beginning of a set time interval, such as academic year, and all credentials published using that seed would be valid. In this way both the hashes and the keys can be stored on the blockchain.

11. CONCLUSION

Our proposed architecture for decentralized credential management and verification provides coverage on the full lifecycle of a credential whilst trying to adhere to the principles of decentralization wherever possible. We posit that the widespread adoption of verification software is crucial for creating a healthy ecosystem among receivers and validators. With this in mind, and in order to maintain high flexibility, we envisioned the protocol as a series of specifications without a hard coded way of implementing it. We see a possible use case for this protocol as an extension of existing services for electronic record exchange such as EMREX, in use cases where there is no direct communication channel between institutions and the transfer has to be done with time delay, and for issuance of user owned copies of issued credentials.

12. REFERENCES

Albeanu, G. (2017). Blockchain technology and education. *On Virtual Learning*, 271.

Crosby, M, et al. (2016). *Blockchain technology: Beyond bitcoin*. Applied Innovation 2, 6-10.

Gräther, W., Kolvenbach, S., Ruland, R., Schütte, J., Torres, C., & Wendland, F. (2018). Blockchain for education: lifelong learning passport. In *Proceedings of 1st ERCIM Blockchain Workshop 2018*. European Society for Socially Embedded Technologies (EUSSET).

Grech, A., & Camilleri, A. F. (2017). Blockchain in education.

Jirgensons, M., & Kapenieks, J. (2018). Blockchain and the future of digital learning credential assessment and management. *Journal of Teacher Education for Sustainability*, *20*(1), 145-156.

Lewenberg, Y, Sompolinsky, Y, & Zohar, A. (2015). *Inclusive block chain protocols*. International Conference on Financial Cryptography and Data Security. Springer, Berlin, Heidelberg

Ocheja, P., Flanagan, B., & Ogata, H. (2018, March). Connecting decentralized learning records: a blockchain based learning analytics platform. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 265-269). ACM.

Robles, K., Appelcline, S. (2016). *Hierarchical Deterministic Keys for Bootstrapping a Self-Sovereign Identity*. Retrieved April 28, 2019 from: https://github.com/WebOfTrustInfo/ID2020DesignWorkshop/blob/master/draft-documents/hierarchical-deterministic-keys-for-bootstrapping-a-self-sovereign-identity.md

Rooksby, J., & Dimitrov, K. (2017). Trustless education? A blockchain system for university grades. In *New Value Transactions: Understanding and Designing for Distributed Autonomous Organisations, Workshop at DIS*.

Skiba, D.J. (2017). *The potential of Blockchain in education and health care*. Nursing education perspectives 38.4 220-221.

Yang, X., Li, X., Wu, H., & Zhao, K. (2017). The application model and challenges of blockchain technology in education. *Modern distance education research*, (2), 6.

Zyskind, G, Oz N. (2015). *Decentralizing privacy: Using blockchain to protect personal data*. Security and Privacy Workshops (SPW), IEEE.

13. AUTHORS' BIOGRAPHIES

Mirko Stanić

Mirko Stanić has a Master's Degree in Information and communication technology from University of Zagreb (2010). He has worked in Central Applications Office since 2011 as the lead software developer on the Croatian Higher Education Admissions system (NISpVU2). His work is divided between working as a developer in software projects and working as a consultant on specialist projects.

Matija Pužar

Matija Pužar is a Senior Security Specialist and developer from Unit, a directorate under the Norwegian Ministry of Education and Research. Matija received his PhD in 2010 from the University of Oslo, where he also worked as a researcher. He has more than 15 years of experience in developing web applications, both back end and front end. In his current position, his focus is on information security and integration services. Matija has been involved in the architectural design and implementation of the EMREX and Erasmus Without Paper solutions.