Securing the Integrity of Time Series Data in Open Science Projects using Blockchain-based Trusted Timestamping

Patrick Wortner[1], Moritz Schubotz[1,2], Corinna Breitinger[3], Stephan Leible[1], Bela Gipp[1]
[1] University of Wuppertal, Germany {lastname@uni-wuppertal.de}
[2] FIZ Karlsruhe, Leibnitz Institute for Information Infrastructure, Germany
[3] University of Konstanz, Germany, corinna.breitinger@uni-konstanz.de

ABSTRACT

The open science movement has become a synonym for modern, digital, and inclusive science. At the same time, open science introduces new challenges for digital libraries, as well as the long-term preservation and quality assurance of open science datasets. According to open science principles, not only researchers but also citizens should be able to contribute data, e.g., so-called 'citizen science projects.' For such democratized projects, securing the *integrity* and *longevity* of research data is a particular concern. We propose an approach capable of securing the integrity of time series data directly as it is generated. The data is automatically stored in a decentralized and tamper-proof manner while using blockchain technology to prevent any subsequent modification. Our prototype demonstrates how time series data recorded by sensors, e.g., temperature, current, and vibration sensors, can be transparently and immutably stored. By demonstrating an inexpensive modular hardware prototype in combination with open source software, we show that the entry barrier is low for implementing open science projects capable of securing data integrity and offering decentralized long-term data storage. Our approach, in turn, can increase the legitimacy of open science datasets and citizen science projects in particular.

1 Introduction

Traditionally, the workflow of researchers and scientists consisted of the following steps: (1) researching literature, (2) formulating a hypothesis, (3) performing experiments and recording measurements, (4) evaluating data, and (5) publishing the results. If the fifth step, however, is repeatedly unsuccessful, scientists cannot succeed in today's research environment. In part due to this pressure to publish, it is not uncommon for scientists to attempt to manipulate steps 3 and 4 to be able to publish [6, 7]. In contrast to the traditional academic research cycle and publishing process, today's *open-science movement* encourages and facilitates the publishing of raw research measurements early in the research cycle and even encourages the publishing of negative results. These developments, in turn, have introduced new challenges to storing and verifying research data throughout the research cycle.

Performing research according to open science principles [12] provides at least two significant advantages. First, data fabrication becomes more complicated if raw data and intermediate results are published. Second, additional insights can be obtained and published by other researchers without the time-consuming experimentation step. The open science movement also blurs the line between scientists and citizens by enabling inclusion of interested individuals, for example, in so-called 'citizen science projects.'

Currently, there is no agreement among scientists on a standardized procedure for reviewing open science datasets in a way that is analogous to today's literature-assessment process. Methods for researching and reviewing scientific literature using content-based and bibliometric measures are well established for traditional publications. While bibliometric measures could also be applied to datasets, a content-based assessment based on the raw data seems hardly feasible. Additionally, the quality of measurement data is influenced by factors, such as (1) well-defined error estimates, (2) accurate meta-data, (3) high redundancy, (4) adequate sampling rate, (5) adherence to all known physical laws, etc. While readers and scientists must manually verify such factors, we believe that scientists could significantly benefit from an automated method to guarantee the integrity and longevity of time series research data immediately as it is generated. Thus, we propose securing time-series research data using a technical solution to protect data against any subsequent changes or manipulation. To achieve this, we make use of decentralized trusted timestamping, which relies on cryptographic hashing and the tamper-proof characteristics of blockchain technology [9].

To demonstrate our proposed solution, we present a simple and inexpensive hardware prototype that can be used by citizens to record measurements for certain physical properties. The corresponding open source software makes use of blockchain technology and decentralized data storage to ensure the immutability of the sensors' time series data. With this contribution, we hope to support the viability of citizen science projects and today's open science movement by making data and entire datasets more trustworthy.

2 Using blockchain for securing sensor measurement data

The blockchain underlying cryptocurrencies, e.g., Bitcoin [11], are offering unique characteristics, such as decentralization, immutability, and trusted timestamping [14, 17]. These properties are making blockchain technology valuable in developing novel applications [4]. Several projects have been proposed to support researchers, including managing academic reputation [15], protecting intellectual property in academic manuscripts submitted for peer review [8], or tracking individual contributions in a collaborative research project [13]. From a technical point of view, there is much literature on blockchain technology and its strengths and weaknesses in different stress tests and use cases [14, 16, 17]. A blockchain can be viewed as a decentralized database without a central authority to manage the data it stores [1]. Data stored on a blockchain is immutable and permanent. We use these characteristics to ensure that each measurement value recorded by a sensor is made tamper-proof.

In the case of citizen science projects, manipulating many sensors in a decentralized network would require significant effort and would be difficult to achieve without being detected. Typically, it is easier to manipulate or prune the data after it was measured or aggregated, but our approach is capable of preventing such retrospective tampering.

Once a hash has been included in the form of a transaction on a blockchain, one can verify that the data associated with the hash (for example, sensor measurement values) were not manipulated after they were collected [5]. We do not propose to store all raw measurement data directly on a blockchain. Instead, we only store the hash value of digital measurements or sets of measurement data. This way is more efficient in terms of performance, cost, and scalability, yet it allows proving that the data existed in a specific format at a certain time, thus increasing data integrity and transparency.

3 Prototype for open science projects

Capturing time series data from sensors and verifiably securing all data with a trusted timestamp requires building an interface between the physical world and a blockchain. To demonstrate our idea, we implemented an inexpensive and easy to use Raspberry Pi prototype that captures vibration, electric current and temperature measurements. The modular design of the prototype allows for easy customizations. For the prototype, we used overall a Raspberry Pi, a general-purpose input/output board, and a set of three sensors. To test the sensors of the model, we designed a testing environment: a rotor with a battery power supply. The rotor consumes power and generates vibrations, which produce signals that can be measured with the sensors we use. There are plenty more use cases like measuring and timestamping earthquake data to publish them in digital libraries or web archives. Further, a Proof of Location method [3] could be added additionally to verify the position where sensors measure the data.

We also implemented open source software to capture the incoming data stream from the sensors, partition the stream into data-chunks, timestamp each chunk, and finally store the chunks. The software and installation manual, as well as links to the first datasets, are available on our GitHub repository[1]. To install the software, the Raspberry Pi needs to be connected to a computer. The software reads the data-stream from the sensors and appends it to an internal buffer. After a user-defined time, or when the data volume size exceeds the chunk size of about 256 kB, a new chunk is created. A hash of every data chunk is computed and uploaded to the trusted timestamping service OriginStamp[2] [10]. From the second chunk onward, a reference to the previous chunk is included in the hash.

However, the hash is meaningless without the associated data. Many services exist to upload data to a central server but services can fail, and servers can cease to exist, or their content censored and tampered. We, therefore, use the InterPlanetary File System (IPFS) [2] to be independent of a central authority for storage and to ensure data verifiability and longevity. This peer-to-peer network is organized in blocks, and the block address is the hash of the file content, which we already used to generate the decentralized trusted timestamp. Therefore, we installed IPFS on the Raspberry Pi to upload the data chunk-by-chunk. For redundancy, we additionally plan to copy the time series measurement data from IPFS to the long-time archival platform Zenodo[3].

4 Conclusion

Measurement values and data streams from sensors are not immune to manipulation or retrospective selective pruning. The ability to securely prove that research data was not manipulated or omitted is especially important for both open science and citizen science projects. In this paper, we proposed a method for independently and securely verifying the time of creation and integrity of sensor data. By storing the data in a decentralized manner using IPFS and by relying on a blockchain-backed solution for storing tamper-proof and decentralized trusted timestamps associated with discrete chunks of measurement values, we showed how sensor data can be made securely verifiable. Our prototype demonstrates how our proposed solution can be easily implemented and used with hardware

[1] https://github.com/FellowsFreiesWissen/Blockchain_Pi

[2] www.originstamp.org

[3] www. zenodo.org

sensors. We argue that enabling any researcher or interested citizen to trace the integrity of measurement values and their time of origin verifiably can significantly strengthen the open science movement and can increase the trustworthiness of citizen science projects.

REFERENCES

[1] Beck, R., Czepluch, J.S., Lollike, N., and Malone, S. Blockchain-the Gateway to Trust-Free Cryptographic Transactions. *ECIS*, (2016).

[2] Benet, J. IPFS - Content Addressed, Versioned, P2P File System. *arXiv:1407.3561*, (2014)

[3] Brambilla, G., Amoretti, M., and Zanichelli, F. Using Blockchain for Peer-to-Peer Proof-of-Location. *arXiv:1607.00174*, (2016).

[4] Casino, F., Dasaklis, T.K., and Patsakis, C. A systematic literature review of blockchain-based applications: current status, classification and open issues. *Telematics and Informatics*, (2018).

[5] Catalini, C. and Gans, J.S. Some Simple Economics of the Blockchain. *MIT Sloan Research Paper No. 5191-16*, (2016).

[6] Fanelli, D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one 4(5): e5738*, (2009).

[7] Fanelli, D. Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PloS one 5(4): e10271*, (2010).

[8] Gipp, B., Breitinger, C., Meuschke, N., Beel, J., and Breitinger, C. CryptSubmit: Introducing Securely Timestamped Manuscript Submission and Peer Review Feedback using the Blockchain. *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, (2017).

[9] Gipp, B., Meuschke, N., and Gernandt, A. Decentralized Trusted Timestamping using the Crypto Currency Bitcoin. *Proceedings of the iConference 2015*, (2015).

[10] Hepp, T., Schoenhals, A., Gondek, C., and Gipp, B. OriginStamp: A blockchain-backed system for decentralized trusted timestamping. *Information Technology 60, 5–6* (2018), 273–281.

[11] Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system. (2008).

[12] Nosek, B.A., Alter, G., Banks, G.C., et al. Promoting an open research culture. *Science 348, 6242* (2015), *1422–1425*.

[13] Schubotz, M., Breitinger, C., Hepp, T., and Gipp, B. Repurposing Open Source Tools for Open Science: a Practical Guide. 2018. *https://doi.org/10.5281/zenodo.2453415*.

[14] Seebacher, S. and Schüritz, R. Blockchain technology as an enabler of service systems: A structured literature review. *International Conference on Exploring Services Science*, (2017), *12–23*.

[15] Sharples, M. and Domingue, J. The blockchain and kudos: A distributed system for educational record, reputation and reward. *European Conference on Technology Enhanced Learning*, (2016), *490–496*.

[16] Yli-Huumo, J., Ko, D., Choi, S., Park, S., and Smolander, K. Where is current research on blockchain technology?—a systematic review. *PloS one 11 (10): e0163477*, (2016).

[17] Zheng, Z., Xie, S., Dai, H., Chen, X., and Wang, H. An overview of blockchain technology: Architecture, consensus, and future trends. *Big Data (BigData Congress), 2017 IEEE International Congress on*, (2017), *557–564*.