Enhancing Integrity and Confidentiality for Secure Distributed Data Sharing Using Open Science Chain

S. Sivagnanam, S. Sakai, K. Lin, F. Garzon, S. Yeu
UC San Diego, SDSC, La Jolla, CA;
{sivagnan, ssakai, klin, ofgarzon, syeu}@sdsc.edu

Motivation

As scientific advancement is an iterative process that builds on top of current and past research, it is critical to maintain data and research integrity.

- How can we ensure research integrity at scale, i.e., among different research groups sharing and working on the same datasets across the world?
- Is it possible to enable quick integrity verification of the exact data sets that were used for a particular published research?
- Can we check the provenance of the metadata used in the research?
- Can we ensure integrity of large collections of data such as imaging data.

Open Science Chain (OSC)

https://www.opensciencechain.org

Open Science Chain is a CI platform that

- utilizes blockchain technologies to securely store information about scientific data including its verification information.
- enables researchers to independently validate the authenticity of datasets, track and view the provenance of the data in an efficient manner.

Features of OSC

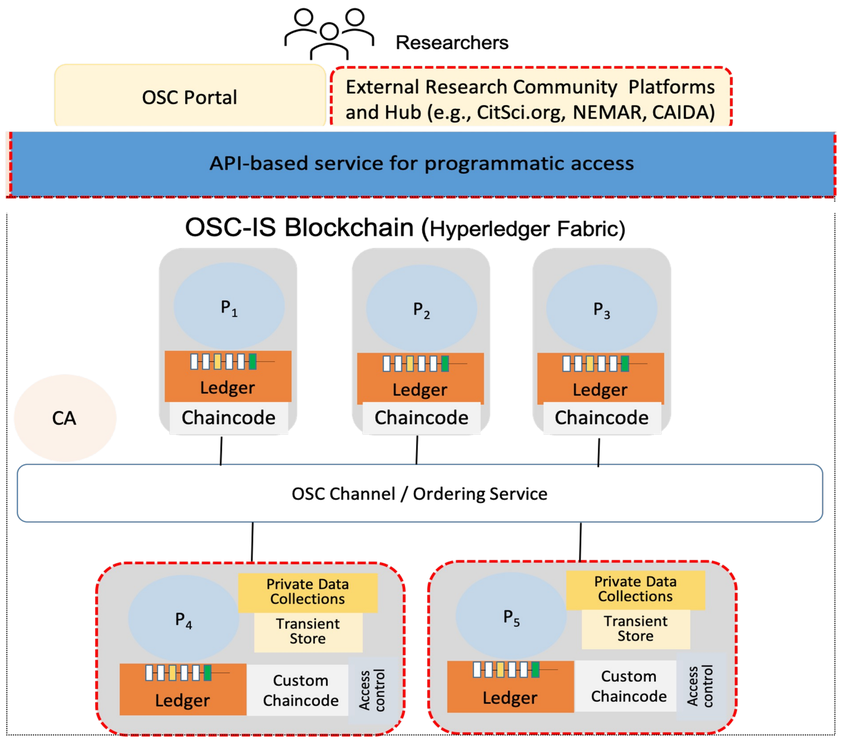
Protecting Integrity and Provenance of Research Data: OSC utilizes Hyperledger Fabric framework to securely store information about scientific data, including its provenance,.

Proof of Existence: The platform provides immutable proof of the existence of research datasets at a given point in time by storing unique identifiers of the data and ownership information on the blockchain

Provenance Tracking: OSC promotes transparency and traceability of research data by tracking and storing all changes made to data on the blockchain.

Data Verification & Validation: Researchers can independently verify the authenticity of scientific data using information stored in the OSC blockchain. They can search, view, and validate scientific datasets, including lineage information

Research Workflows: OSC allows researchers to create "research workflows" that link entries in the OSC blockchain with other repositories like GitHub, documenting an auditable record of the data workflow process behind research hypotheses



OSC – Integrity Services (OSC-IS)

API based services to

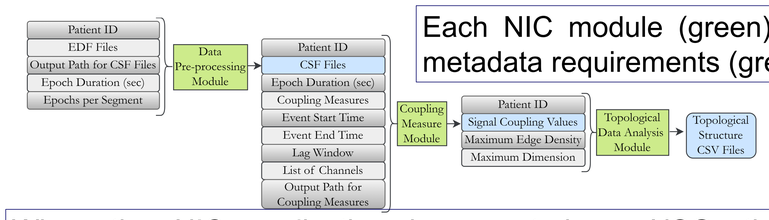
- provide support for metadata formats and standards adopted by disciplines and projects
- segregate and encrypt the metadata and provenance information of private datasets in the blockchain to allow only authorized entities to access the information
- provide controls for restricting access to information stored in the blockchain

Implementation details:

- Chaincode manages and enforces permissions based on group membership and user actions.Rules for transaction submission defined in chaincode including validation checks
- Data template (schemas) using JSON schema conforming to a meta schema.
- Private data in transient database and only hash shared publicly
- Developed a simple REST to translates HTTP requests to chaincode invocation
- Developing test suites with unit tests and wiki/documentation

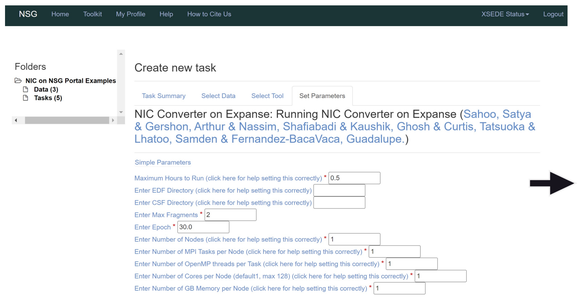
Use Case 1: Neuroscience Gateway integration with Open Science Chain

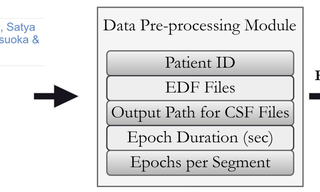
We store the standardized metadata information associated with the data generated by the Neuro-Integrative Connectivity (NIC) tool, which is one of the software application available within Neuroscience Gateway (NSG) framework, into the OSC to study the reuse and reproducibility of the application.



Each NIC module (green) has its own distinct set of metadata requirements (grey) to generate output (blue)

When the NIC application is executed on NSG, the metadata elements along with the information about the job execution is stored on OSC's blockchain. Any subsequent changes to the pipeline execution can be tracked and audited using the blockchain.





Execution on NSG → CSF Files

Information about Job execution → OSC OPEN SCIENCE CHAIN

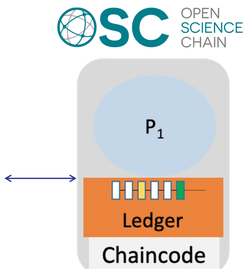
Data and Metadata from NIC execution on NSG stored on OSC.

Use Case 2: CitSci integration with Open Science Chain

Utah Water Watch project within CitSci framework has over 2000 data entries with nearly 20-30 data fields per entry. The data is contributed by over 200 citizen scientists which gets updated and needs to be tracked. We store the entries from Utah Water Watch project into OSC to study the reuse and reproducibility of the application as well as track the provenance changes associated with the data entries

Create Group/Users | Create/Update Schema | Create/submit/ update artifacts

API - Curl



Publications

[1] Sivagnanam, S., et al. Towards building a trustworthy pipeline integrating Neuroscience Gateway and Open Science Chain. Database (2024) Vol. 2024: article ID baae023; DOI: https://doi.org/10.1093/database/baae023 (accepted)

[2] Sivagnanam, S., Nandigam, V. and Lin, K., 2019. Introducing the Open Science Chain: Protecting Integrity and Provenance of Research Data. In Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning) (pp. 1-5)

https://www.opensciencechain.org
Contact : sivagnan@sdsc.edu

This research is supported by NSF grants –, 1840218, 2114202.