# Artifact

Eduardo Oliveira

February 21, 2025

## 1 Open Science Platform

### 1.1 Overview

The Open Science platform aims to empower researchers and members of the scientific community by providing a secure, transparent, traceable and tamper-proof environment for sharing project artifacts and data. Building on this objective, the platform leverages decentralized technologies to ensure the integrity and reliability of shared information.

### 1.2 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.

## 1.3 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.

## 1.4 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.

## 1.5 Entity-relationship model

User Accounts Table

| **Attribute** | **Description** heightfoaf:name |
|---|---|
| Researcher's full name foaf:mbox | Researcher's email foaf:organization |
| Affiliated institution schema:identifier | Unique identifier (e.g., ORCID) foaf:h |
| Online account details (ID, role, public key) schema:linked$_{p}roject$ | Associated research project ID height |

Table 1: User Account Attributes

Research Projects Table

| **Attribute** | **Description** heightschema:identifier |
|---|---|
| Unique project identifier schema:publicKey | Cryptographic key schema:name |
| Project name dc:abstract | Research abstract schema:keywords |
| Keywords describing the project schema:startDate | Start date of the project schema:endDate |
| End date of the project schema:funding | Funding organization details schema:location |
| Project location schema:metadataCID | IPFS metadata content ID schema:linked$_{u}ser$ |
| Associated researcher ID height | |

Table 2: Research Project Attributes

## 1.6 Blockchain Operations

## 1.7 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.

### 1.8 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.

### 1.9 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.

## 2 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.