preprint of publication:

P. Mytis-Gkometh, G. Drosatos, P. S. Efraimidis and E. Kaldoudi, Notarization of Knowledge Retrieval from Biomedical Repositories using Blockchain Technology, ICBHI 2017: International Conference on Biomedical and Health Informatics, Thessaloniki, Greece, 18-21 November 2017

Notarization of Knowledge Retrieval from Biomedical Repositories using Blockchain Technology

P. Mytis-Gkometh[1], G. Drosatos[2], P. S. Efraimidis[1] and E. Kaldoudi[2]

[1] Department of Electrical and Computer Engineering, Democritus University of Thrace, Kimmeria, Xanthi 67100, Greece
[2] School of Medicine, Democritus University of Thrace, Dragana, Alexandroupoli 68100, Greece

***Abstract*— Biomedical research and clinical decision depend increasingly on a number of authoritative databases, mostly public and continually enriched via peer scientific contributions. Given the dynamic nature of data and their usage in the sensitive domain of biomedical science, it is important to ensure retrieved data integrity and non-repudiation, that is, ensure that retrieved data cannot be modified after retrieval and that the database cannot validly deny that the particular data has been provided as a result of a specific query. In this paper, we propose the use of blockchain technology in combination with digital signatures to create smart digital contracts to seal the query and the respective results each time a third-party requests evidence from a reference biomedical database. The feasibility of the proposed approach is demonstrated using a real blockchain infrastructure and a publicly available medical risk factor reference repository.**

***Keywords*— Biomedical repositories, cryptographic techniques, blockchain, integrity, non-repudiation.**

I. Introduction

Biomedical research and clinical practice relies increasingly on authoritative data gathered and curated in reference biomedical databases. Examples include: clinical databases (registries or academic clinical databases) that hold clinical data on patient cohorts [1]; biomedical databases [2] with current data on pharmaceuticals [3], metabolomics [4], inheritance data and other omics (for example, the rich collection available from the European Bioinformatics Institute at http://www.ebi.ac.uk/services); publication repositories and other medical evidence repositories [5], either general purpose (the most prominent example being PubMed service by the National Library of Medicine, USA) or high evidence quality, such as Cochrane Library reports.

Biomedical references databases are continually updated to include new data sets (e.g. PubMed included ~1M new records in 2016), and are often validated and, if necessary, updated to correct existing data. At any given point in time, these data are heavily accessed by humans (clinicians, patients and researchers alike) and software (via appropriate application programming interfaces) to establish current evidence and inform clinical acts and biomedical research. As such, it is important to ensure that data cannot be manipulated retrospectively and that data ‘consumers’ can have a proof of what data were retrieved from the database at a given point in time as a result of a specific query.

A reliable knowledge retrieval service has to fulfill at least the following two important requirements; integrity and non-repudiation. Integrity, means that the query and the retrieved data cannot be modified (either by accident or deliberately), once the retrieval operation completes. Non-repudiation, in this context means that given any past retrieval operation, the knowledge retrieval service cannot validly deny that the exact data have been provided by the service as a response to the given query at the specific time. An interesting solution that satisfies the above requirements can be found in the emerging field of blockchain infrastructures. Blockchains inherently ensure the integrity of each recorded transaction. Moreover, non-repudiation can also be accomplished if blockchains are for example combined with digital signatures.

In this paper, we propose the use of blockchain technology to create smart digital contracts to seal the query and the respective results each time a third-party requests evidence from a reference biomedical database. The proposed approach is demonstrated on the powerful Ethereum blockchain platform [6] with a retrieval service for the publicly available CARRE risk factor reference repository [7]. The repository has been developed in the context of the European Union funded FP7-ICT project CARRE (Grant no. 611140), which researched and developed novel personalised decision support services for managing comorbidities associated with cardiorenal disease.

II. Background

Blockchain is a distributed, incorruptible transaction management technology without one single trusted party. Each new transaction is broadcasted to a distributed network of nodes; once all nodes agree the transaction is valid, the transaction is added to a block. Every block contains a timestamp and the hash (cryptographic seal) of the previous block and the transaction data, thus creating an immutable, append-only chain. Copies of the entire blockchain are maintained by each participating node.

The first blockchain was proposed for and implemented in Bitcoin [8], a distributed infrastructure where users can make financial transactions without the need of a regulator (e.g. a bank). Nowadays, other blockchain infrastructures are emerging, for example the Ethereum [6], where everyone can participate in the blockchain generation, and the Hyperledger Fabric [9], where only approved parties can post to the blockchain. In permissionless blockchains like Bitcoin and Ethereum, all transactions are public, however, no direct links to identities exist. When applied to financial transactions, this privacy preserving features can be enhanced even further [10]. However, in applications that require non-repudiation, identity should be irrevocably maintained; this can be ensured by the appropriate use of public key infrastructure solutions [11],[12].

A recent systematic review on current state, limitations and open research on blockchain technology [13] discusses a number of blockchain applications that extend from crypto-currency to Internet of things, smart contracts, smart property, digital content distribution, Botnet, and P2P broadcast protocols. Currently, there is considerable optimism that blockchain technology will revolutionize the healthcare industry [14]. Indeed, blockchain technology has been proposed as a solution for privacy-preserving control and sharing of patient personal healthcare data [15],[16],[17] and for record management in clinical trials to ensure that data is fully published and not tampered with [18],[19].

III. Query Notary Service

In this paper, we propose a lightweight wrapper for conventional databases that uses blockchain technology to offer database query notary services to data consumers (humans and programs alike). The proposed notary service administers contracts that seal a query placed to a database and the returned results. The service offers irrevocable proof of data retrieved by a specific query placed by a specific consumer, thus establishing query transaction integrity and non-repudiation. In this way, the proposed system assures that the consumer is protected against a service that may accidently or intentionally try to repudiate or alter a past query transaction.

The overall architecture is presented in Fig. 1. The main component is the blockchain contract service that acts as a mediator between conventional biomedical databases and data consumers. The structure of the biomedical knowledge could be any database model (i.e. SQL or NoSQL databases), or even semantic repositories (i.e. RDF stores).

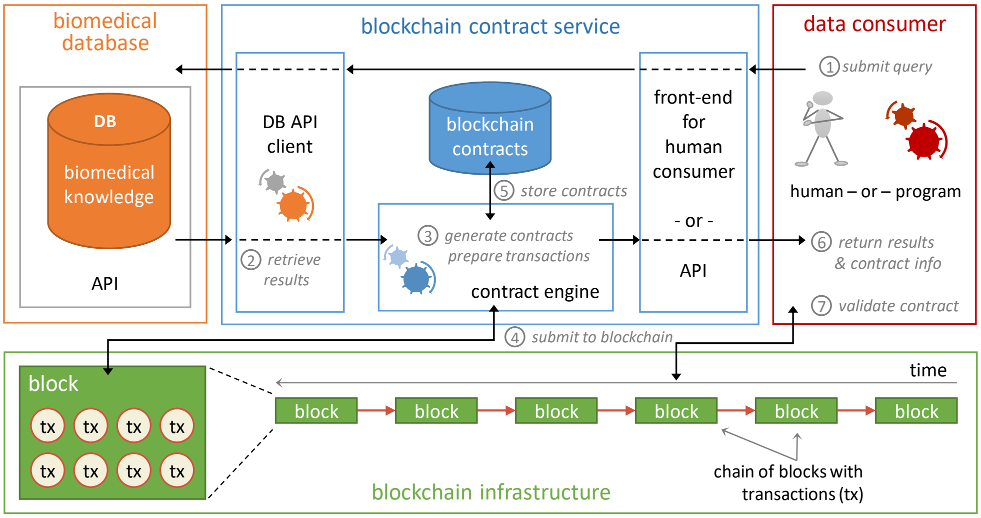


Fig. 1 The general architecture of our query notary service.

The proposed notary service exhibits three computational layers: (a) a data consumer front-end, which can be either an interface for human data consumers or an application programming interface (API) for 3rd party programs that request data from a biomedical database; (b) an interface to communicate with biomedical database interfaces, which is specific to each database API; and (c) the contract generation engine, which collates the query/results data and the consumer, prepares transactions and submits them to a blockchain infrastructure, and stores contract information (contract address and its application binary interface).

The workflow of the notary service is as follows. First, the data consumer front-end undertakes the communication with the party placing the query to the database. In its simplest version, the query is forwarded to the database API via the database API client. As an added-value, the query can also be signed by a public key infrastructure to verify later the identity of the data consumer. The API client places the query via the database API and retrieves the results; both (signed) query and results are forwarded to the blockchain contract service. Subsequently, these data are hashed (e.g. using SHA256 [20]) and the hash is included in a smart contract that is deployed to a blockchain infrastructure.

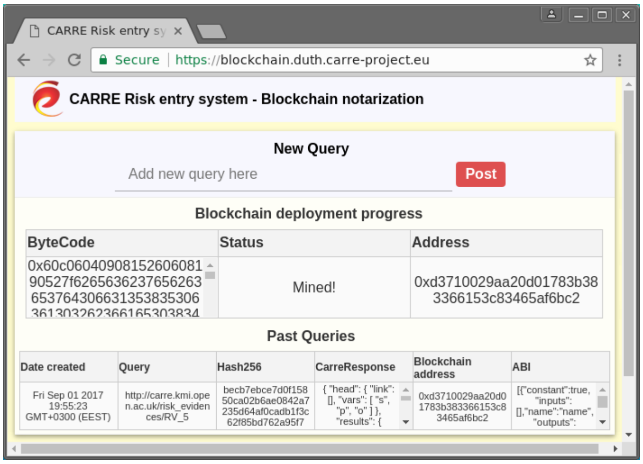
The contract generation engine then returns the query results to the data consumer via the front-end, accompanied by the smart contract's address on the blockchain, the application binary interface (ABI) to interact with the contract, and the (signed) query and its results. A respective entry is also made into the local contract database. The packet returned to the data consumer contains also database certification information to verify the identity of the database and thus ensure query transaction non-repudiation (for example, the database blockchain public key signed by a digital certification authority). The consumer archives the query transaction (query and signed response) in a local database for future reference.

At any later time, the data consumer or any third party can verify the query transaction dataset by retrieving the respective contract from the blockchain infrastructure and comparing the retrieved hash of the original data with a new hash of the claimed (signed) query and respective results.

IV. Implementation & Experimental Results

The proposed architecture was implemented to provide query notary services for the CARRE risk factor reference repository [7], an open, online database collecting current high-level evidence on risk factors for the cardiorenal syndrome and related comorbidities. In this repository, risk factors are described in a structured way following the CARRE risk factor ontology [21]. Risk evidence descriptions are manually entered by authorized medical experts following a collaborative literature survey process by which appropriate medical publications of high level medical evidence are identified in PubMed and used to extract state-of-the-art medical evidence on risk factors related to cardiorenal disease. The resulting risk factor descriptions are available as Linked Data, following the Resource Description Framework (RDF) format (http://www.w3.org/TR/rdf-syntax), via an open access RDF repository. Currently the CARRE risk factor repository describes more than 100 different risk factors corresponding to 250 risk associations between more than 50 medical conditions related to cardiorenal disease as retrieved from 65 scientific publications.

In this demonstration, blockchains were implemented using the Ethereum infrastructure [6], which in addition to the transaction introduces a programmable logic into the blocks. This functionality allows for smart programmable contracts that have the ability to work autonomously in an if-this-then-that fashion. The use of the Ethereum blockchain infrastructure requires running an Ethereum node using the Geth client (version 1.5.9). Smart contracts are implemented in the Solidity language (https://solidity.readthedocs.io), while a MongoDB database (https://www.mongodb.org) is deployed for the local storage of contracts and respective information.



CARRE Risk entry system - Blockchain notarization

New Query

Blockchain deployment progress

ByteCode	Status	Address
0x60c06040908152606081 90527f62656362376562 63 6537643066313538353 06 361303262366165303834	Mined!	0xd3710029aa20d01783b38 3366153c83465af6bc2

Past Queries

Date created	Query	Hash256	CarreResponse	Blockchain address	ABI
Fri Sep 01 2017 19:55:23 GMT+0300 (EEST)	http://carre.kmi.ope n.ac.uk/risk_eviden ces/RV_5	becb7ebce7d0f158 50ca02b6ae0842a7 235d64af0cadb1f3c 62f85bd762a95f7	{ "head": { "link": [], "vars": ["s", "p", "o"] }, "results": {	0xd3710029aa20d0 1783b383366153c8 3465af6bc2	[{"constant":true, "inputs": [],"name":"name", "outputs":

Fig. 2 Snapshot of contract deployment process in blockchain.

The front-end was implemented using JavaScript and Ajax asynchronous requests to establish communication with the CARRE RDF repository via its SPARQL endpoint client (https://devices.duth.carre-project.eu/sparql). The Meteor web framework (https://www.meteor.com) was used to connect the front-end with the backend of the query notary service.

Fig. 2 shows a snapshot of the notary service as implemented for CARRE risk factor repository. When a new query is placed via the front end, the notary service communicates with the CARRE repository and creates a smart contract out of the query and the returned results. The contract is compiled locally and then the bytecode is deployed on the Ethereum blockchain. The front end provides information on the status of the procedure and information on past queries and their respective application binary interface on the Ethereum blockchain.

Experimental verification used the Ropsten Ethereum test network (https://testnet.etherscan.io), which simulates the blockchain environment but is provided for free. At the time of testing in July 2017, the transaction confirmation delay was 30 to 50 seconds, and the cost of placing a transaction into the Ethereum blockchain was 0.00302 Ether (Ethereum's cryptocurrency). With an exchange rate of 1

Ether = 171.64€ (31 Jul 2017, www.worldcoinindex.com), the cost of one transaction was about 0.52€.

V. Discussion

This paper proposes a query notary for biomedical data consumers (humans or programs alike) who need to retrieve accurate and certified data from reference biomedical databases. The proposed approach utilizes blockchain technology and is implemented using a real blockchain infrastructure and a publicly available medical reference repository. The notary is realized following the approach of software-as-a-service, thus bringing the cost to the data consumer on a needs basis. Additionally, a private blockchain network maintained by health regulators, such as healthcare establishments and medical research organizations (similar to that proposed in [19]) could be established to alleviate this cost. Work in progress includes design and development of a mechanism that exploits the blockchain to provide irrefutable version control of database content and give undisputable proof that the results returned via a query correspond to the most recent update. Additional work addresses smart contracts involving dynamic graph data (e.g. Linked Open Data cloud datasets), where the question is to combine certified sub-graphs (for example from different repositories) in order to validate larger, integrated data graphs.

Acknowledgment

This work was partly supported by the European Commission FP7-ICT project CARRE (Grant No. 611140).

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Williams WG (2010) Uses and limitations of registry and academic databases. Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu 13:66-70
2. Duck G, Nenadic G, Filannino M et al (2016) A Survey of Bioinformatics Database and Software Usage through Mining the Literature. PloS One 11:e0157989
3. Vaughan K, Scolaro KL, Anksorus HN et al (2014) An evaluation of pharmacogenomic information provided by five common drug information resources. J Med Libr Assoc 102:47
4. Go EP (2010) Database Resources in Metabolomics: An Overview. J Neuroimmune Pharmacol 5:18-30
5. Falagas ME, Pitsouni EI, Malietzis GA et al (2008) Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses. FASEB J 22:338-342
6. Buterin V (2014) A next-generation smart contract and decentralized application platform. White Paper, http://www.ethereum.org/pdfs/EthereumWhitePaper.pdf
7. CARRE P (2016) CARRE risk factor reference repository. FP7 EU project (FP7-ICT-611140). Available at: https://www.carre-project.eu/innovation/carre-risk-factor-entry-system
8. Nakamoto S (2008) Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf
9. Cachin C (2016) Architecture of the Hyperledger blockchain fabric. Workshop on Distributed Cryptocurrencies and Consensus Ledgers (DCCL); Chicago, Illinois, USA
10. Herrera-Joancomartí J (2015) Research and challenges on bitcoin anonymity. Data Privacy Management, Autonomous Spontaneous Security, and Security Assurance. LNCS, vol. 8872: Springer, Cham:3-16
11. Housley R, Ford W, Polk W et al (1998) Internet X. 509 public key infrastructure certificate and CRL profile. 2070-1721
12. Matsumoto S, Reischuk RM (2016) IKP: Turning a PKI Around with Blockchains. IACR Cryptology ePrint Archive 2016:1018
13. Yli-Huumo J, Ko D, Choi S et al (2016) Where Is Current Research on Blockchain Technology?-A Systematic Review. PLoS One 11:e0163477
14. Mettler M (2016) Blockchain technology in healthcare: The revolution starts here. 18th Int Conf on e-Health Networking, Applications and Services; Munich, Germany, pp 1-3
15. Yue X, Wang H, Jin D et al (2016) Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control. J Med Syst 40:218
16. Roehrs A, da Costa CA, da Rosa Righi R (2017) OmniPHR: A distributed architecture model to integrate personal health records. J Biomed Inform 71:70-81
17. Azaria A, Ekblaw A, Vieira T et al (2016) Medrec: Using blockchain for medical data access and permission management. Int Conf on Open and Big Data (OBD) pp 25-30
18. Carlisle BG (2014) Proof of prespecified endpoints in medical research with the bitcoin blockchain. Web Blog, The Grey Literature https://www.bgcarlisle.com/blog/2014/08/25/proof-of-prespecified-endpoints-in-medical-research-with-the-bitcoin-blockchain/. Accessed 29 August, 2017
19. Nugent T, Upton D, Cimpoesu M (2016) Improving data transparency in clinical trials using blockchain smart contracts. F1000Res 5:2541
20. Penard W, van Werkhoven T (2008) On the secure hash algorithm family. National Security Agency, Tech. Rep.
21. Third A, Kaldoudi E, Gkotsis G et al (2015) Capturing scientific knowledge on medical risk factors. 1st Int. Workshop on Capturing Scientific Knowledge, collocated with the 8th Int. Conf. on Knowledge Capture (K-CAP); Palisades, NY, USA

Corresponding author:

Author: George Drosatos
Institute: School of Medicine, Democritus University of Thrace
Street: University Campus, Dragana
City: Alexandroupoli
Country: Greece
Email: gdrosato@ee.duth.gr