EXPLORING BLOCKCHAIN TO SUPPORT OPEN SCIENCE PRACTICES

Iksha Gurung,Slesa Adhikari, Abdelhak Marouane, Rajesh Pandey, Satkar Dhakal- 2, Manil Maskey - 3[*]

University of Alabama in Huntsville, 2 - Vanderbilt University, 3 - NASA Marshall Space Flight Center

ABSTRACT

Open science aims to foster transparent sharing of scientific processes including open access, incentivization, provenance, open source code and tools, metrics, and resource sharing. However, effective management of these processes remains a challenge. This paper explores the application of blockchain technology to address these key aspects of open science. Blockchain offers a decentralized and secure platform for information exchange and verification. By leveraging blockchain, open science can enhance transparency and reproducibility. In this paper, we present an implementation of blockchain for Earth science data synchronization across organizations, enabling tracking of data copying, citation, and download. The findings highlight the potential of blockchain in supporting open science objectives.

Index Terms— Blockchain, Open Science

1. INTRODUCTION

Open science is a transformative movement that advocates for the sharing of scientific resources, promoting collaboration, and enhancing reproducibility [1]. It entails making research data, methodologies, and findings openly accessible to the broader scientific community and the public. While open science has gained momentum in recent years, its successful implementation faces various challenges, including the management of scientific resources, data verification, proper attribution, and ensuring transparency throughout the research process.

One emerging technology that shows great promise in addressing these challenges is blockchain [2]. Blockchain is a decentralized and secure technology that enables the exchange, verification, and immutable recording of information across a network. It is renowned for its application in cryptocurrencies like Bitcoin, but its potential extends far beyond financial transactions.

In the context of open science, blockchain offers several advantages that align with the core principles of open science. By leveraging blockchain, scientists can establish a decentralized environment for sharing and verifying authoritative data, ensuring traceability of scientific resources, and enabling proper attribution and credit within the realm of open science. Blockchain's inherent features, such as decentralization and cryptographic techniques, make it difficult to alter data without consensus, thereby ensuring the authoritativeness and integrity of scientific resources.

Moreover, blockchain technology allows researchers to be incentivized and fosters collaboration. Through blockchain-based incentive mechanisms, scientists can securely and openly share their data and be rewarded for their contributions. This not only encourages the dissemination of scientific knowledge but also promotes more efficient and productive research practices.

While the potential benefits of blockchain in open science are significant, its successful integration into existing scientific workflows requires careful consideration of various factors. Policy frameworks for access, citation, and data management need to be established to ensure fair and ethical use of blockchain-based systems. Furthermore, the additional costs associated with managing blockchain infrastructure and the need for a cultural shift towards embracing this technology pose additional challenges that need to be addressed.

In this paper, we explore the practical insights and implementation of blockchain technology for advancing open science. We present a case study where an Amazon Web Services (AWS) hosted blockchain [3] was utilized to synchronize Earth science data among multiple organizations. The blockchain was configured to track data copying, citation, and download; a monitoring dashboard was developed to visualize the blockchain's activities. The findings from this study contribute to the understanding of how blockchain can address the challenges faced in open science, provide a foundation for future research and development in this domain, and pave the way for a more efficient and accountable scientific ecosystem.

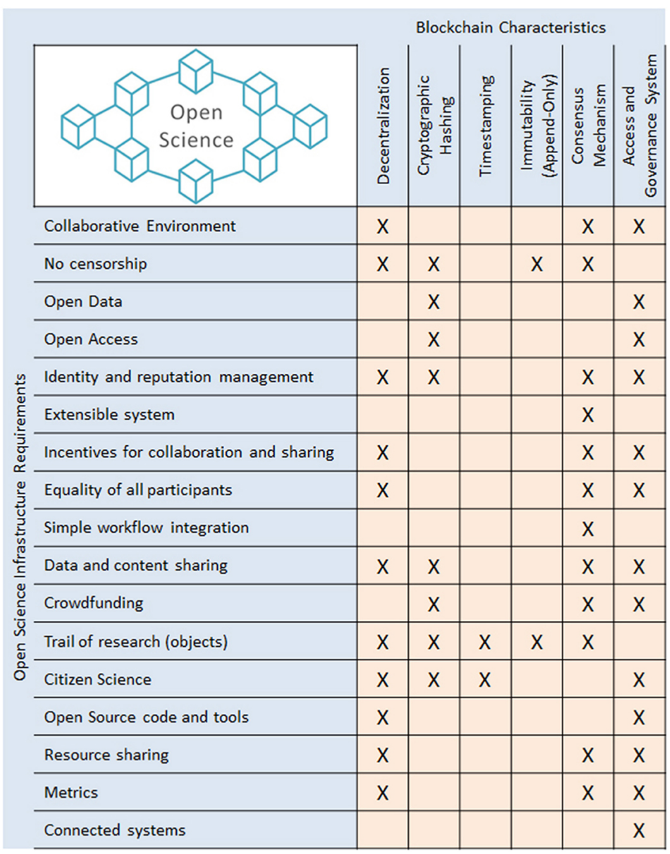
2. BLOCKCHAIN

Blockchain is a decentralized, distributed ledger technology. It provides a secure and transparent way to record and verify transactions or data. Information (i.e., a transaction) is stored in the form of blocks. These blocks are chained together to provide an immutable ledger of transactions performed. Due to the decentralized, and distributed nature of blockchain, it does not rely on a central authority for verification or con-

[*]This research is supported by NASA Grant NNM11AA01A as part of the IMPACT project.

trol. Instead, it uses a consensus mechanism to validate and add new blocks to the chain. This ensures the integrity and immutability of the data stored on the blockchain.

Blockchain technology is the backbone of crypto-currencies such as Bitcoin and Etherium. There are different types of blockchains, including public, private, and consortium blockchains. Crypto-currencies utilize open blockchains. These blockchains are open to the public and allow anyone to participate in the network. Private blockchains, on the other hand, are restricted to specific participants where the control of the network is more centralized. Consortium blockchains are a hybrid, where multiple organizations collaborate and jointly control the blockchain network. To maintain proper channels of data distribution while still providing open access to organizations and contributors, we utilized an AWS managed blockchain, which is a form of a consortium blockchain.



Open Science Infrastructure Requirements	Decentralization	Cryptographic Hashing	Timestamping	Immutability (Append-Only)	Consensus Mechanism	Access and Governance System
Collaborative Environment	X				X	X
No censorship	X	X		X	X	
Open Data		X				X
Open Access		X				X
Identity and reputation management	X	X			X	X
Extensible system					X	
Incentives for collaboration and sharing	X				X	X
Equality of all participants	X				X	X
Simple workflow integration					X	
Data and content sharing	X	X			X	X
Crowdfunding		X			X	X
Trail of research (objects)	X	X	X	X	X	
Citizen Science	X	X	X			X
Open Source code and tools	X					X
Resource sharing	X				X	X
Metrics	X				X	X
Connected systems						X

Fig. 1. Mapping of open science infrastructure requirements and blockchain characteristics from [2]

Blockchain has multiple characteristics that make it suitable for open science. Fig. 1 outlines multiple open science infrastructure requirements which could be fulfilled by blockchain [2]. First, blockchains are decentralized in that information is not controlled by a central entity and the stored information is immutable. This allows for an open collaborative environment, fosters citizen science, enables resource sharing, and allows for data and content sharing. Second, it offers transparency. All the transactions are recorded and publicly available for all members. This transparency helps in building trust and ensuring accountability. Third, blockchain uses cryptographic hashing and timestamping, meaning all the information recorded in the blockchain is unique. This feature allows scientists to view trails of research and enable data and content sharing. Additionally, blockchain utilizes a consensus mechanism, meaning all members in the network must approve of the information proposed or stored in the network. This enables efficient and accurate metrics calculations, such as the number of downloads or citations per dataset, the verification of dataset metadata without intermediaries, etc.

3. OPEN SCIENCE BLOCKCHAIN USECASE

As a use case for an Open Science Blockchain (OSBC), we used the Visualization, Exploration, and Data Analysis (VEDA) project. VEDA is an initiative from National Aeronautics and Space Administration (NASA)'s Earth Science division that utilizes open-source tools and resources to create an open-source science cyberinfrastructure for data processing and geographic information systems (GIS) capabilities [4]. VEDA allows scientists from different organizations and teams to contribute datasets. These datasets are then used to create reproducible analysis and discoveries that highlight multiple events happening around the globe.

In the context of VEDA, blockchain can serve as a decentralized entity that enables researchers to validate the authoritativeness of VEDA datasets. Blockchain allows researchers to track the provenance of the VEDA datasets, ensuring their origin and integrity. This allows for greater trust and reliability in the research findings. Additionally, blockchain can facilitate data citation, making it easier to attribute credit and recognition to the original creators of the datasets.

NASA scientists use NASA's High-End Computing program (HEC) [5] for large scale modeling and simulation. The same scientists from HEC consume and contribute multiple datasets and discoveries to VEDA. HEC also maintains its own storage and authentication method which are not discussed in detail here. For our use case we demonstrate how HEC, as a different entity, is able to contribute to the VEDA data store while following the principles of open science powered by OSBC.

3.1. Architecture

AWS provides a managed blockchain as a service [3]. It provides pre-configured templates and configurations which allow for faster adaptation of the blockchain technology. This AWS blockchain service provides two different open source blockchain frameworks, namely Hyperledger Fabric [6], and Ethereum [7]. While the Ethereum offering on AWS is fully

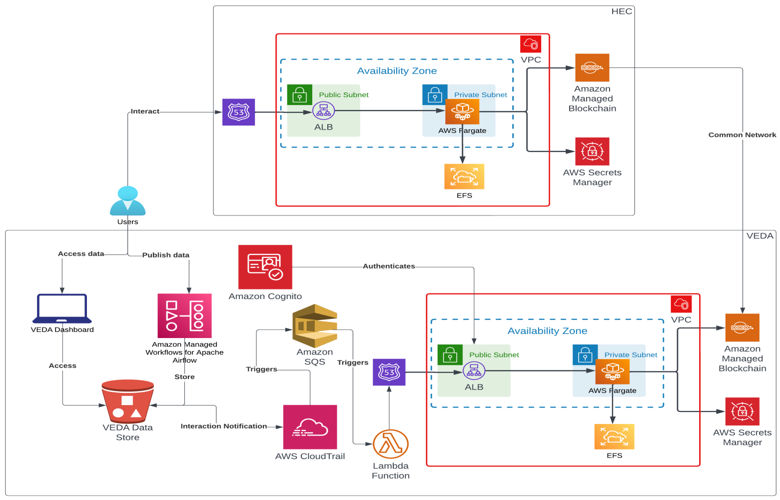


Fig. 2. OSBC implementation architecture for the VEDA and HEC use case

public, Hyperledger Fabric is private, and member organizations can be proposed to the network. In the case of VEDA, all datasets are openly accessible, but contribution is limited to member organizations. As such, Hyperledger Fabric was better suited for our use case.

We designed a workflow utilizing the VEDA data store and this blockchain service provided by AWS. Fig. 2 demonstrates the architecture of the implementation. VEDA and HEC both host their own instances of blockchain in AWS but have a shared network. Each hosted blockchain have 2 peer nodes. Peer nodes store the local copy of the ledger, run the chaincodes, and endorse transactions [8]. The blockchain services are hosted in their own virtual private network. The blockchain service is accessed by other applications and users via a Elastic Container Service (ECS) Fargate task. Each of these tasks hosts a representational state transfer (REST) application programming interface (API). The ECS is paired with an application load balancer (ALB) to maintain availability and scalability. Any requests made to the REST API is authenticated using VEDA auth backed by AWS Cognito. Furthermore, all interactions to the VEDA data store is tracked using AWS CloudTrail. The logs from CloudTrail are sent to AWS Simple Queue Service (SQS). Each entry to the SQS triggers the AWS lambda function which then makes API requests to the REST API for registering any interactions with the VEDA data store.

3.2. Use case

In this section, we discuss different scenarios of interactions with HEC's store and the VEDA data store.

Use Case 0 and Use Case 1 in Table 1 demonstrate how OSBC ensures data provenance tracking and promotes transparency. Users of HEC copied Nitrogen Dioxide observation data from VEDA, establishing HEC as the authoritative source. The OSBC tracks this information, granting scientists and users access to the authoritative data from VEDA or HEC. Data from other sources is not considered authoritative.

In Use Case 2, the OSBC facilitates collaboration, reproducibility, and accountability. HEC utilized the copied monthly Nitrogen Dioxide data to create a new dataset showing monthly differences, published to VEDA. Scientists provide citations for the original data, and the OSBC tracks linkages between the difference files and originals, enabling future reproduction.

In Use Case 3, the OSBC enables users to verify data authoritativeness and prevents duplication in the data store. Users receive an error when attempting to publish a file already present in either the VEDA or HEC data store. Only authorized copies, as described in Use Case 0, are allowed as duplicates.

Data in VEDA and HEC use cases is publicly readable, but data write access is restricted to VEDA and HEC members. Trusted organizations can join the network if VEDA and HEC agree to grant them access to the data stores, facilitated by the OSBC (Use Case 4).

Additionally, the architecture described in Section 3.1 automatically captures any interactions with VEDA data store in the OSBC. All the information stored in OSBC is tamper proof, and transparent.

4. CONCLUSION

Blockchain technology offers promising solutions for addressing challenges in open science, promoting transparency, reproducibility, and accountability. It enables the decentralized and secure sharing of authoritative data, traceability of scientific resources, and proper attribution. Blockchain-based

Use Case 0	HEC Copies data from VEDA to become authoritative distributors of the data
Description	HEC copies data from VEDA, recording the operation in the OSBC. The data is copied to HEC's data store, establishing a link between HEC and VEDA blocks for the same data. This maintains authoritative and non-authoritative linkages within the network.
Use Case 1	Differentiate between authoritative and non-authoritative data
Description	Data is authoritative if it was downloaded/accessed through: • VEDA data store or HEC data store Data is not authoritative if it was downloaded/accessed through: • Any other sources that are not VEDA or HEC • Any modified version of the same data not hosted by VEDA or official copy from HEC
Use Case 2	Citation is enabled to support open science
Description	Derived products can be cited with the original source when pushed to the network, stored in the chain. However, citation alone does not authenticate the derived product. To be authoritative, it must first be hosted by VEDA (Use Case 1). The chain preserves the dataset's evolution for easy viewing.
Use Case 3	Re-uploads to the network are not permitted
Description	Newly pushed data undergoes a hash check against existing datasets to ensure its uniqueness. If the hash already exists, the operation is rejected, maintaining data authoritativeness and providing a validation mechanism.
Use Case 4	Adding a new organization to the network
Description	New organizations can be "recommended" to the network but are not added directly. In our case, VEDA and HEC both need to be in agreement for any additional organization to be allowed to access the network.

Table 1. Use Cases

incentives foster collaboration and efficient research practices. However, successful adoption requires policy frameworks, data management approaches, cost considerations, and cultural shifts. The presented case study demonstrates practical implementation, contributing insights for further research. Blockchain has the potential to revolutionize open science, enhancing collaboration, and benefiting researchers and society as a whole. However, we recognize that full adoption of OSBC will take a long time as policies around data citations, and accesses are needed. There is also an additional cost of maintenance and development resources for blockchain. Furthermore, decentralized data storage and distribution has yet to be explored to enable proper open science. Continued research is essential to unlock blockchain's full potential for advancing open science.

5. REFERENCES

[1] Rahul Ramachandran, Kaylin Bugbee, and Kevin Murphy, "From open data to open science," *Earth and Space Science*, vol. 8, no. 5, pp. e2020EA001562, 2021, e2020EA001562 2020EA001562.

[2] Stephan Leible, Steffen Schlager, Moritz Schubotz, and Bela Gipp, "A review on blockchain technology and blockchain projects fostering open science," *Frontiers in Blockchain*, vol. 2, 2019.

[3] "Blockchain on amazon web services (aws)," https://aws.amazon.com/blockchain/, Accessed: 2023-06-02.

[4] "Visualization, exploration, and data analysis (veda) project," https://www.earthdata.nasa.gov/esds/veda, Accessed: 2023-06-02.

[5] "Nasa high-end computing," https://hec.nasa.gov/, Accessed: 2023-06-04.

[6] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, Srinivasan Muralidharan, Chet Murthy, Binh Nguyen, Manish Sethi, Gari Singh, Keith Smith, Alessandro Sorniotti, Chrysoula Stathakopoulou, Marko Vukolić, Sharon Weed Cocco, and Jason Yellick, "Hyperledger fabric: A distributed operating system for permissioned blockchains," in *Proceedings of the Thirteenth EuroSys Conference*, New York, NY, USA, 2018, EuroSys '18, Association for Computing Machinery.

[7] Gavin Wood, "Ethereum: A secure decentralised generalised transaction ledger," .

[8] "Hyperledger fabric peer nodes," https://hyperledger-fabric.readthedocs.io/en/latest/peers/peers.html, Accessed: 2023-06-04.