Timestamping Metadata Using Blockchain: A Practical Approach

Tassos Kolydas[1][0000-0001-9907-658X]

[1] National and Kapodistrian University of Athens, Athens, Greece
kolydart@music.uoa.gr

Abstract. Long-term preservation of digital information requires confidence in the credibility and ability of digital archives and systems to consistently provide accessible and usable content. Ensuring that the provided information has remained unchanged over time is a particular challenge. Trusted timestamping is an effective method that allows anyone to prove without any doubt that specific content existed at a particular date and time. A practical approach for trusted timestamping using the Ethereum blockchain is presented here. A complete metadata record is stored as transaction input data along with a document hash digest. The approach is uncomplicated, human and machine readable, self-explanatory, and modular. It supports metadata preservation and copyright protection of digital documents applying verification without disclosure. The approach aims at extending current digital archives and systems using existing, well-tested technology.

Keywords: Trusted Timestamping, Ethereum Blockchain, Digital Preservation, Copyright Protection.

1 Introduction

Long-term preservation of digital information requires confidence in the credibility and ability of digital archives and systems to consistently provide accessible and usable content. Information generated today must survive long-term changes in storage media, devices, and data formats [1]. Ensuring that the provided information has remained unchanged over time is a particular challenge. Trusted timestamping is an effective method that allows anyone to prove without any doubt that specific digital content existed at a particular date and time [2]. Blockchain technology has emerged as a means of storing information on decentralized networks, secured from tampering and revision [3]. Ethereum is an open-source, globally decentralized computing infrastructure that executes programs called “smart contracts” [4]. It uses a blockchain to synchronize and

Tassos Kolydas, “Timestamping Metadata Using Blockchain: A Practical Approach”, Emmanouel Garoufallou, Francesca Fallucchi & Ernesto William De Luca (ed.), *Metadata and Semantic Research. MTSR 2019. Communications in Computer and Information Science,* 1057, Springer, Cham, 2019.

This is this is an author-created version of the preprint. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-36599-8_42.

store the system's state changes, along with a cryptocurrency called "ether" to meter and constrain execution resource costs [4].

A practical approach for trusted timestamping using the Ethereum blockchain is presented here. A complete metadata record is stored in transaction input data, along with a document hash digest. The approach is uncomplicated, human readable as well as machine readable, self-explanatory, and modular. It supports metadata preservation and copyright protection of digital documents without revealing the content of the documents. The approach aims at extending current archives and systems using existing, well-tested technology.

2 Related Work and Background

Timestamping is a technique used to prove the existence of certain digital data prior to a specific point in time. When one uses a certain timestamping service, he should confirm in advance that its security level sufficiently meets his security requirements [2]. However, before the advent of blockchain technology, timestamping schemes were generally so complicated that it was not easy to evaluate their security levels accurately [2]. More recently, authenticity of information has become a core issue due to the spread of false news. Research on the subject has showed that falsehoods diffuse significantly farther, faster, deeper, and more broadly than the truth in all categories of information [5].

Blockchain technology can be used to address issues associated with information integrity in the present and near term, assuming proper security architecture and infrastructure management controls. It does not, however, guarantee reliability of information in the first place, and would have several limitations as a long-term solution for maintaining trustworthy digital records [6]. Recently, the European Union launched the International Association of Trusted Blockchain Applications, aiming to bring blockchain and distributed ledger technology into the mainstream [7]. Blockchain technology has the potential to change the practices and systems used for archival functions, both for the storage of digital resources and of the metadata describing them [3].

2.1 Overview

One of the most well-known services is *OriginStamp*, which is a trusted timestamping service that enables anyone to prove ownership of information on a specific date and time. It is provided both as a free service [8] saving data on blockchain once a day or as a

paid service for instant entry [9]. It utilizes bitcoin [10], "because it has achieved the largest market capitalization and has attracted the highest number of participating computing nodes" [11], while more blockchains were added recently. A unique SHA-256 hash is generated from all content and is contained in a bitcoin transaction [12]. The free service keeps the transaction costs in the blockchain to a minimum by collecting all hashes received over a 24-hour period and computing a single aggregate SHA-256 hash from the list of hashes [13]. The service exclusively stores which hashes were included in which transaction. This information allows verification of any hash using the blockchain [11]. This means that trust in a third party and/or firm cryptographic knowledge is required.

Proof of Existence is a similar solution providing only paid service to store the document proof on the bitcoin blockchain [14]. The document is certified via embedding its SHA-256 digest in the bitcoin blockchain. This is done by generating a special bitcoin transaction that encodes/contains the hash via an OP_RETURN script. Thus, no other information is provided except the document digest. The service costs much more than *OriginStamp* and has no free plan. Thus, the same limitations apply.

García-Barriocanal et al. proposed an approach in which a blockchain combined with other related technologies can be arranged in a particular way to obtain a decentralized solution for metadata supporting key functions [3]. The proof-of-concept implementation of a decentralized metadata system considers the different functions of metadata as points of departure, thus building the solution based mainly on blockchain.

Lemieux has presented a synthesis of original research documenting several cases of the application of blockchain technology, considering the different types of solutions in relation to implications for record keeping and long-term preservation of authentic records [15].

The proposed solution in this paper belongs in the first type of system – according to Lemieux – the "mirror" type, where the blockchain serves as a repository of "digital fingerprints" of the records in an originating system [15, p. 2273]. Including a documents digest, the original file content has essentially been encoded into the blockchain, and the blockchain can serve as a document registry [16, p. 37].

3 Timestamping (meta)Data on Blockchain

3.1 Description

The method consists of storing a metadata record along with any document hash digest on the Ethereum blockchain. The record, stored as a JSON object, is inserted in the “input data” field of a zero-value transaction. Each transaction contains a single record.

Verifying the metadata record is as simple as retrieving the transaction content and decoding the input data from hexadecimal to UTF-8. Some blockchain explorers, like *EtherScan*, provide the function on their web applications (see Fig. 1). Validating the document requires an extra step of hashing the original bitstream and comparing it with the included digest.

The process from the digital archive’s point of view includes the secure storage of each digital document *for the specific transaction* in which it is included as well as the identifier of the transaction, known as the “transaction hash”. Having these two, anyone can verify the metadata record or the document’s content for any changes against the timestamped content. Therefore, the responsibility for metadata creation and storage remains in the archive’s system.

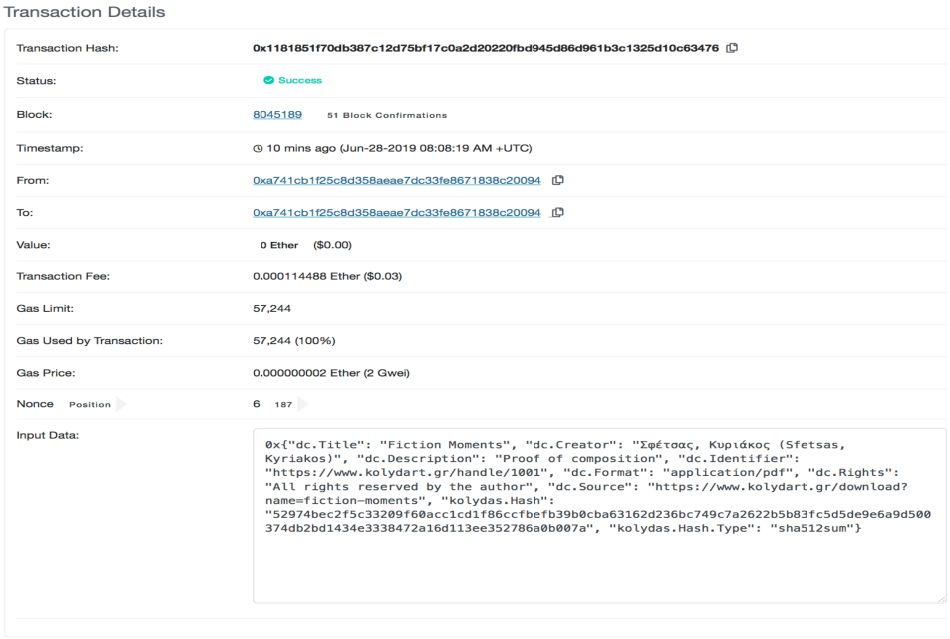

Fig. 1. Transaction viewed from EtherScan.io blockchain explorer [17].

The simplicity of the approach adheres to the design and philosophy of the Ethereum blockchain as described in the white paper: “An average programmer should ideally be able to follow and implement the entire specification so as to fully realize the unprecedented democratizing potential that cryptocurrency brings and further the vision of Ethereum as a protocol that is open to all” [18]. Also, the transaction input data field accepts any data. Transactions contain an optional data field that has no function by default [18]. The term “transaction” is used in Ethereum to refer to the signed data package that stores a message to be sent from an externally owned account.

3.2 Implementation

Multiple packages are available for interacting with the Ethereum blockchain. The Python package web3.py (https://github.com/ethereum/web3.py) was selected for this implementation. The code is available at:
https://www.kolydart.gr/download?name=mtsr-2019-kolydas-code

Creating a New Transaction:

```
from web3 import Web3, HTTPProvider
w3 = Web3(HTTPProvider(provider)) # connect to blockchain network node
transaction_content = dict(
   nonce = w3.eth.getTransactionCount(address),
   gasPrice = w3.eth.gasPrice,
     gas = w3.eth.estimateGas({'to': address, 'from': address, 'value': 0, 'data':
   record_hex}),
   to = address,
   value = 0,
   data=record_hex,
) # prepare transaction content
signed_txn = w3.eth.account.signTransaction(transaction_content, pk) # sign transaction
transaction_hash = w3.eth.sendRawTransaction(signed_txn.rawTransaction) # send
   transaction
print(transaction_hash.hex()) # retrieve transaction hash
```

Retrieving a Stored Record’s Element:

```
from web3 import Web3, HTTPProvider
import json
w3 = Web3(HTTPProvider(provider)) # connect to blockchain network node
transaction = w3.eth.getTransaction(transaction_hash) # get transaction
inputData = transaction.input # get record from transaction input data
json_data = bytearray.fromhex(inputData[2:]).decode() # decode hex to utf-8
print(json.loads(json_data)['dc.Title']) # retrieve record element
print(json.loads(json_data)['kolydas.Hash']) # retrieve record element
```

Transaction.
The proposed solution was used to create the transaction hash 0x1181851f70db387c12d75bf17c0a2d20220fbd945d86d961b3c1325d10c63476. For improved security, the SHA-512 algorithm was used for the document hash value. The metadata record is available from:
https://www.kolydart.gr/handle/1001.

4 Evaluation

The suggested approach contains most of the well-known benefits of blockchain technology. Stored records are immutable, every transaction is timestamped, and data are interconnected in such a way that upon tampering, the structure becomes invalid.

Centralization is a process where the authority to make decisions lies in the hands of only a few. In other words, centralization is the consistent and systematic way of entrusting authority to people who are in the center of the organization [19]. Using blockchain in a decentralized network, everyone holds the same information. Interacting with anyone in the network is possible without the need for any third party. Furthermore, the proposed solution interacts directly with the blockchain and is platform-independent, without any third-party involvement. Actually, entrusting the timestamping process to a third party to store it for you on a decentralized blockchain is a kind of paradox.

Access to the data is public, and information is freely available. Each entity is responsible for the information it stores on the blockchain. At the same time, digital document content remains private, and verification of the authenticity of a document takes place without disclosing its content.

Simplicity is a key feature of the approach. The principle of least effort (PLE) is having a growing influence on library and information science research, becoming increasingly important in various subfields [20]. It postulates that people, and even well-designed machines, will naturally choose the path of least resistance or “effort.” Thus, a lightweight approach implemented with a few lines of code might in many cases be preferred.

Portability and interoperability are key elements in the life cycle of metadata. Using JSON formatted data, and being able to manually verify authenticity using a single hash function is an important advantage. Also, it is possible to use different hash algorithms, depending on the circumstances, as long as it is documented in the metadata.

Document size does not affect the proposed timestamping procedure because the hash digest’s string length is fixed. The size of the metadata record is the same, whether

describing a small or large document. File size is a factor that could be examined during hashing when processing resources are limited, so the above-mentioned flexibility to use different hash algorithms is a major advantage over other solutions. When speed is a priority, a less resource-intensive hash function could be used to quickly compute the hash output of a large document. Alternatively, if security is a concern, a state-of-the-art collision-resistant hash function could be used – like the proposed SHA-512 – in order to make it infeasible to produce the same hash value from two different documents.

Privacy is another important aspect of the proposed method. While we all reap the benefits of a data-driven society, there is a growing public concern about user privacy [21]. Centralized organizations – both public and private – amass large quantities of personal and sensitive information [21]. Using a third-party platform that requires some kind of user registration poses a security risk when data breach incidents occur or users' data are misused. The proposed method ensures that users own and control their personal data. No interaction with third parties is involved, and storing data on the blockchain requires only the cryptographic key of a digital wallet and some ether.

The cost of the solution is variable, depending on the STARTGAS and GASPRICE values used in the transaction and also the size of the input data field [18, "Messages and Transactions"]. The cost is competitive compared to other blockchain solutions and extremely low compared to non-blockchain solutions. Since each transaction contains a single metadata record stored in real time, the solution provides a service comparable to the above-mentioned platforms *OriginStamp* and *Proof of Existence*. Using the code mentioned previously, the cost of storing the record on the Ethereum blockchain was 0.000114488 ether (about $0.025); validating the document digest is free. *Proof of Existence* requires a fixed payment of 0.00025 btc (about $2.4) per transaction [14]; validation is also free. *OriginStamp* has multiple price plans; a free plan is sufficient for uses such as creating eight timestamps and downloading the corresponding proofs each month (50 credits). Paid plans vary from $39 for 1,000 credits per month to $249 for 100,000 credits per month [9]. The Hellenic Copyright Organization has developed a non-blockchain online timestamping service that charges based on the size of the uploaded file. Prices range from €10 for timestamping a 3MB file to €70 for a 2GB file; the verification process costs one hundred euros (€100) [22].

The proposed solution provides added value to digital libraries and archives, which can offer copyright protection as a service to their content providers. This creates an additional incentive for digital content creators to trust their creations to digital archives since they will automatically have their copyright protected.

5 Implications

Since it is so easy to apply a trusted timestamp to your data, you *have to immediately timestamp your original content*. Otherwise, you risk having your content claimed by others who take advantage of their knowledge of the timestamping procedure and benefits. Protecting original content implies that timestamping should be performed even on works in progress.

Timestamping works in progress and re-timestamping updated info add complexity to the design of the archives because multiple transaction hashes have to be stored in order to follow intermediate states of information.

Last but not least, privacy is an important factor that must be considered. Private key management is essential in a system that relies upon cryptography, such as blockchain. This includes the generation, exchange, storage, use, and replacement of keys, which is difficult to achieve in practice [6, p. 129]. Effective key management, including system policy, user training, and organizational and departmental interactions, is a critical factor for the success of the solution.

6 Conclusion

A practical approach for trusted timestamping metadata records and digital documents has been demonstrated. Most digital library packages lack such an implementation for the moment. But interest in blockchain solutions for metadata is rapidly gaining traction. The effect that the web “fabric” had on the world during the 1990s seems comparable to the impact of the information “locking” that blockchain provides today.

Future research could explore timestamping large collections – even complete archives – in a single transaction using a *Merkle tree*, also known as a *binary hash tree*, which is a data structure used for efficiently summarizing and verifying the integrity of large sets of data [23]. This structure could be used to summarize all documents in a collection, producing an overall digital fingerprint of the entire set while providing an efficient process for verifying whether a document was included in a collection on a specific date and time.

References

1. Lorie, R. A.: Long term preservation of digital information. Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries. pp. 346–352, (2001). https://dl.acm.org/citation.cfm?id=379726. Last accessed 2019/8/29.

2. Une, M.: The Security Evaluation of Time Stamping Schemes: The Present Situation and Studies. In: IMES Discussion Papers Series 2001-E-18 (2001). https://citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.23.7486&type=ab. Last accessed 2019/6/30.
3. García-Barriocanal, E., Sánchez-Alonso, S., Sicilia, M.-A.: Deploying Metadata on Blockchain Technologies. In: Research Conference on Metadata and Semantics Research – MTSR, pp. 38–49. Springer, Heidelberg (2017/11).
4. Antonopoulos, A. M., Wood, G.: Mastering ethereum: building smart contracts and dapps. O'Reilly Media (2018).
5. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science 359(6380) 1146–1151 (2018).
6. Lemieux, V. L.: Trusting records: Is Blockchain technology the answer? Records Management Journal 26(2), 110–139 (2016).
7. Zmudzinski, A.: European Union Launches International Association of Trusted Blockchain Applications. CoinTelegraph (2019/04/03). https://cointelegraph.com/news/european-union-launches-international-association-of-trusted-blockchain-applications. Last accessed 2019/6/30.
8. OriginStamp. https://originstamp.org. Last accessed 2019/6/30.
9. OriginStamp. https://originstamp.com. Last accessed 2019/6/30.
10. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System. The Cryptography Mailing List. https://bitcoin.org/bitcoin.pdf. Last accessed 2019/6/30 (2008/10/31).
11. Gipp, B., Meuschke, N., Gernandt, A.: Decentralized Trusted Timestamping using the Crypto Currency Bitcoin. In: Proceedings of the iConference 2015 (υπό έκδοση), Newport Beach, CA, USA, Mar. 24–27. http://www.gipp.com/wp-content/papercite-data/pdf/gipp15a.pdf. Last accessed 2019/6/30 (2015).
12. Gipp, B., Meuschke, N., Beel, J. & Breitinger, C.: Using the Blockchain of Cryptocurrencies for Timestamping Digital Cultural Heritage. Bulletin of IEEE Technical Committee on Digital Libraries (TCDL) 13(1) https://www.gipp.com/wp-content/papercite-data/pdf/gipp2017a.pdf, last accessed 2019/6/30 (2017).
13. Gipp, B., Breitinger, C., Meuschke, N., Beel, J.: CryptSubmit: Introducing Securely Timestamped Manuscript Submission and Peer Review Feedback using the Blockchain. In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) (2017).
14. Proof of Existence. https://www.proofofexistence.com/. Last accessed 2019/6/30.
15. Lemieux, V. L.: A Typology of Blockchain Recordkeeping Solutions and Some Reflections on their Implications for the Future of Archival Preservation. Big Data. In: 2017 IEEE International Conference – BIGDATA, (Arles 2017). https://ieeexplore.ieee.org/abstract/document/8258180. Last accessed 2019/6/30.
16. Swan, M.: Blockchain: Blueprint for a new economy. O'Reilly Media (2015).
17. etherscan.io: Transaction Details. https://etherscan.io/tx/0x1181851f70db387c12d75bf17c0a2d20220fbd945d86d961b3c1325d10c63476. Last accessed 2019/6/30.
18. A Next-Generation Smart Contract and Decentralized Application Platform. Ethereum White Paper. https://github.com/ethereum/wiki/wiki/White-Paper#ethereum. Last accessed 2019/6/30.
19. Kaushik, A., Choudhary, A., Ektare, C., Thomas, D., Akram, S.: Blockchain – Literature Survey. In: 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (2017). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8256979. Last accessed 2019/6/30.

20. Chang, Y.-W.: Influence of human behavior and the principle of least effort on library and information science research. Information Processing & Management 52(4) 658–669 (2016). https://doi.org/10.1016/j.ipm.2015.12.011. Last accessed 2019/6/30.
21. Zyskind, G., Nathan, O., Pentland, A. S.: Decentralizing Privacy: Using Blockchain to Protect Personal Data. 2015 IEEE Security and Privacy Workshops. pp. 180–184, IEEE (2015). https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7163223. Last accessed 2019/8/30.
22. Hellenic Copyright Organization: Electronic Timestamping Service. https://www.timestamp.gr/en/information. Last accessed 2019/6/30.
23. Antonopoulos, A. M.: Mastering Bitcoin: Programming the Open Blockchain. O'Reilly Media (2017).