Towards building a Fault Tolerant and Secure Open Science Chain

Manu Shantharam, Scott Sakai, Kai Lin, Subhashini Sivagnanam
San Diego Supercomputer Center, University of California San Diego
(mshantharam, ssakai , klin, sivagnan) @sdsc.edu

SDSC SAN DIEGO SUPERCOMPUTER CENTER

Open Science Chain (OSC) attempts to address known credibility and reproducibility issue in scientific research. [1]

Many funding agencies now require that the data be made available post research phase in order to increase confidence and trust in the research work. [2,3]

Open Science Chain Background

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. Open Science Chain (OSC) attempts to answer the following questions:

- 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 verification of the exact data sets that were used for a particular published research?
- Can we check the provenance of the data used in the research?
- Can we ensure integrity of large collections of data such as imaging data.

What is Open Science Chain?

http://www.opensciencechain.org

Open Science Chain is a CI platform that

- utilizes Hyperledger Fabric (HLF) 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.

Open Science Chain consists of

1. **OSC blockchain** : Consists of three peers (with ledgers), three orderers in raft configuration, security module and a certificate authority. Running three peers and orderers on independent hosts improve the fault tolerance of the system.
2. **OSC Portal**: Client web portal that uses CILogon for authentication and provides easy to use interfaces to contribute metadata information related to artifacts and workflows used in publication.

OSC Resources

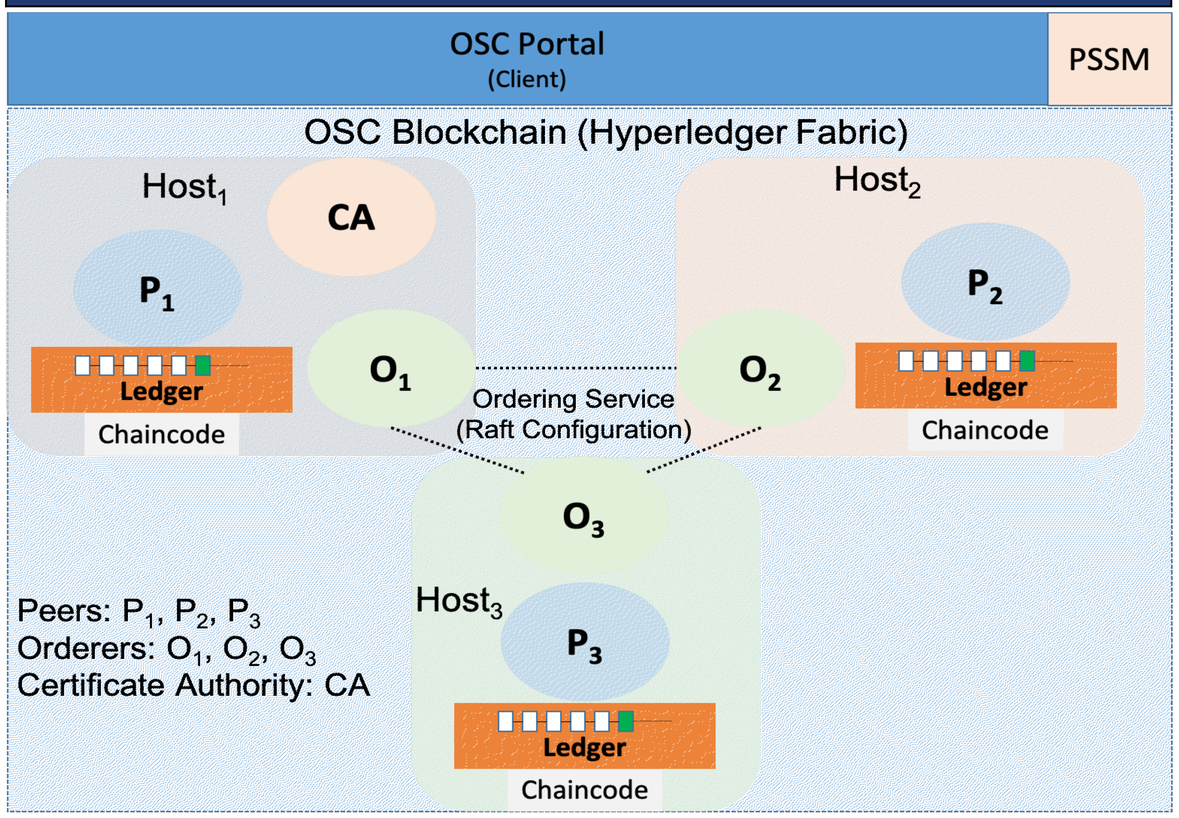
Contact

- info@opensciencechain.org
- Twitter :@OpenSciChain

Websites

- http://www.opensciencechain.org
- https://portal.opensciencechain.sdsc.edu/

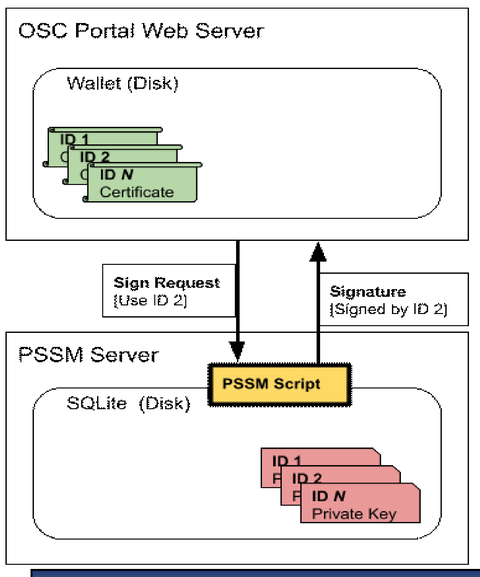
OSC Architecture



Enhanced Security

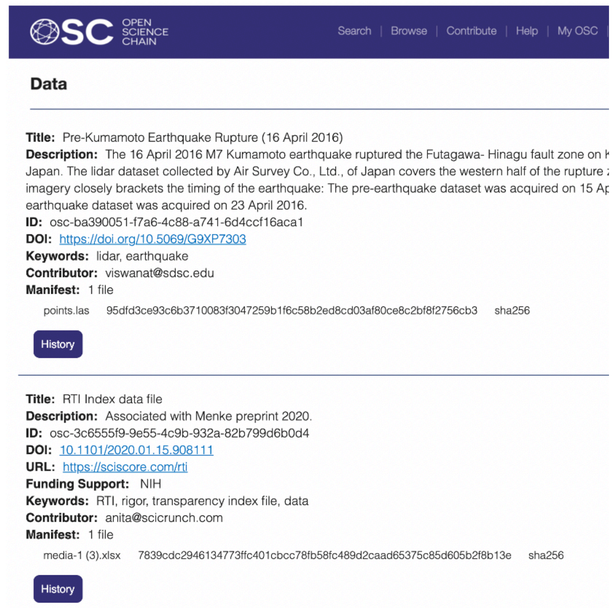
The OSC Portal creates an identity (with (private key and certificate) in OSC's HLF for each identity registering through the OSC Portal. These are stored locally on the host running the portal. Following are new security enhancements implemented in OSC:

- OSC Portal is run within a docker container
- Two separate identities for performing read and write operations to the blockchain instead of a single "god" admin identity.
- Portal Software Security Module (PSSM) to store private keys of identities



OSC PSSM module

- Similar to software HSM
- Keys are stored outside the reach of OSC Portal (Client) reducing the exposure of private keys
- Interactions between the portal and PSSM is using a REST interface and requires mutual TLS authentication
- Implemented in perl as a CGI script



OSC Portal

https://portal.opensciencechain.sdsc.edu/

Using Open Science Chain portal, researchers can

- Contribute metadata and verification information related to their scientific datasets
- Update the metadata and verification information as the dataset changes
- Refer to a specific version of an existing dataset
- Search and view datasets and obtain verification information
- Independently verify and validate other scientific datasets
- Create workflows linking multiple sources of data and computational code [4]

Future Work

- Track evolving datasets (change notifications)
- Ability to provide feedback to contributors
- Show the success rate of using the dataset on OSC
- Client tools for programmatic access from compute resources

REFERENCES

[1] Sivagnanam. S, Nandigam.V, and Lin.K. "Introducing the Open Science Chain: Protecting Integrity and Provenance of Research Data." Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines. ACM, 2019
[2] 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
[3] NSF. 2011 (accessed April 11, 2019). Digital Research Data Sharing and Management.
[4] V. Nandigam, K. Lin, M. Shantharam, S. Sakai, and S. Sivagnanam. 2020. Research Workflows - Towards reproducible science via detailed provenance tracking in Open Science Chain. In Practice and Experience in Advanced Research Computing (PEARC '20). Association for Computing Machinery, New York, NY, USA, 484–486. DOI:https://doi.org/10.1145/3311790.3399619

This research is funded by the NSF: 1840218