DMS-XT: a blockchain-based document management system for secure and intelligent archival

Edlira Martiri
Lecturer / Blockchain architect
Department of “Statistics and Applied Informatics”,
FE, UT, Tirana, Albania
edlira.martiri@unitir.edu.al

Gentjana Muça
Blockchain developer
Ambrogio, sh.p.k.,
Tirana, Albania
gentjanamuca@yahoo.com

Abstract

First areas where the blockchain technology dominated were financial sectors for the secure trading, exchanging or supply-chain of assets. Then cryptocurrencies started to exchange not only money, but also objects. They developed the concept of DApps (decentralized applications) introducing the third blockchain generation. Despite all different areas where blockchain can be used today, in this paper we are focused in secure document management. The idea of the system we present, called DMS-XT, is to store not the whole content of a document, but after properly getting an extract from the unstructured content of pdf documents using Information Extraction techniques, and encrypting it, then storing it in the blockchain. Whoever wants to verify the ownership and content, can do so by retrieving and decrypting this information-view stored in the blockchain. To test the system accuracy and performance we suggest applying it in Education, for the secure storage and quality assurance of diplomas for authorship right protection and against plagiarized content.

1. Introduction

We continuously feel we live in times of fake news or alternative truths. Cases when intellectual property is stolen or misused are also frequent feeds we read from different sources in our everyday life. More precisely, fake documents implying fake certifications, degrees or other documents are documents one could easily find with one simple “search” over the internet. For example, one such case is a “diploma mill” with center in Pakistan where thousands of British nationals bought fake degrees. During 2013-2014 were sold around 3,000 qualifications, including master’s and doctorate degrees.

The above problem stands for diplomas issued on paper, and it is valid for digital attestations too [Bbc18]. Up to now, the education system worldwide has tried to protect their documents and make them trustworthy by applying cryptographic mechanisms, such as digital signatures. This mechanism guarantees integrity (the document is not tampered with), authenticity (the owner can be easily verified), and non-repudiation (the owner can not deny he is the owner). These are all necessary properties of a system that stores, manages and protects documents, but if we consider attestations we should think of very long-term functionality, thing that digital signatures cannot offer without a high degree of technical and procedural complexity, with the additional disadvantage of heavily relying on central authorities.

One possible solution to all these issues, is the introduction of blockchain in Document Management Systems (DMS). The blockchain can be considered as a distributed ledger, or a database, containing a list of continuous records, called blocks, connected as a tree structure and secured by cryptographic algorithms (hash functions) [Gat17]. The underlying mechanisms of the blockchain strongly rely on cryptographic apparatus, and mathematical mechanisms. We will briefly describe some of them in Section 2.

The main goal of this paper is to provide the architecture of a developing system for diploma management having as a back-up a blockchain-based solution for document content verification. One of the main features of the system are the inclusion of a plagiarism tool, and a statistics module. The idea is presented in the following main steps: (1) unstructured information from documents in .pdf formats is extracted; (2) information is converted to a structured form resulting in a compact table including necessary fields from the documents; (3) the table data is encrypted and stored in the blockchain. Further details will be given in Section 3. In Section 4 conclusions and further work will be treated.

1. 1. State-of-art

Being a distributed ledger shared among all the nodes in a network, the blockchain was first used in the financial sector for exchanging and trading assets in a secure way and very efficiently because of the short execution time of the transactions. The variety of applications ranges from currency exchange, payments, remittances, loans, crowdfunding to stocks and shares, digital bonds, gold, etc. Other sectors include healthcare, insurance, communications, peer-to-peer storage, identity management, and every day new areas are exploring the inclusion of blockchain in their information systems.

In 2014 blockchain developers made possible the arise of a new network where everyone could enter the global economy allowing them to exchange without intermediaries. Since then the interest in blockchain became important at the government level, starting with Next in 2016 from the Russian Federation, Singapore in 2016 in collaboration with IBM, World Economic Forum in November 2016 to discuss government models developed in blockchain, in 2017 Harvard suggested blockchain as a groundbreaking technology, etc. Blockchain and cryptocurrencies give to developing countries a great opportunity to advance their economy. It would be a great opportunity for our country or similar ones too, even though by now blockchain is not seriously considered by their respective governments [Mar18].

The first HEI (Higher Education Institution) that stored academic certificates was the University of Nicosia, Cyprus. They actually use the bitcoin network [Sha16]. Malta is another country who has adopted blockchains for academic and professional certifications [Csm18]. Their project relies on a successful initiative, Blockcerts, developed by MIT Media Lab Learning Initiative. Blockcerts is an open-source ecosystem for creating, sharing, and verifying blockchain-based educational certificates [Mit18].

Blockchain is extending fast in the Education area. A blockchain system based on Ethereum is implemented in University of Glasgow Scotland, UK to store student grades [Roo17]. Other solutions include TrueRec from SAP, to manage certificates of online courses [Trr18].

All these solutions store in the blockchain all validated transactions, whereas data are not stored. In fact storing large amounts of data in the blockchain would be very costly for the data owner. For that reason, there were created many storage solutions in order to offer cheaper, faster, more secure, more distributed and independent that cloud solutions. Some of the most popular solutions are: (1) ***StorJ***: it is a blockchain-based cloud storage, and it secures documents by encrypting them [Sto18]. Documents are split into partitions where every part is a peer in the network. (2) ***IPFS***: is a blockchain-based File-System, distributed in all peers of the network [Ipf18]. It offers a similar to BitTorrent file exchanging mechanism and versioning similar to Git.

2. Generic blockchain characteristics and technicalities

There are some important characteristics of the blockchain technology, showing its importance and future perspective. Not only the fact of being decentralized, but also terms such as smart contracts, consensus, unchangeability, open source, peer-to-peer, are at the root of the algorithms on which blockchain is built. All these characteristics are essential in document management systems. Some definitions of the characteristics are [Swa15]:

1. *Distribution*: the design allows distribution of blocks and synchronization in the network.
2. *Smart contracts*: are pieces of code executed on the blockchain, consisting of complex instructions written in a programming language and determines the rights of each party in the network.
3. *Consensus*: prior to executing a transaction there exists a consensus between parties, verifying that the transaction is valid.
4. *Data unchangeability*: after a transaction is recorded it cannot change afterword.
5. *Transparency*: in the internet stack, blockchain has added a new layer, the layer of trust, which is a characteristic able to be coded and included in the algorithms of this technology. Basically, trust can be achieved in a trusted network of nodes, guaranteeing its security.
6. *Integrity*: file or data integrity is a very important security principle and in the blockchain world it allows its users to verify data version is unchanged.

2.1 Hash functions

A hash function (H) is a mathematical function that has three attributes [Gat17], [Wan18]. It can take any string as input and produces a fixed-size output. Secondly, it must be efficiently computable, meaning given a string, in a reasonable length of time, one can figure out what the output is. In out blockchain explanation we need hash functions that are cryptographically secure. The cryptographic properties of hash functions are many, but we will mention some in particular: (i) the function is collision-free; (ii) has the hiding property; (iii) it is puzzle-friendly [Men96].

The first property that we need from a cryptographic hash function is that it's collision free. And what that means is that it's impossible, nobody can find values x and y, such that x and y are different, and yet the hash of *x* is equal to the hash of *y* [Mat18]. Collisions in fact exist in every hash function, but it is impossible to find within considerable time, with regular machines and computational power two same hash values from two different inputs. Until now there are no known ways to find faster collisions in a hash function beyond a certain space cardinality.

The second property of hash functions is that they are hiding, i.e. given the output of the hash function H(x), then there is no feasible way to find x [Bak95]. This means the function is irreversible, the output is completely different from the original data, which in turn is hidden or safe even if its hash is exposed.

The binary tree structure is called a Merkle tree after Ralph Merkle who invented it. An important feature of the Merkle trees is that if a data block needs to be proved if it belongs to a Merkle tree, it can be verified if the hashes of the blocks match, from the given block to its parent, from that parent to its parent, until the root is reached. This way it can be verified that the block belongs to the tree. This membership verification can be done in logarithmic time (O (log n)).

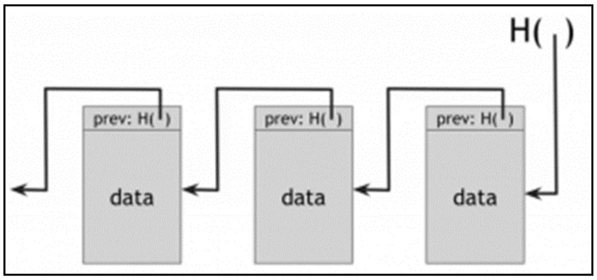
The third property is to be *puzzle-friendly*. This means unpredictability and randomness, i.e. given any data there is no way to tell the value of the hash without calculating it [Tha17].

2. 2. Hash pointers

A hash pointer is a kind of data structure that is used a lot in the systems implemented on blockchain. It is basically a simple structure where except the fact that it contains the address where to points, but it stores this information in a hashed form. Whereas a regular pointer gives a way to retrieve the information, a hash pointer allows to get the information back and verify that the information hasn't changed. So, a hash pointer tells us where something is and what its value was [Med18].

With hash pointers can be built all kinds of data structures, from linked lists to binary trees, if they don't have cycles. It suffices to substitute the regular pointers with hash pointers. This is the data structure called "blockchain", a tree using hash pointers, as in Fig. 1.

Fig. 1. Diagram of hash pointers. (Source from



[Cou18])

It can be easily seen that we can add data at the ending leaves of the tree. If anybody messes the data earlier in the levels of the tree, it will be immediately detectable. That's why blockchains are "tamper-evident".

3. DMS-XT: system architecture and component functionalities

The DMS-XT system aims to make electronic management of student diploma (in the first and second cycle of studies) more coordinate, simpler, and to increase the quality of their content by detection of possible plagiarized content from previously stored thesis. This system is conceptualized to help especially universities in countries with a low level of informatization in higher education.

For example, in Albania the lack of an anti-plagiarism tool in some of the main universities has become a heavy load for lecturers, in their attempt to understand the originality of the submitted material. Not only this is an issue, but also the frequency of certain topics is an observed phenomenon (even though not documented), or the quality of the references to these topics. Moreover, the existence of centers that provide ready-made works is on the rise and is a problematic phenomenon, evidenced by various media chronicles. As the easy-to-find "diploma mill", in the space of social networks, one can find thesis' for an affordable prize.

The system is designed not to archive the whole material (thesis book) but will serve as a common platform between lecturers and students to better coordinate the mentorship process and quality assurance of the thesis. The student can read suggested topics of department lecturers and then decide which research topic to pick. Lecturers can add, delete or edit research topics; check submitted drafts for plagiarized content; and approve a final thesis prior to defense day. Another role in the system is the department secretary. He will assign mentorship to students and administer the process.

3. 1. System flow

In figure 2, is shown the process of a thesis registration. The steps are as follows:

(1) ***File Upload***. Secretary uploads a thesis file in .pdf format. It is supposed that even though such files represent unstructured data, they all follow the same template, as approved accordingly by inner regulatory bodies of every university. This policy will be the main drive for the system logic.

(2) ***PDF Parser***. File will be analyzed by the component which will try to find the defined fields as according to the policy. The tool is grammar-free and it will be able to detect certain field names such as: Student Name, Mentor Name, Year, Program degree, Abstract, Keywords, and References.

(3) ***Information Extractor***. The found fields from the parser will help the IE module in retrieving the correct information and storing it in a table. The module will automatically crate the information view by means of Natural Language Processing (NLP).

(4) ***Information-view Builder***. The module will process the extracted information and store it in a structured form. The module is responsible for passing the view to the browser. The secretary confirms the correct creation of the view and by selecting the name of the mentor he transfers this content to her account.

(5) ***Database***. The view is stored to a central database. This module will serve not only for the storage of the extracted information, but also during the plagiarized content check. Plagiarism will be checked based on three fields: abstract, keywords, and reference list.

(6) ***Information-view encryption***. The view will be encrypted, and only this encrypted content will be stored in the blockchain. after a student successfully defends and passes his/her thesis. The blockchain will serve as a backup system for every time a verification is requested. In this case, the view will be decrypted, compared with the view in the database, and a decision is made if it is verified or not.

For the process of verification, a similar flow is followed. We are not presenting a full picture in this paper, but we can say that the difference stands in the database logic. The requested student thesis will be retrieved from the blockchain, will be decrypted, then a search in the database index will provide the record which will be further compared with the decrypted view. A positive match means the student is verified and that he/she is the owner of the document.

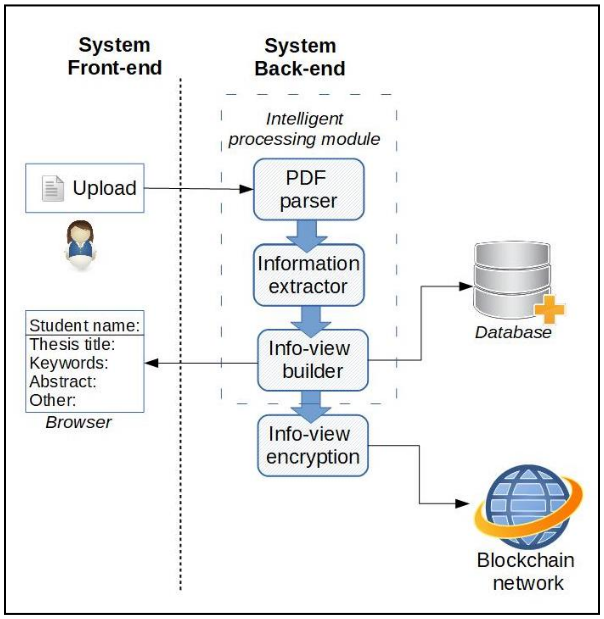


Fig. 2. Thesis registration and information-view creation, storage and encryption.

4. Conclusions and future work

Managing and verifying the documents/diploma authenticity is very difficult and due to the very high technicalities digital signatures cannot be a long-term solution. Blockchain technology offers a very stable solution in the document management. It offers stability and security as it strongly relies on cryptographic apparatus and mathematical mechanisms. There are a lot of blockchain characteristics that make it a good match for the documents management system such as: distribution, being permanent and not mutable, open source, consensus among the interested parties, etc.

The system will include a traditional approach by using database and a distributed approach by using blockchain. The main feature of the system is the plagiarism tool that will check the source of the information by extracting the information on three core parts such as: abstract, introduction and references. DMS-XT is a system that aims to manage the diploma of bachelor and master students. All the process will follow a simple flow to allow the document verification. In the blockchain will be stored only the encrypted information-view which later will be used for ownership verification. The view will be created automatically by the means of Natural Language Processing (NLP).

We provided the architecture of the system and described the main components in the registration scenario. The actual system has finished the design phase of the SDLC and smart contracts implementation and testing is the first step in the implementation phase, following the other components of Intelligent processing module as according to Fig. 2, starting with the PDF parser, Information extractor, Information-view builder and encryption module.

As generic literature suggests and with the rapid developments of the blockchain technology, we believe the integration of DMS-XT with other system is optimistic. After a successful testing of the system in the context of HEI diplomas we will further adopt the solution to other areas within areas in need for a secure and transparent document management process, such as public administration, human resource management, etc.

A final word to mention is the fact that the blockchain is not without its warts. New networks are growing, but current mechanisms have slowed their transaction speed. Blockchain, as a technology still needs to be appropriately regulated in both Europe and USA. When these regulations will be fully developed and introduced probably the costs will increase. Nevertheless, with all the skepticism it evoked, blockchain is absolutely an important achievement and a milestone in the technology development.

References

[Bak95] Bakhtiari, Shahram, Reihaneh Safavi-Naini, and Josef Pieprzyk. Cryptographic hash functions: A survey. Vol. 4. Technical Report 95-09, Department of Computer Science, University of Wollongong, 1995.

[Bbc18] https://www.bbc.co.uk/news/uk-42579634, accessed September 2018.

[Cou18] “Bitcoin and Cryptocurrency Technologies”, Princeton University online course. Accessed March, 2018. (https://www.coursera.org/learn/cryptocurrency/home/welcome)

[Csm18] Case Study Malta|Learning Machine. https://www.learningmachine.com/casestudies-malta.

[Gat17] Gates, Mark. Blockchain: Ultimate guide to understanding blockchain, bitcoin, cryptocurrencies, smart contracts and the future of money. CreateSpace Independent Publishing Platform, 2017.

[Ipf18] IPFS: https://ipfs.io/, accessed August 2018.

[Mar18] Martiri, Edlira; Muca, Gentjana,, “A blockchain eco-system analysis for the Western Balkans countries and an economic perspective”,7thInternational Conference on Computer Science and Communication Engineering, UBT, Kosovo, October 2018.

[Mat18] Mathworld website: accessed June 2018.

http://mathworld.wolfram.com/Collision-FreeHashFunction.html.

[Med18] Medium, “Hash pointers and data structures”, accessed online, June 2018, https://medium.com/@zhaohuabing/hash-pointers-and-data-structures-f85d5fe91659

[Men96] Menezes, Alfred J.; van Oorschot, Paul C.; Vanstone, Scott A (1996). Handbook of Applied Cryptography. CRC Press. ISBN 0849385237.

[Mit18] Certificates, Reputation, and the Blockchain – MIT MEDIA LAB. http://certificates.media.mit.edu/

[Roo17] Rooksby, John, and K. Dimitrov. "Trustless Education? A Blockchain System for University Grades." New Value Transactions Understanding and Designing for Distributed Autonomous Organisations Workshop at DIS 2017.

[Sha16] Sharples, Mike et al. 2016. Innovating pedagogy 2016: Open University innovation report.

[Sto18] StorJ: https://storj.io/, accessed August 2018.

[Swa15] Swan, Melanie. Blockchain: Blueprint for a new economy. " O'Reilly Media, Inc.", 2015.

[Tha17] Thakur, Mukesh. "Authentication, Authorization and Accounting with Ethereum Blockchain.", Master thesis, University of Helsinky (2017).

[Trr18] Meet TrueRec by SAP: Trusted Digital Credentials Powered by Blockchain. Retrieved March 22, 2018 from https://news.sap.com/meet-truerec-by-sap-trusteddigital-credentials-powered-by-blockchain/

[Wan18] Wang, Maoning, Meijiao Duan, and Jianming Zhu. "Research on the Security Criteria of Hash Functions in the Blockchain." Proceedings of the 2nd ACM Workshop on Blockchains, Cryptocurrencies, and Contracts. ACM, 2018.