Introducing the Open Science Chain - Protecting Integrity and Provenance of Research Data

Subhashini Sivagnanam*
Viswanath Nandigam*
Kai Lin
sivagnan@sdsc.edu
viswanat@sdsc.edu
klin@sdsc.edu
University of California San Diego
San Diego Supercomputer Center
9500 Gilman Dr MC 0505, La Jolla, California

ABSTRACT

Data sharing is an integral component of research and academic publications, allowing for independent verification of results. Researchers have the ability to extend and build upon prior research when they are able to efficiently access, validate, and verify the data referenced in publications. Despite the well known benefits of making research data more open, data withholding rates have remained constant. Some disincentives to sharing research data include lack of credit, and fear of misrepresentation of data in the absence of context and provenance. While there are several research data sharing repositories that focus on making research data available, there are no cyberinfrastructure platforms that enable researchers to efficiently validate the authenticity of datasets, track the provenance, view the lineage of the data and verify ownership information. In this paper, we introduce and provide an overview of the NSF funded Open Science Chain, a cyberinfrastructure platform built using blockchain technologies that securely stores metadata and verification information about research data and tracks changes to that data in an auditable manner in order to address issues related to reproducibility and accountability in scientific research.

CCS CONCEPTS

• **Information systems** → **Data provenance**; *Integrity checking*;
• **Security and privacy** → *Cryptography*; Trust frameworks.

KEYWORDS

Data Reproducibility, Data Provenance, Data Integrity, Cryptography, Blockchain, Distributed Ledger Technology

ACM Reference Format:
Subhashini Sivagnanam, Viswanath Nandigam, and Kai Lin. 2019. Introducing the Open Science Chain - Protecting Integrity and Provenance of Research Data. In *Practice and Experience in Advanced Research Computing (PEARC '19), July 28-August 1, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3332186.3332203

*Both authors contributed equally to this research.

PEARC '19, July 28-August 1, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-7227-5/19/07…$15.00
https://doi.org/10.1145/3332186.3332203

1 INTRODUCTION

There is a known credibility and reproducibility issue in scientific research [2, 7–9, 11, 12, 14, 20, 21, 25] that is a direct consequence of low research data sharing rates. The credibility of the research findings comes into question when the results cannot be replicated with limited available data or when data is not made available at all. Results from a recent study [16] show that many published Computer Science research hypothesis were not replicable within the context of the referenced computational code. Out of the 613 articles published in 13 top-tier systems research conference publications, the study found that only 25% of the results from the articles are replicable due to various reasons including modified data or unavailable data. Similarly, researchers [13] have found that almost 90% of microarray studies from a leading genetics journal are not fully reproducible, mostly due to incomplete availability of the data and methods.

Facilitating the future reuse of data in a secure and independently verifiable manner is critical to the advancement of research. Realizing the importance of making research data available for the community, several funding agencies now require that the data be made available post research phase in order to increase confidence and trust in the research work [17, 18] and to enable reusability in future scientific research. Some publishers now have made data sharing mandatory and a prerequisite for inclusion in publications [19].

Data sharing policies by funding agencies and publishers along with the availability of third party data repositories encourage sharing of research datasets. However, a cross-sectional study of data sharing and withholding in the life sciences shows that the percentage of data withheld remained constant [26]. Disincentives to sharing research data include lack of credit and concerns about intellectual property [6], protecting scientific lead [26], and fear of misrepresentation or misuse of data, especially if the provenance of the data is not available to provide context.

While there are several research data sharing repositories (e.g. figshare, Dryad Digital Repository) that focus on making research data available, there are no cyberinfrastructure (CI) platforms that enable researchers to efficiently validate the authenticity of datasets,

track the provenance, view the lineage of the data or verify ownership information.

2 BACKGROUND

Data referenced in scientific publications are typically stored on cloud or local resources and made accessible via a public URL (e.g. [3, 15, 22]). Data stored on publicly accessible resources have a risk of being lost if the underlying resources change. A majority of research data generated by researchers or by using science gateway resources is typically shared without any form of persistent identifiers. While Digital Object Identifiers (DOIs) are being more commonly used in publications, they have their limitations especially with regard to scientific research datasets. DOIs cannot verify if the content pointed to by the identifier changes. As research datasets evolve, versioning and provenance becomes an issue.

2.1 Blockchain Technology

Blockchain, a type of distributed ledger technology, offers a secure cryptographically protected record of transactions (blocks). Apart from the transaction information, a typical entry in the ledger consists of a unique cryptographic signature, a timestamp of the transaction and a previous block's hash. Blockchain's "append only" structure prevents altering or deleting previously entered data. Data in the blockchain ledger is therefore verifiable and immutable, which is essential for reproducibility and audits. In a distributed ledger environment, the ledger is distributed across multiple nodes or peers. Updates to the ledger are done independently at each of the peers and are based on a consensus algorithm. Once consensus is reached at all the peers, the ledgers get updated simultaneously. The distributed nature not only prevents a single point of failure but also allows each node that participates in the network to verify the ledger data, further enhancing security.

Blockchain networks can be public, private or consortium-based [5]. A public blockchain is truly decentralized, permissionless, open to anyone and secured by cryptoeconomics such as Bitcoin. However cryptoecomonics have a computationally expensive overhead called mining to ensure trust among anonymous users by legitimizing transactions. Fully private blockchains are centralized to a single organization with restricted access to members of that organization. Consortium blockchains are semi-private in the sense that consensus in the network is controlled by a limited set of nodes, all participants have known identities and transactions do not require cryptoeconomics which improves transaction performance. Consortium blockchain is ideally suited for research and academic environments where organizational collaboration is common and security efficiencies and scalability are preferred over the computational cryptoeconomics and anonymity.

3 OPEN SCIENCE CHAIN

The Open Science Chain (OSC - www.opensciencechain.org) project, recently funded by the National Science Foundation is building a CI platform, using consortium blockchain technologies, with the goal of enabling a broad set of researchers to efficiently share, verify and validate the authenticity of scientific data while preserving the provenance and proof of existence and ownership.

Verification information about the data (SHA256 hash) is stored as a manifest in the blockchain along with the metadata. The actual data is stored off-chain. Storing large amounts of data in the blockchain is inefficient especially since some scientific datasets tend to be in the multi-terabytes size range. Storing only the comprehensive metadata of a dataset enables researchers to share large datasets or sensitive data that are stored off-chain, yet verifiable with the information stored on-chain with the OSC. Envisioning the future growth of OSC with additional member organizations joining the network, having actual data off-chain ensures that every transaction executed on every individual node with varying compute capabilities remains efficient.

OSC is agnostic to the type of data whose verification information is stored in the blockchain. In the initial implementation of the OSC data model, we are requiring the location of the dataset (e.g DOI, Accession number, URL), dataset verification information (SHA256 Hash) and information about the contributor as mandatory parts of the metadata but will gradually expand to include other optional metadata elements based on feedback and usage analysis. OSC will generate verification identifier for the information stored on the chain that uniquely ties together metadata elements such as contributor information, location of the data, and cryptographic hash of the data.

Providing an efficient web based portal with seamless user interfaces and multi-platform client tools for interacting with the underlying blockchain platform is critical to lowering the complexity barrier for use of this technology and promoting adoption by the wider scientific community. We are building the OSC web portal to interact with the OSC blockchain as described in section 4.2.

OSC requirements are gathered from several use cases encompassing a variety of science disciplines and research communities including science gateways [23] and research labs. Many science gateways are driven by the explosion of data in their domain and offer platforms for data distribution and computational analysis on these datasets to their communities. Numerous peer-reviewed articles, publications and graduate students' research work have resulted from using gateway resources and this number is only growing [24]. Preserving provenance of derived data as well as the algorithms from these gateways, especially those used in publications is vital due to the dynamic nature of the resources and data themselves. Similarly research labs typically span research over a long time period and produce data during various stages that continues to evolve with new breakthroughs, and changes in technologies. These data become more valuable over time and retaining the provenance and lineage of previous data iterations is important, especially when it is used at various stages of its evolving lifecycle in publications.

4 OSC ARCHITECTURE

The main architecture components of the OSC, as described below, are the consortium blockchain platform, middleware services, and the application space that includes OSC web portal and client tools. In OSC, users authenticate using the CILogon [4] federated identity and are mapped to individual identities within the blockchain network. A conceptual overview of the OSC architecture is shown in figure 1.

4.1 OSC Consortium Blockchain Platform

At the core of OSC architecture is the OSC consortium blockchain platform implemented using Hyperledger Fabric [1], an open source blockchain framework from the Linux Foundation. Hyperledger Fabric is highly modular allowing for creation of pluggable components like an identity management system and other applications on top of this architecture. The OSC consortium blockchain network comprises of two endorsing peer nodes with the ledger and an ordering node that is customized to support OSC use case requirements.

A peer can be either an endorsing peer or a committing peer. An endorsing peer accepts transaction requests initiated from the application level clients (e.g. OSC portal), validates and verifies it authority and executes the chaincode against the current ledger. A committing peer simply maintains the ledger by committing transactions. The two endorsing peer nodes is set up to avoid single point of failure. The ordering node provides a communication channel for clients and peers, serving as a broadcast service for messages and guaranteeing delivery and order of transactions. The ledger holds record of all transactions or blocks and a state database.

A block is comprised of three segments - the header, block metadata and data. Each block contains a cryptographic record linking it to the previous block. The header and block metadata segments are standard components of a Hyperledger Fabric block structure [10]. The header will comprise of a unique block number, the hash of the previous block's header and hash of the current block's data segment. Block metadata segment stores entries added by the ordering service and information of number of transactions in the block. The data segment is the main portion of the block that is customized for OSC. This includes transaction data, signatures, peer certificates, metadata and other information based on the OSC data model.

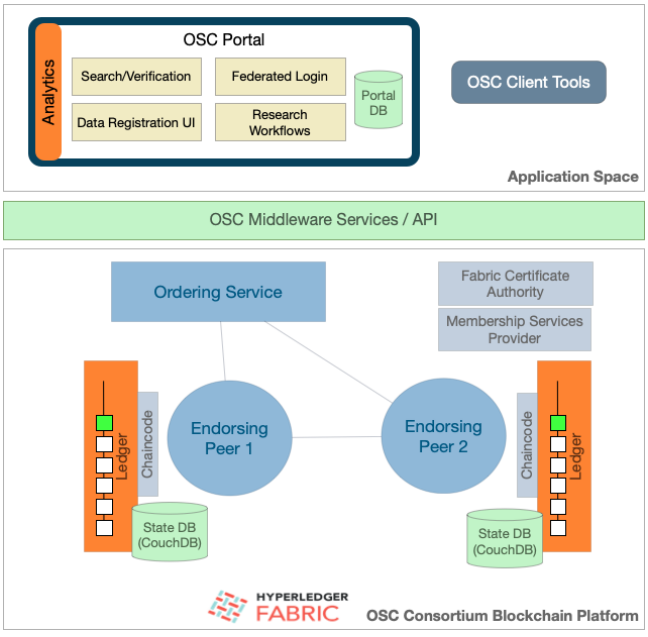

Figure 1: OSC Architecture

Transactions represent a modification to the ledger. For example, if a user updates the size or location of a research dataset, a new transaction gets created with the new information and a cryptographic signature of the new information that identifies who made the change. In OSC, individual contributors rather than a common organization account cryptographically sign each transaction. This ensures an auditable record that associates changes to a data asset with individual contributors via cryptographically signed transactions. This will also allow for future OSC functionality that enables users to continue their work on existing research data in OSC by migrating their credentials from the previous institutional account to the new one.

Transactions are processed by chaincode, which then commits the transaction to the ledger. OSC chaincode rules also restrict updates to original contributors of the record in the blockchain as well as restrict duplicate entries of the same data or data collections. The ledger also contains a state database (CouchDB) to maintain current state. The ledger using the state database will represent the latest values for all keys ever included in the blockchain and provide a verifiable history of all valid and invalid transactions occurring during the operation of the system. By using CouchDB together with JSON, we plan to take advantage of rich querying capabilities and will build ledger data query applications on top of it. The built-in membership services provider (MSP) and Fabric certificate authority (CA) generates the certificates and keys for the peers and orderers to digitally sign transactions and prove their membership in a network. All user transactions requests are signed by these certificates and executed within the network.

4.2 OSC Web Portal

The OSC portal will include applications to facilitate user registration, data registration, data analytics, and search capabilities. Extensive user guides and documentation to use the OSC portal and applications will be available for the researchers. The OSC portal comprises of user-friendly interfaces for registering metadata and to search and verify datasets. During the dataset registration and update process, the web application guides users through an intuitive user interface, including metadata entry for datasets, auto calculating the SHA256 checksum for the data with a drag and drop feature for smaller datasets. When updates are made to a dataset or data collection, all metadata changes including the SHA256 checksums for each and every file in that data collection are tracked, enabling users to view a detailed evolution history of that dataset over time.

Researchers also have the ability to develop "research workflows" linking data entries in the ledger creating an auditable record of the data workflow process behind the research findings (e.g referring to specific versions of the source data and algorithms in the hypothesis). While the state database always has the most latest values of the dataset information, the ledger will contain detailed prior transaction information including what changes were made, at what time and who made the changes. Researchers will have the ability to use the search interface to look up all these detailed dataset information including transaction details. Researchers can also provide feedback to the owner of the data especially when the data registered with OSC cannot be validated.

4.3 OSC Middleware Services

We are developing middleware services and application programming interfaces (API) to support client tools for multiple platforms. These client tools will enable researchers to register their data directly from their systems and other lab environments. For example, using the client tools, researchers will be able to register derived datasets from high performance computing simulations. Making these API public and widely available will also promote independent third party software tool development that will conversely see an increase in usage.

5 OSC TRANSACTION WORKFLOW

Prior to users interacting with the OSC blockchain (via the OSC portal or client tools), the data model and endorsement policy is defined and chaincode is deployed on the endorsing peers. The orderer and peers are registered with the CA. The following steps outline the transaction flow that happens within the OSC blockchain platform when a researcher uses an application level client, such as the OSC web portal or client tools, to either create or update a transaction. High level transaction workflow is also shown in figure 2 below.

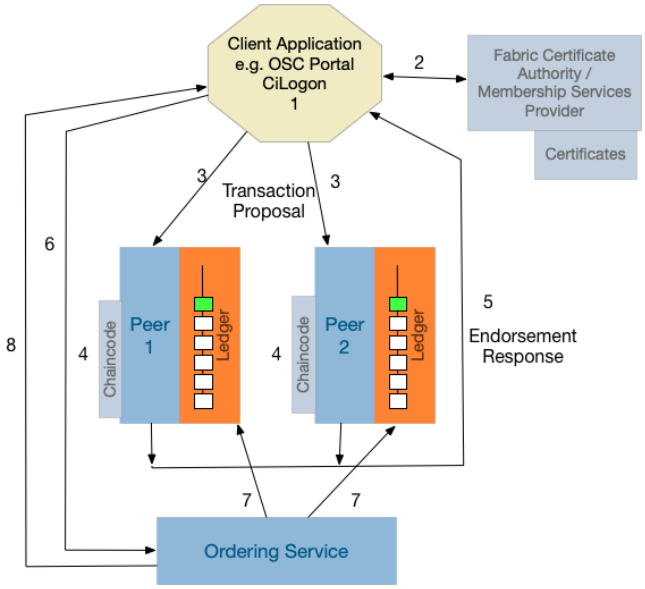

Figure 2: OSC Transaction Flow

(1) Users authenticate using federated identity such as CILogon on the client end points.
(2) Users are mapped to individual identities within the OSC blockchain network.
(3) The OSC client end point submit a transaction request (e.g. updating the location of a dataset) on behalf of the user to the endorsing peers.
(4) The endorsing peers executes the OSC chaincode (e.g. if user is original contributor and allowed to update; ensuring uniqueness of data asset) against the transaction request, validates the transaction request format and certificate details to ensure that the user has the authority to update the block.
(5) If transaction is valid, each endorsing peer returns an endorsement response back to the client without updating the ledger at this point. An invalid transaction is returned without further action.
(6) The client sends the approved transaction to the ordering node that properly orders and includes it in a block. The transactions within the block are validated to ensure endorsement policy is fulfilled and then sent to all the nodes in the network atomically.
(7) The individual peers in the network then update their ledgers with the latest block (for e.g adding information containing the updated location).
(8) The client application is notified with a response whether the transaction has been appended to the chain.

As we get more organizations joining the OSC consortium, we anticipate an corresponding increase in number of peers in the network. While having an extended network with more peers leads to a more independent and decentralized system, additional challenges arise related to choosing the right consensus mechanism, evaluating transaction performance, etc.

6 CONCLUSION

The main thrust of this paper is to introduce the Open Science Chain project whose primary mission is to provide a CI framework where scientific data can be securely tracked and independently verified regardless of the domain science. OSC provides researchers the ability to register information about their scientific data and provides immutable proof of existence of research data at a given point in time by storing unique identifiers of the data and ownership information on the blockchain. OSC also promotes transparency and traceability of research data by tracking and storing all changes made to data on the blockchain. Other researchers have the ability to independently verify authenticity of scientific data as well as its lineage using information stored in the OSC blockchain. OSC aims to increase the confidence of the scientific results and enhance data sharing and reuse, which will result in greater research productivity and reproducibility.

7 ACKNOWLEDGMENTS

Open Science Chain is supported by the National Science Foundation under Award Number 1840218. Authors would like to thank Dmitry Mishin at the San Diego Supercomputer Center, University of California San Diego for his technical contributions to the project.

REFERENCES

[1] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, et al. 2018. Hyperledger fabric: a distributed operating system for permissioned blockchains. In *Proceedings of the Thirteenth EuroSys Conference*. ACM, 30.
[2] Marcia Angell. 2009. Drug companies & doctors: A story of corruption. *The New York Review of Books* 56, 1 (2009), 8–12.
[3] Lei Bao, Minya Pu, and Karen Messer. 2014. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics* 30, 8 (2014), 1056–1063.
[4] Jim Basney, Terry Fleury, and Jeff Gaynor. 2014. CILogon: A federated X. 509 certification authority for cyberinfrastructure logon. *Concurrency and Computation: Practice and Experience* 26, 13 (2014), 2225–2239.

[5] Vitalik Buterin. 2015 (accessed April 11, 2019). *On Public and Private Blockchains.* https://blog.ethereum.org/2015/08/07/on-public-and-private-blockchains/

[6] Stephanie OM Dyke and Tim JP Hubbard. 2011. Developing and implementing an institute-wide data sharing policy. *Genome medicine* 3, 9 (2011), 60.

[7] Daniel Engber. 2016 (accessed April 11, 2019). *Cancer Research Is Broken.* https://slate.com/technology/2016/04/biomedicine-facing-a-worse-replication-crisis-than-the-one-plaguing-psychology.html

[8] Leonard P Freedman, Iain M Cockburn, and Timothy S Simcoe. 2015. The economics of reproducibility in preclinical research. *PLoS biology* 13, 6 (2015), e1002165.

[9] Ben Goldacre. 2013. Are clinical trial data shared sufficiently today? No. *Bmj* 347 (2013), f1880.

[10] IBM Research India. 2017 (accessed April 11, 2019). *Hyperledger Fabric V1.0: Block Structure (Part 1).* https://blockchain-fabric.blogspot.com/2017/04/hyperledger-fabric-v10-block-structure.html

[11] John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine* 2, 8 (2005), e124.

[12] John PA Ioannidis. 2015. How to make more published research true. *Revista Cubana de Información en Ciencias de la Salud (ACIMED)* 26, 2 (2015), 187–200.

[13] John PA Ioannidis, David B Allison, Catherine A Ball, Issa Coulibaly, Xiangqin Cui, Aedín C Culhane, Mario Falchi, Cesare Furlanello, Laurence Game, Giuseppe Jurman, et al. 2009. Repeatability of published microarray gene expression analyses. *Nature genetics* 41, 2 (2009), 149.

[14] John PA Ioannidis, Sander Greenland, Mark A Hlatky, Muin J Khoury, Malcolm R Macleod, David Moher, Kenneth F Schulz, and Robert Tibshirani. 2014. Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet* 383, 9912 (2014), 166–175.

[15] Andy Jarvis, Hannes Isaak Reuter, Andrew Nelson, and Edward Guevara. 2008. Hole-filled SRTM for the globe Version 4. (2008).

[16] Gina Moraila, Akash Shankaran, Zuoming Shi, and Alex M Warren. 2014. *Measuring reproducibility in computer systems research.* Technical Report. Technical report, University of Arizona.

[17] NIH. 2003 (accessed April 11, 2019). *Final NIH Statement in Sharing Research Data.* https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

[18] NSF. 2011 (accessed April 11, 2019). *Digital Research Data Sharing and Management.* https://www.nsf.gov/nsb/publications/2011/nsb1124.pdf

[19] PLOS ONE. 2011 (accessed April 11, 2019). *Data Availability.* https://journals.plos.org/plosone/s/data-availability

[20] pubpeer. 2014 (accessed April 11, 2019). *A crisis of trust.* https://blog.pubpeer.com/publications/EE46E14F6AA97049928835DBC6B908

[21] Leonid Schneider. 2016 (accessed April 11, 2019). *Voinnet aftermath: ethical bankruptcy of academic elites.* https://forbetterscience.com/2016/04/07/voinnet-aftermath-ethical-bankruptcy-of-academic-elites/

[22] Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. 2010. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 3 (2010), 431–432.

[23] XSEDE. 2019 (accessed April 11, 2019). *XSEDE Science Gateways.* https://www.xsede.org/ecosystem/science-gateways

[24] XSEDE. 2019 (accessed April 15, 2019). *XSEDE Publications.* https://portal.xsede.org/publications/

[25] Neal S Young, John PA Ioannidis, and Omar Al-Ubaydli. 2008. Why current publication practices may distort science. *PLoS medicine* 5, 10 (2008), e201.

[26] Darren E Zinner, Genevieve Pham-Kanter, and Eric G Campbell. 2016. The changing nature of scientific sharing and withholding in academic life sciences research: trends from national surveys in 2000 and 2013. *Academic medicine: journal of the Association of American Medical Colleges* 91, 3 (2016), 433.