Artifact

Eduardo Oliveira

February 27, 2025

1 Open Science Platform

1.1 Overview

The proposed decentralized model is designed to facilitate seamless sharing, collaboration, and validation of research data, addressing key challenges in reproducibility and transparency. Traditional centralized systems often suffer from data silos, lack of verifiability, and risks of manipulation, which hinder open and trustworthy scientific practices. By leveraging blockchain technology and IPFS-based storage solutions, this model enables researchers to securely share their data while maintaining control over its metadata and provenance.

The Open Science platform empowers researchers and the broader scientific community by providing a secure, transparent, traceable, and tamper-proof environment for sharing project artifacts and data. It achieves this through a decentralized architecture built upon blockchain, IPFS, and smart contracts, ensuring the integrity and reliability of shared information. Validation mechanisms verify user input at the point of upload, preventing unauthorized modifications and enhancing trust in scientific outputs. This approach supports Open Science principles by promoting verifiable, long-term access to research artifacts without reliance on centralized intermediaries.

2 Introduction

This chapter presents the design and implementation of the Open Science Platform, a decentralized system integrating blockchain, IPFS, and smart contracts to enhance research reproducibility. The platform facilitates transparent and verifiable research artifact management by leveraging decentralized storage and immutable records while incorporating off-chain services to provide indexing, metadata extraction, and automated workflow execution.

3 Technology Stack

The Open Science Platform is developed using a hybrid architecture that combines decentralized and off-chain technologies to ensure secure, traceable, and

efficient data management.

3.1 Decentralized Components

- **Hyperledger Iroha v1 Blockchain:** Acts as the core infrastructure for managing user and project accounts, recording transactions, and enforcing business rules via smart contracts to ensure secure and transparent data exchange.
- **InterPlanetary File System (IPFS):** Provides decentralized, tamper-proof storage for research artifacts and metadata, ensuring persistent and verifiable access to shared data.

3.2 Off-Chain Components

- **Jupyter Notebooks (Python):** Serves as the front-end interface, automating and displaying execution steps related to research workflows.
- **Apache Tika:** Extracts metadata from uploaded files, enhancing artifact organization and searchability.
- **Whoosh:** Facilitates efficient indexing and keyword-based search for stored artifacts.

4 Platform Operations

The platform encompasses multiple core operations that govern user interaction and data management.

4.1 User Self-Enrollment

Users enroll in the platform by generating public/private keys conforming to ED25519 and SHA-3 standards. During registration, they provide identity information (e.g., name, institution, email, ORCID, role). The user's metadata is stored as a JSON file in IPFS, with the corresponding Content Identifier (CID) recorded on the blockchain.

4.2 Project Registration

Registered users can create projects by providing a descriptive name, an abstract, relevant keywords, start and end dates, funding agency, and location. A blockchain account is automatically assigned to the project, establishing a bi-directional link between the project owner and the project metadata stored on IPFS.

4.3 Artifact Management

4.3.1 File Upload

Users upload research artifacts such as papers, datasets, and images. Each file is stored on IPFS, generating a unique CID that ensures traceability.

4.3.2 Metadata Extraction and Storage

Simultaneously with the upload process, Apache Tika extracts metadata from the file. The extracted metadata is formatted in JSON, stored on IPFS, and its CID is linked to the corresponding project account on the blockchain.

4.3.3 Indexing and Search

The system indexes both files and metadata to enable efficient search functionality. Users can search for artifacts using keywords, with results displaying metadata and retrieval details.

4.4 Verification and Access

4.4.1 File Validation

To ensure data integrity, the platform verifies whether the CID of a stored file on IPFS matches the CID recorded on the blockchain. Any mismatch indicates tampering or corruption.

4.4.2 File Download

Validated files can be retrieved from IPFS and downloaded to the local file system.

5 Security and Integrity Considerations

This section discusses security mechanisms ensuring the authenticity and integrity of stored research artifacts. It covers cryptographic hashing, blockchain immutability, and IPFS redundancy, as well as potential attack vectors and mitigation strategies.

6 Conclusion

This chapter has detailed the technological underpinnings and operational workflows of the Open Science Platform. By leveraging blockchain, IPFS, and smart contracts alongside off-chain indexing and metadata extraction, the platform enhances research reproducibility through immutable data storage, verifiable metadata, and automated validation mechanisms.

6.1 Technology Stack

The Open Science platform is built upon a robust technical foundation, comprising:

- Hyperledger Iroha v1 Blockchain: The core infrastructure for account management and transaction recording and business rules enforcement through Smart Contracts ensuring secure and transparent data exchange.
- IPFS (InterPlanetary File System): The decentralized storage for project artifacts and metadata, guaranteeing tamper-proof and persistent access to shared information.

Aside from the decentralized technologies above, the platform also relies on the following off-chain, centralized components:

- Jupyter Notebooks in Python: The front-end interface of the platform leverages Jupyter Notebooks in Python to automate and display the execution steps of the activities in the platform.
- Apache Tika: Utilized for extracting file metadata, enhancing the platform's ability to manage and describe artifact content.
- Woosh: For efficient indexing and search capabilities for artifacts stored on the platform.

6.2 Operations

The Open Science platform is comprised of the following operations:

- User self-enrollment: Any user can self-enroll in the platform, with only a set of public/private keys conformant with standard ED25519 and SHA-3 as a requirement. The user must provide identity information such as full name, institution, email, ORCID, and role (e.g., author, publisher, reviewer). A JSON formatted representation of the user metadata is stored on IPFS, and the generated CID (Content Identifier) is stored in the blockchain of the user account.
- Project registering: Once enrolled, a user can register a project by providing a descriptive name for the project, an abstract summarizing the scope and goals of the project, keywords related to the project, start and end dates, funding agency, and location. The system automatically assigns an account in the blockchain for the project and links the user as the project owner bi-directionally. A JSON formatted representation of the project metadata is stored on IPFS, and the generated CID (Content Identifier) is stored in the blockchain of the project account.

- File upload: Users can upload artifacts such as papers, reports, images, datasets, etc., from their local machine to the platform. These files are stored securely on IPFS. A unique identifier (CID - Content Identifier) is generated for each artifact uploaded, which is used to track data provenance.
- Metadata extraction: In tandem with the upload process, metadata information from each uploaded file is extracted.
- Metadata upload: A JSON formatted representation of the metadata extracted from a file is uploaded on IPFS. The generated CID (Content Identifier) is stored in the blockchain of the project account.
- File indexing: The system indexes the files and their corresponding metadata, enabling efficient search functionality for users.
- Keyword search: Any user can perform searches based on keywords. Positive occurrences are displayed along with metadata information from the files.
- File validation: The file validation is performed. A file is considered valid if the CID in IPFS and the CID stored on the blockchain match exactly.
- File download: The valid file is downloaded to the local file system.

6.3 Data Model

The data model that supports the platform is comprised of two main classes User and Project. The User class contains attributes for user identity information, while the Project class contains attributes for project metadata. A many-to-many relationship exists between Users and Projects where, a single user can be associated with multiple projects.

To describe the attributes of each entity in the data model, three main ontologies were considered: FOAF (Friend of a Friend), Dublin Core and Schema.org. These standard vocabularies provide a common language for describing metadata information and can potentially easy the integration with other systems adopting W3C standards for semantic Web, like knowledge graphs, for instance.

6.4 Entity-relationship model

6.5 Blockchain Operations

On the context of blockchains, smart contracts are verifiable piece of code

6.6 Benefits

The Open Science platform offers numerous benefits for researchers and members of the scientific community, including:

- Secure data sharing: By utilizing blockchain technology and IPFS, the platform ensures tamper-proof data exchange.
- Transparent data management: The use of smart contracts and decentralized storage guarantees transparency in data access and modification history.
- Collaborative research environment: The platform enables researchers to collaborate on projects, share artifacts and results, and track progress.

6.7 Challenges

The Open Science platform faces several challenges, including:

- Scalability: As the number of users increases, the platform needs to be able to handle a growing amount of data and transactions efficiently.
- Interoperability: Ensuring seamless integration with existing research platforms and tools is crucial for widespread adoption.
- User Adoption: Educating researchers about the benefits of decentralized technologies and the Open Science platform can be an uphill battle.

6.8 Future Work

The Open Science platform has several areas for future development, including:

- Integration with existing research platforms: Collaborations with established research platforms to expand the platform's reach and user base.
- Enhanced security measures: Implementing additional security protocols to protect against potential threats and maintain the integrity of shared information.
- User interface improvements: Enhancing the web interface to make it more user-friendly and accessible for researchers from diverse backgrounds.

7 Conclusion

The Open Science platform is a comprehensive solution for secure, transparent, traceable, and tamper-proof data sharing and collaboration. By leveraging decentralized technologies, the platform empowers researchers to share project artifacts and data in a reliable and trustworthy manner.