
d-OSP – DECENTRALIZED OPEN SCIENCE PLATFORM



A DECENTRALIZED INFRASTRUCTURE FOR
TRANSPARENT AND REPRODUCIBLE RESEARCH

MASTER'S CANDIDATE:
EDUARDO COSTA DE OLIVEIRA

REGIONAL UNIVERSITY OF NORTHWESTERN RIO GRANDE DO SUL

MASTER'S DISSERTATION

SUPERVISOR:
DR. RAFAEL ZANCAN FRANTZ

JOINT SUPERVISOR:
DR. THIAGO GOMES HECK



APRIL, 2025

First published in April 2025 by
Applied Computing Research Group - GCA
Department of Exact Sciences and Engineering
Rua Lulu Ilgenfritz, 480 - São Geraldo
Ijuí, 98700-000, Brazil.

Copyright © MMXII Applied Computing Research Group
<http://www.gca.unijui.edu.br>
gca@unijui.edu.br

In keeping with the traditional purpose of furthering science, education and research, it is the policy of the publisher, whenever possible, to permit non-commercial use and redistribution of the information contained in the documents whose copyright they own. You however are *not allowed* to take money for the distribution or use of these results except for a nominal charge for photocopying, sending copies, or whichever means you use redistribute them. The results in this document have been tested carefully, but they are not guaranteed for any particular purpose. The publisher or the holder of the copyright do not offer any warranties or representations, nor do they accept any liabilities with respect to them.

Support: This research is partially funded in Brazil by the Coordination for the Improvement of Higher Education Personnel (CAPES) and the National Council for Scientific and Technological Development (CNPq) through the following projects: 309425/2023-9 and 402915/2023-2.

Regional University of Northwestern Rio Grande do Sul

The committee in charge of evaluating the dissertation entitled “ d-OSP – Decentralized Open Science Platform: A Decentralized Infrastructure for Transparent and Reproducible Research”, presented by Master’s Candidate:

Eduardo Costa de Oliveira in the final exam for the degree of Master’s Degree in Mathematical and Computational Modeling, hereby provides recommendations and requires adjustments so that the final version has to be presented within _____ days.

Dr. Rafael Zancan Frantz

UNIJUÍ

(Supervisor)

Dr. Thiago Gomes Heck

UNIJUÍ

(Joint Supervisor)

Dr. Fabio Paulo Basso

UNIPAMPA

Dr. Sandro Sawicki

UNIJUÍ

I dedicate this work to my beloved wife and sons, whose unwavering support, patience, and encouragement sustained me throughout this challenging and rewarding journey. Their love was a constant source of strength and inspiration.

Contents

Acronyms Index

API - Application Program Interface

BCT - Blockchain Technologies

BFT - Byzantine Fault Tolerance

CODATA - Committee on Data of the International Science Council

dApp - Distributed Application

DLT - Digital Ledger Technologies

ECDSA - Elliptic Curve Digital Signature Algorithm

FAIR - Findable, Accessible, Interoperable, and Reusable

FOAF - Friend of a Friend

ER - Entity-Relationship

EVM - Ethernet Virtual Machine

GRPC - High performance Remote Procedure Call (RPC)

HTTP - Hyper Text Transfer Protocol

IPFS - InterPlanetary File System

JSON - JavaScript Object Notation

JSON-LD - JSON Linked Data

LEARN - Leveraging European Research Data Network

OpenAIRE - Open Access Infrastructure for Research in Europe

P2P - Peer to Peer

PBFT - Practical Bizantine Fault Tolerance

PoS - Proof of Stake

PoW - Proof of Work

RDA - Research Data Alliance

RDM - Research Data Management

RSD - Research Data Storage

RSA - Rivest Shamir Addleman

WSV - World State View

YAC - Yet Another Consensus

Acknowledgements

Iwould like to acknowledge the important contributions of my advisor, Dr. Rafael Z. Frantz, whose guidance, encouragement, and steady mentorship were fundamental throughout this research. I am equally grateful to my co-advisor, Dr. Thiago Heck, for his thoughtful insights and persistent questioning, which helped sharpen the focus and depth of this work. I also extend my sincere thanks to Dr. Carlos Molina Jiménez for his generous guidance and for continuously challenging my thinking in ways that significantly enriched this study.

Abstract

*The first principle is that you must not fool yourself
and you are the easiest person to fool.*

Richard Feynman, American Theoretical Physicist (1918 - 1988)

The scientific community is confronted with a critical challenge: the high rate of failed research reproductions, which calls into question the validity and reliability of scientific results. Although the causes of irreproducibility are diverse, issues surrounding data access, transparency, and integrity are particularly significant. The lack of effective mechanisms to support reproducibility contributes to reduced accuracy, inflated research costs, and declining trust in scientific progress. To address this challenge, we propose a Decentralized Open Science Platform that leverages decentralized technologies such as blockchain, smart contracts, and the InterPlanetary File System to advance the adoption of Open Science principles. This study examines how these technologies can improve data sharing and support reproducibility across various scientific fields. By developing an artifact, we evaluate both the benefits and the technical hurdles of integrating these technologies, aiming to foster a more transparent, reliable, and trustworthy scientific ecosystem. The findings from this research offer insight into how decentralized infrastructures can help close reproducibility gaps and reinforce the integrity of scientific inquiry.

Resumo

*O primeiro princípio é que você não deve enganar a si mesmo
e você é a pessoa mais fácil de enganar.*

Richard Feynman, Físico Teórico Americano (1918 – 1988)



A comunidade científica enfrenta um desafio crítico: a elevada taxa de falhas na reprodução de pesquisas, que compromete a validade e a confiabilidade dos resultados científicos. Embora as causas da irreprodutibilidade sejam diversas, questões relacionadas ao acesso, transparência e integridade dos dados desempenham um papel especialmente relevante. A ausência de mecanismos eficazes para sustentar a reprodutibilidade contribui para a redução da precisão, o aumento dos custos de pesquisa e a perda de confiança no progresso científico. Para enfrentar esse desafio, propomos uma Plataforma Descentralizada de Ciência Aberta que emprega tecnologias descentralizadas, como blockchain, contratos inteligentes e o InterPlanetary File System, com o objetivo de fortalecer a adoção dos princípios da Ciência Aberta. Este estudo investiga como essas tecnologias podem melhorar o compartilhamento de dados e apoiar a reprodutibilidade em diferentes áreas científicas. Por meio do desenvolvimento de um artefato, avaliamos tanto os benefícios quanto os desafios técnicos da integração dessas tecnologias, buscando promover um ecossistema científico mais transparente, confiável e robusto. Os resultados desta pesquisa oferecem subsídios sobre como infraestruturas descentralizadas podem contribuir para mitigar as lacunas de reprodutibilidade e reforçar a integridade da investigação científica.

Resumen

*El primer principio es que no debes engañarte a ti mismo
y tú eres la persona más fácil de engañar.*

Richard Feynman, Físico Teórico Estadounidense (1918 - 1988)

La comunidad científica se enfrenta a un desafío crítico: la alta tasa de fallos en la reproducción de investigaciones, lo que pone en duda la validez y la fiabilidad de los resultados científicos. Aunque las causas de la irreproducibilidad son diversas, los problemas relacionados con el acceso, la transparencia y la integridad de los datos desempeñan un papel especialmente importante. La falta de mecanismos eficaces para respaldar la reproducibilidad contribuye a una menor precisión, mayores costes de investigación y una pérdida de confianza en el avance científico. Para abordar este desafío, proponemos una Plataforma Descentralizada de Ciencia Abierta que emplea tecnologías descentralizadas como blockchain, contratos inteligentes y el InterPlanetary File System, con el fin de promover la adopción de los principios de la Ciencia Abierta. Este estudio examina cómo estas tecnologías pueden mejorar el intercambio de datos y fomentar la reproducibilidad en distintas disciplinas científicas. A través del desarrollo de un artefacto, evaluamos tanto los beneficios como los desafíos técnicos de integrar estas tecnologías, con el objetivo de contribuir a un ecosistema científico más transparente, fiable y sólido. Los resultados de esta investigación ofrecen perspectivas sobre cómo las infraestructuras descentralizadas pueden ayudar a cerrar las brechas de reproducibilidad y fortalecer la integridad de la investigación científica.

Chapter 1

Introduction

*Quis custodiet ipsos custodes?
Who will guard the guards themselves?*

Juvenal, Roman Poet (2 AC)

The purpose of this chapter is to introduce the context, motivation, and scope of the research presented in this dissertation. It outlines the reproducibility challenges faced in contemporary scientific practice and frames the relevance of decentralized technologies in addressing these issues. By situating the problem within the broader movement of Open Science, this chapter defines the research objectives and presents the foundational concepts that guide the development of the proposed solution. Provide a clear understanding of the problem domain, the research questions addressed, and the overall structure of the work. The chapter begins by establishing the research context in Section ?? . Section ?? presents the motivation behind the study, followed by the main and specific objectives outlined in Section ?? . The methodological approach adopted is described in Section ?? . Section ?? introduces the research hypothesis along with its rationale, while Section ?? summarizes the key contributions. Finally, the structure of the dissertation is presented in Section ?? , providing an overview of the subsequent chapters.

1.1 Research Context

The reproducibility crisis in science arises when researchers fail to replicate the results of a study, even when using the same methods and materials. This issue is prevalent across scientific disciplines where replication is essential for validating findings, advancing knowledge, and maintaining public trust in research.

Concerns about transparency and reproducibility have intensified in recent years. Studies indicate that a significant proportion of published research does not withstand rigorous scrutiny when retested, raising doubts about the reliability of scientific knowledge. This crisis results in wasted resources, misdirected efforts, and diminished confidence in research findings. Despite various efforts to improve reproducibility and better reporting guidelines such as PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) [?], ARRIVE (Animal Research Reporting of In Vivo Experiments) [?] and FAIR (The FAIR Guiding Principles for scientific data management and stewardship) [?], significant challenges remain in ensuring that research findings are consistent, transparent, and verifiable.

Several factors contribute to this challenge, including flawed study designs, data quality issues, unreliable measurement tools, and inconsistent research practices. Poorly designed studies may overlook critical variables, while inadequate measurement tools can produce misleading results. Furthermore, incomplete reporting or the omission of key methodological details can hinder independent verification of findings [?].

A promising approach to addressing reproducibility issues lies in the adoption of Open Science [?], both as a movement and as a methodological framework that fosters transparency, collaboration, and accessibility in research. Open Science promotes unrestricted access to research outputs, including publications, datasets, code, methodologies, and protocols, enabling verification and reuse while reducing barriers to scientific progress. By advocating for systematic, reproducible, and accessible research practices, Open Science enhances accountability and trust within the scientific community. However, despite its transformative potential, its current implementations encounter significant technical and structural limitations, particularly in scalability, consistency, and automation. Existing infrastructures often rely on centralized repositories and fragmented systems that lack transparency, are vulnerable to manipulation or demand extensive manual oversight to ensure verifiability.

To overcome these challenges, decentralized technologies offer a robust proposition that aligns with the principles of Open Science while addressing its technical constraints. Blockchain technology ensures the immutability and integrity of research records, preventing unauthorized modifications, tampering and traceability. IPFS (InterPlanetary File System) provides decentralized and persistent data storage, eliminating single points of failure and ensuring long-term accessibility of research outputs. Smart contracts facilitate the automated enforcement of predefined research-sharing policies reducing the need for intermediaries and fostering trust in collaborative environments. By integrating these decentralized technologies, it becomes possible to construct an infrastructure that not only upholds Open Science principles but also strengthens reproducibility by delivering secure, transparent and verifiable mechanisms for data sharing and validation across scientific disciplines.

Building upon these foundations, this dissertation proposes the development of a structured artifact designed to provide a transparent framework for documenting, validating and disseminating research outputs. This artifact will accommodate diverse materials, including reports, protocols, datasets, images and videos, ensuring secure and accessible record-keeping. By leveraging decentralized technologies, the proposed artifact targets the enhancement of methodological rigor, support independent verification of scientific findings and ultimately contribute to a more trustworthy and reproducible scientific ecosystem.

The scientific community has long recognized the importance of reproducibility in ensuring the validity and accuracy of research findings. Reproducibility refers to the ability of researchers to replicate another researcher's study or experiment, with minimal changes, to verify its results and conclusions [?]. However, despite widespread efforts to address this issue, failed reproductions continue to be a persistent problem across various scientific domains, which not only compromises the credibility of individual studies but also undermines our collective understanding of scientific phenomena.

Studies have shown that reproducibility rates vary widely across disciplines, but even among those fields with a strong tradition of rigor, replication rates often fall short. For instance, a study published in the journal PLOS Biology found that only 22% of papers published in top-tier journals were replicable [?]. Similarly, a review of studies on reproducibility in Nature revealed that only 15% of studies reported high levels of reproducibility [?].

Several interrelated factors contribute to these poor rates of reproducibility. Methodological flaws, such as sampling biases or inadequate control

groups, often result in difficulties replicating findings [?]. Additionally, a lack of transparency in research practices, such as insufficient reporting on methods, materials, and results, hinders independent verification and replication efforts [?]. Poor data management practices further exacerbate these challenges, as issues like inadequate data documentation, lack of standardized metadata, and the absence of persistent identifiers make it difficult to retrieve, interpret, and reuse data [?]. The use of proprietary or inaccessible data formats, as well as data loss due to improper storage or insufficient version control, further limits reproducibility [?]. Moreover, limited resources, including small research teams or restricted funding, can prevent the rigorous experimentation necessary to ensure high reproducibility levels, particularly when researchers lack access to robust data infrastructure and long-term archival solutions [?].

The need for transparency and openness in scientific research has led to the development of the Open Science movement. Open Science is a broad initiative that advocates for the transparency, accessibility and collaboration of research processes and outputs. Its tenets include making research datasets, publications, code and methodologies publicly accessible, enabling independent verification and reuse of scientific work. Open Science encourages practices such as publishing raw data, sharing code and providing comprehensive methodological descriptions that allow others to reproduce and extend findings.

Despite the substantial contributions of Open Science, the current systems and tools supporting these practices remain fragmented and often lack the scalability and consistency required for widespread adoption. As such, significant challenges in ensuring reproducibility persist, particularly when dealing with large, complex datasets, proprietary information or sensitive research materials.

Given these challenges, the integration of decentralized technologies such as blockchain [?], IPFS [?] and smart contracts [?] presents a potential solution to address the limitations of current Open Science frameworks. Blockchain technology offers a robust mechanism for ensuring the immutability and integrity of records, making it easier to track data provenance and verify results over time. IPFS provides a scalable, decentralized infrastructure for storing large datasets, ensuring they remain accessible and tamper-proof. Smart contracts, meanwhile, automate the verification and validation process, making it more efficient and secure. Together, these technologies enable a more transparent and accessible research ecosystem that enhances the reproducibility of scientific findings across various domains.

1.2 Motivation

The scientific community faces a critical challenge in ensuring the integrity and validity of research findings. Reproducibility, the ability to independently verify and replicate the results of a study, is fundamental to scientific progress. However, studies have shown that a significant proportion of published research findings cannot be reliably reproduced, with some fields reporting irreproducibility rates exceeding 50% [?]. This crisis threatens the credibility of individual researchers, academic institutions, and entire disciplines, ultimately slowing the advancement of knowledge and reducing public trust in science.

Despite ongoing efforts to improve reproducibility, the problem persists due to systemic barriers embedded within traditional research practices. Conventional publication systems prioritize novelty over transparency, funding mechanisms rarely incentivize data sharing, and peer-review processes lack the infrastructure to systematically verify research integrity. Consequently, even well-intentioned researchers struggle to ensure their work is reproducible due to inadequate tools for transparent data management, persistent identifier tracking, and independent verification.

The Open Science movement has emerged as a response to these challenges, advocating for unrestricted access to research outputs, including publications, datasets, software, and methodologies. By fostering transparency, collaboration, and standardization, Open Science aims to improve reproducibility through practices such as open data sharing, pre-registered study protocols, and transparent peer review. However, while Open Science provides a conceptual foundation for improving reproducibility, its practical implementation remains hindered by the lack of scalable, automated solutions for verifying research integrity in complex, data-intensive environments.

This dissertation proposes a decentralized solution that integrates blockchain technology, the InterPlanetary File System (IPFS), and smart contracts into scientific workflows to address reproducibility challenges within the Open Science framework. Blockchain technology ensures the immutability and verifiability of research records, preventing data manipulation and guaranteeing long-term integrity. IPFS provides a decentralized and tamper-proof storage mechanism, facilitating efficient and secure access to research outputs. Smart contracts automate validation and verification processes, reducing reliance on manual intervention and minimizing errors.

By leveraging these decentralized technologies, this research aims to establish a more transparent, secure, and reproducible scientific ecosystem. Beyond addressing current reproducibility challenges, this work seeks to strengthen the broader Open Science framework by providing tools that enable data integrity, facilitate collaboration, and support independent verification across scientific disciplines.

1.3 Objectives

We present the objectives that have directed our research in the context of Open Science and the reproducibility of scientific research, which resulted in the contributions of this dissertation. This work explores the potential of distributed technologies, specifically blockchain, IPFS, and smart contracts, to improve research transparency, data integrity, and reproducibility.

1.3.1 Main objective

Develop a platform that integrates blockchain, IPFS, and smart contracts to enhance the reproducibility in alignment with Open Science principles.

1.3.2 Specific objectives

- Analyze the limitations of current research data management practices with respect to provenance and reproducibility;
- Investigate the extent to which blockchain and decentralized storage can support immutable, persistent, and verifiable research data;
- Design an architecture that integrates blockchain, IPFS, and smart contracts to enable transparent storage, access control, and provenance recording;
- Implement the proposed architecture in a functional platform that supports the secure publication and retrieval of datasets, metadata, and experiment records;
- Conduct experiments to evaluate the platform's behavior in scenarios involving the registration, retrieval, and validation of experiment data, including associated metadata;
- Compare the reproducibility, availability, and integrity of research workflows executed with and without the use of the platform;

1.4 Methodology

This study follows a structured approach grounded in the development and evaluation of a technological artifact designed to address challenges in scientific reproducibility, transparency, and data sharing. To ensure methodological rigor and practical relevance, we employ a research strategy that supports iterative refinement and empirical validation.

This research is guided by the Design Science Research (DSR) paradigm [?], which emphasizes the creation and assessment of artifacts as a means of generating knowledge. The objective is to design a decentralized application that facilitates scientific collaboration and improves the auditability and reproducibility of research outputs. By leveraging blockchain technology, decentralized file system, and smart contracts, the application enables researchers to securely share data, methods, and results in a transparent and verifiable manner.

The methodology consists of multiple phases. The first phase involves conceptual modeling, where the system's architecture, data structures, and user interfaces are defined. This conceptual groundwork guides the subsequent prototyping phase, during which functional prototypes are implemented and iteratively refined based on feedback and testing.

Following prototype refinement, the final system undergoes evaluation to assess its capacity to support reproducible and transparent research practices. This includes both functional validation and alignment with Open Science principles.

The DSR methodology guiding this work incorporates three essential principles. The first is innovation, through which the research introduces a novel solution to an identified problem. The second is improvement, focusing on advancing current practices in decentralized data management. The third is validation, ensuring that the artifact undergoes testing to demonstrate its effectiveness in meeting the research objectives.

1.5 Research Hypothesis and Rationale

The central hypothesis of this research is that combination of these decentralized technologies: blockchain, smart contracts, and the InterPlanetary File System (IPFS), can improve reproducibility in scientific research. This

hypothesis is grounded in the premise that many of the reproducibility challenges in contemporary science experiments stem from data management issues: lack of transparency in experimental protocols and study designs, restricted access to primary data, insufficient metadata, and the absence of reliable provenance tracking mechanisms.

To evaluate this hypothesis, a functional artifact will be developed and implemented, integrating blockchain, IPFS, and smart contracts within the d-OSP. The platform will support the registration, sharing, validation and provenance of research assets, such as datasets, code, and protocols and reports. The evaluation will proceed in two phases. First, a functional assessment will determine whether the system supports the reproducibility-oriented capabilities it claims to provide. Second, a scenario-based evaluation will examine how the artifact performs when used to register and verify simulated scientific workflows. These scenarios will reflect common reproducibility failure points and will be used to observe whether the system can mitigate or eliminate them.

Through this dual evaluation strategy, the research will demonstrate whether decentralized infrastructure can operationalize key principles of reproducible science, and to what extent such technologies can be practically implemented to support Open Science goals. Should the system succeed in these evaluations, the hypothesis will be considered supported.

1.6 Summary of Contributions

In this section, we outline the principal contributions of this dissertation and situate them in relation to our earlier published research.

- A decentralized infrastructure that supports the transparent sharing and verification of scientific data is presented. By integrating blockchain for immutable logging, IPFS for distributed storage, and smart contracts for automated governance, we facilitate the secure and auditable dissemination of research outputs. This technological configuration strengthens reproducibility by making both data and metadata from experimental procedures verifiable and tamper-resistant.
- The artifact's value is demonstrated through a structured validation process. By employing iterative cycles of design, implementation, and evaluation within a DSR framework, we explore how decentralized technologies can reduce variability in research outcomes. We test the artifact through simulated use cases that highlight its capacity to mitigate reproducibility issues in data-intensive domains.

- Build upon our earlier research, specifically the peer-reviewed publication *On the Use of Blockchain Technology to Improve the Reproducibility of Preclinical Research Experiments* [?], presented at the 25th International Conference on Enterprise Information Systems. That foundational work examined blockchain’s potential for improving scientific transparency. In this dissertation, we extend those insights by incorporating complementary technologies and refining their orchestration within Open Science infrastructures.

Taken together, our contributions advance the discourse on Open Science by providing a potential solution to current limitations in research data sharing. We demonstrate how decentralized architectures can uphold the principles of transparency, accountability, and accessibility offering a practical response to the ongoing reproducibility crisis in scientific research.

1.7 Dissertation Structure

This section provides an overview of the organization of this dissertation and the content of each chapter.

- **Chapter 1: Introduction** – Establishes the overall context of the research, articulating the underlying motivation and situating the study within the broader discourse on the application of decentralized technologies and Open Science principles to enhance reproducibility. It presents the adopted methodology, introduces the research hypothesis, and provides a summary of the main contributions. The chapter concludes with an overview of the dissertation’s structure.
- **Chapter 2: Background** – Presents the technical and conceptual foundations that support the research. It examines core concepts including decentralized and distributed systems, blockchain technologies, cryptography, smart contracts, IPFS, and the Open Science movement.
- **Chapter 3: Literature and Research Area Review** – Synthesizes relevant scholarly work, situating the research within ongoing debates and developments. It reviews previous initiatives that leveraged decentralized technologies to enhance reproducibility and establishes the conceptual grounding for the proposed model. .
- **Chapter 4: Proposed Decentralized Model** – Introduces the principal technical contribution of the dissertation: the d-OSP, a decentralized

infrastructure represented as an artifact designed to improve reproducibility in scientific research. It outlines the platform's architecture, data model, implementation details, and the rationale behind key design decisions. The chapter also presents an empirical evaluation of the platform's performance and functionality.

- **Chapter 5: Conclusions** – Summarizes the key findings of the research and reflects on their broader implications. It revisits the hypotheses in light of the presented evidence and discusses potential avenues for future research.
- **Appendices and References** – The appendices include supplementary materials such as implementation details, code excerpts, and technical documentation that complement the main chapters. The references section compiles all cited scholarly sources, upholding the standards of academic transparency and rigor.

Chapter 2

Background

Quis custodiet ipsos custodes? Who will guard the guards themselves?

Roman Poet Juvenal 2 AC

To support the development of the proposed Decentralized Open Science Platform, this chapter presents the technical and conceptual foundations underpinning the research. Section ?? examines decentralized and distributed systems, Section ?? introduces the main concepts of blockchain technologies, and Section ?? defines and contextualizes smart contracts. The InterPlanetary File System is discussed in Section ??, while the synergies between blockchain and IPFS are explored in Section ?. Lastly, the Open Science movement is examined in Section ?. Together, these elements form the foundational framework upon which the platform's architecture and governance mechanisms are built.

2.1 Decentralization and Distributed Systems

Decentralization and distributed systems are fundamental paradigms that redefine how data is processed, stored, and verified across various domains. Unlike traditional architectures that rely on a centralized entity to manage operations, decentralized systems distribute control among multiple participants [?]. This approach enhances fault tolerance, ensures system resilience, and mitigates risks associated with single points of failure. Distributed systems, in turn, rely on a network of interconnected nodes to collectively maintain and process data, enabling scalability and redundancy [? ?]. These principles underpin various modern technologies, including blockchain [?] and decentralized file storage networks [?], which eliminate reliance on centralized intermediaries and foster transparency and security.

2.1.1 The Building Blocks for Decentralization

Blockchain technology embodies decentralization by ensuring that no single entity controls data integrity and transaction validation. Transactions are recorded in a distributed ledger that is collectively maintained by network participants through consensus mechanisms [?]. Using cryptographic techniques [?] and game-theoretic incentives [?], the blockchain achieves trustless verification, preventing unauthorized modifications while ensuring transparency.

Similarly, the InterPlanetary File System (IPFS) uses distributed system principles to provide decentralized data storage [?]. Unlike conventional file storage, which relies on centralized servers, IPFS distributes files across a peer-to-peer network, addressing them based on their content rather than location. This approach not only ensures data persistence but also enhances accessibility by allowing multiple nodes to host and retrieve the same content. In contrast to blockchain, which primarily records transactions and state changes, IPFS enables efficient and scalable storage of large data objects, complementing blockchain's immutability with a robust storage layer.

Both blockchain and IPFS exemplify the synergy between decentralization and distributed computing. Blockchain secures and verifies data integrity, while IPFS ensures scalable, redundant storage, collectively forming the foundation for decentralized applications, provenance tracking, and secure data sharing.

2.1.2 Decentralized Data Management

Blockchain technology and the InterPlanetary File System (IPFS) represent complementary solutions in decentralized systems. Blockchain serves as a distributed ledger that ensures immutability, transparency, and security through cryptographic hashing and consensus mechanisms, excelling at maintaining an auditable record of transactions and verifying data integrity [?]. In contrast, IPFS is a peer-to-peer distributed file system that addresses data storage challenges by organizing and retrieving content based on its hash rather than its location, enhancing resilience and efficiency by enabling distributed data hosting across a network of nodes [?].

2.2 Blockchain

Blockchain is a decentralized, immutable ledger technology designed to facilitate secure and transparent transactions within distributed networks. Initially conceptualized for the Bitcoin blockchain [?], this technology has since evolved into a multi-purpose infrastructure featuring use cases in various domains, including finance, supply chain management, and digital identity verification.

The fundamental characteristics of blockchain, including decentralization, immutability, transparency, and security, make it a powerful tool for enhancing scientific reproducibility. By providing an auditable and tamper-proof record of research data and workflows, blockchain ensures data integrity and long-term verifiability. Its foundation in cryptographic principles and distributed systems enables it to support a robust infrastructure that enhances the reliability and transparency of scientific research.

2.2.1 Merkle Trees

The development of blockchain, however, did not occur in isolation. The concept of a cryptographically secured chain of blocks predates Bitcoin and draws from earlier research on distributed consensus and cryptographic techniques. A key component in blockchain structures is the Merkle tree, a concept introduced by Ralph Merkle in the 1980s [?]. These trees enable efficient data integrity verification by organizing hashes in a hierarchical structure, which is crucial for maintaining the integrity of blockchain data.

2.2.2 Secure Time Stamping

Building on these foundational cryptographic concepts, Stuart Haber and W. Scott Stornetta proposed a method for securely time-stamping digital documents in 1991 [?]. This innovation was significant because it prevented backdating and tampering, laying the groundwork for immutable records. In 1992, Haber, Stornetta, and Bayer further refined this approach by incorporating Merkle trees into their time-stamping system, thereby improving efficiency and strengthening security [?]. These advancements not only contributed to the development of blockchain but also highlighted the potential of decentralized, immutable ledgers for maintaining verifiable records.

2.2.3 Public, Private and Hybrid Blockchains

Blockchains can be categorized based on their access control and governance mechanisms into public, private, and hybrid models. Each type offers distinct advantages and trade-offs in terms of decentralization, security, scalability, and transparency.

- **Public Blockchains:** Allow unrestricted participation, enabling any user to join the network, validate transactions, and maintain a copy of the ledger. These blockchains emphasize decentralization and security, ensuring data integrity through cryptographic hashing and consensus mechanisms such as Proof of Work (PoW) and Proof of Stake (PoS) [?]. Prominent examples include *Bitcoin* and *Ethereum*, which are designed to function in open, trustless environments. A defining feature of public blockchains is their transparency, as all transactions are recorded on a publicly accessible ledger. This openness makes them well-suited for scenarios requiring verifiability and censorship resistance, though it typically incurs trade-offs in scalability and performance, as each node must process and store all transactions.
- **Private or Permissioned Blockchains:** Restrict participation to authorized entities, making them particularly suitable for enterprise and institutional applications. Frameworks such as Hyperledger Fabric and Hyperledger Iroha are designed for regulated environments where identity verification, compliance, and access control are critical considerations [?]. Since private blockchains operate under a governing authority that controls access and consensus rules, they offer enhanced

scalability and efficiency compared to public blockchains. Unlike PoW-based systems, permissioned blockchains often employ consensus mechanisms such as Practical Byzantine Fault Tolerance (PBFT) and Raft, which reduce computational overhead and transaction latency [?]. However, the trade-off is a reduction in decentralization, as control is concentrated within a predefined group of participants.

- **Hybrid Blockchains:** Integrate elements of both public and private blockchains, enabling organizations to leverage the benefits of decentralization while maintaining privacy for sensitive data. These architectures allow entities to manage a private ledger while selectively interacting with public networks for verification, auditability, or interoperability. Hybrid approaches are particularly valuable in sectors where confidentiality and transparency must coexist, such as supply chain management, financial services, and healthcare [?]. By connecting permissioned chains to public networks, hybrid models offer a balance between data privacy, operational efficiency, and the advantages of immutable public verification.

2.2.4 Mining and Block Validation

Mining is the process by which transactions are validated and added to a blockchain. In PoW-based systems, miners compete to solve cryptographic puzzles, while in PoS-based systems, validators are selected to propose and confirm blocks based on their stakes. Mining serves two key purposes: securing the network by making attacks computationally expensive and issuing new tokens as rewards. This incentive structure aligns participant behavior with the network's security goals [?]. For example, Bitcoin employs PoW mining, while Ethereum 2.0 uses PoS validation.

2.2.5 Consensus Mechanisms

Consensus mechanisms are fundamental to blockchain networks, ensuring agreement among distributed nodes without requiring centralized authority. These mechanisms validate transactions and maintain the integrity of the ledger, preventing issues such as double-spending and malicious attacks.

- **Proof of Work (PoW):** Requires participants, known as miners, to solve computationally intensive cryptographic puzzles. The first miner to

solve the puzzle earns the right to append a new block to the blockchain and receives a block reward. This mechanism ensures network security by making it economically and computationally infeasible to manipulate the ledger. However, its reliance on high energy consumption has raised sustainability concerns [? ?]. A difficulty adjustment algorithm modulates puzzle complexity to maintain a consistent block production rate in response to changes in total network computational power.

- **Proof of Stake (PoS):** Offers an energy-efficient alternative to PoW by replacing mining with a staking-based validation process. In PoS systems, participants, called validators, are selected to propose and attest to new blocks based on the amount of cryptocurrency they lock as collateral. This design reduces energy consumption while preserving network security and decentralization. Rather than expending computational effort, validators are incentivized to behave honestly through economic rewards and penalties. Contemporary PoS protocols often include enhancements such as dynamic staking models and slashing mechanisms, which penalize malicious behavior and reinforce protocol robustness [?].

2.2.6 Byzantine Fault Tolerance

Byzantine Fault Tolerance (BFT) refers to the ability of a distributed system to function correctly even when some nodes fail or behave maliciously [?]. Classical BFT algorithms, such as Practical Byzantine Fault Tolerance (PBFT) [?], require $2f + 1$ honest nodes out of $3f + 1$ total nodes to tolerate up to f Byzantine faults. While PBFT offers strong consistency and finality, it assumes a fixed set of validators, making it particularly suitable for permissioned blockchain environments.

In this work, we focus on the BFT-based consensus mechanism employed by Hyperledger Iroha: Yet Another Consensus (YAC).

YAC

Yet Another Consensus (YAC) is a lightweight BFT algorithm tailored for permissioned blockchain networks, where validators are pre-approved entities [?]. YAC emphasizes efficient communication and low-latency block validation, making it well-suited for enterprise-grade applications that demand both performance and reliability.

Consensus Algorithm for YAC

The YAC (Yet Another Consensus) algorithm ensures Byzantine Fault Tolerance (BFT) through a voting-based process, as follows:

Fault Tolerance

The network must satisfy:

$$n \geq 3f + 1$$

Where:

- n : Total number of nodes in the network.
- f : Maximum number of Byzantine (malicious or faulty) nodes tolerated.

This ensures that the network can tolerate up to f Byzantine nodes while still achieving consensus.

Supermajority Agreement For a block to be finalized, a supermajority of nodes must agree on its hash:

$$v > \frac{2n}{3}$$

Where:

- v : Number of votes for the block hash.
- n : Total number of nodes.

This condition guarantees both safety and liveness in consensus.

Voting Process

The voting process involves two phases:

- i. Proposal Phase: A leader node proposes a block hash.
- ii. Voting Phase: Nodes vote on the proposed hash, and votes are collected until the supermajority condition is met.

Finality Once a block achieves supermajority agreement, it is considered finalized and added to the blockchain. This ensures that all honest nodes accept the same block, preventing forks.

Consensus mechanisms, mining, and fault tolerance strategies are critical in ensuring blockchain security and functionality. The choice between PoW, PoS, and BFT-based approaches impacts the efficiency, decentralization, and security of blockchain networks. Furthermore, the distinction between public, private, and hybrid blockchains influences their applications, with public blockchains prioritizing trustlessness and private blockchains emphasizing control and efficiency. Understanding these foundational aspects enables the development of blockchain-based solutions tailored to the needs of Open Science and research reproducibility.

Hyperledger Iroha integrates YAC to ensure secure and robust consensus through a voting-based mechanism. This approach enables fault tolerance and streamlined consensus, aligning with the system's goals of scalability and operational efficiency.

2.2.7 Cryptography in Blockchain

Public key cryptography plays a fundamental role in securing blockchain networks by enabling secure transactions, identity verification, and data integrity without requiring a centralized authority [?]. It forms the foundation for digital signatures, key management, and encryption mechanisms that ensure trust and security in decentralized environments.

The Role of Public Key Cryptography in Blockchain

Blockchain networks rely on asymmetric cryptography, also known as public key cryptography, to authenticate and authorize transactions [?]. Each participant holds a pair of cryptographic keys: a public key, which functions as an address to receive transactions, and a private key, which is used to sign transactions and assert ownership. When a user initiates a transaction, they produce a digital signature using their private key. This signature enables other participants to verify the transaction's authenticity without exposing the private key itself [?].

This mechanism ensures that only the rightful owner of an asset can authorize its transfer, preventing fraud and unauthorized access [?]. Additionally, cryptographic hashing techniques complement public key cryptography by ensuring data integrity and linking transactions in an immutable ledger [?].

Commonly Used Cryptographic Ciphers and Standards

Several cryptographic ciphers and standards are widely used in blockchain implementations to provide strong security guarantees:

- **RSA (Rivest-Shamir-Adleman):** A traditional public key cryptosystem based on the difficulty of factoring large prime numbers [?]. While RSA is widely used in general cryptographic applications, its key sizes are relatively large compared to modern alternatives, making it less practical for blockchain applications.
- **Elliptic Curve Cryptography (ECC):** A more efficient asymmetric cryptography scheme that provides the same level of security as RSA but with significantly smaller key sizes [?]. This efficiency makes ECC the preferred choice for blockchain applications.
- **ECDSA (Elliptic Curve Digital Signature Algorithm):** A widely adopted digital signature scheme based on ECC, used in Bitcoin and Ethereum to secure transactions [?].
- **EdDSA (Edwards-curve Digital Signature Algorithm):** A modern alternative to ECDSA, known for its faster signature verification and improved security properties [?]. It is used in newer blockchain protocols like Monero.
- **X25519:** A secure key exchange protocol based on Curve25519, commonly used in cryptographic operations for secure communication in blockchain applications [?].
- **Ed25519 with SHA-3:** Hyperledger Iroha natively employs Ed25519, a variant of EdDSA, combined with SHA-3 for enhanced security. This cryptographic combination ensures efficient key pair generation, digital signatures, and verification processes tailored to Iroha's permissioned blockchain environment [?].

Elliptic Curve Cryptography (ECC) in Blockchain

Elliptic Curve Cryptography (ECC) is a type of public key cryptography that leverages the mathematical properties of elliptic curves over finite fields to provide strong security with smaller key sizes [?]. The security of ECC is based on the Elliptic Curve Discrete Logarithm Problem (ECDLP), which is computationally hard to solve [?].

In blockchain systems, ECC is primarily used for:

- **Digital Signatures:** ECC enables the creation of compact and secure digital signatures, such as those used in Bitcoin (ECDSA) and newer blockchain protocols (EdDSA and Ed25519) [? ?].

- **Key Pair Generation:** Blockchain wallets generate private-public key pairs using elliptic curves, ensuring that users can securely sign and verify transactions [?]. For example, Hyperledger Iroha specifically utilizes Ed25519 with SHA-3 for key pair generation, providing robust security and efficiency [?].
- **Scalability and Efficiency:** Due to its small key size and lower computational requirements, ECC allows blockchain networks to process transactions more efficiently while maintaining security [?].

Bitcoin, for instance, uses the secp256k1 elliptic curve for key generation and signing, which provides a 256-bit key length offering high security with lower processing overhead compared to traditional cryptographic methods [?]. In contrast, Hyperledger Iroha leverages Ed25519 with SHA-3 to ensure a balance between performance, security, and compatibility with modern cryptographic best practices [?].

Public key cryptography is a cornerstone of blockchain security, enabling authentication, digital signatures, and secure communication. The adoption of ECC and its derivatives, such as ECDSA and EdDSA, has significantly improved the efficiency and scalability of blockchain networks, making them resilient to attacks while minimizing computational and storage costs. Future blockchain advancements may incorporate more advanced cryptographic techniques, including post-quantum cryptography, to further enhance security in decentralized systems.

2.3 Smart Contracts

Smart contracts represent a fundamental technological innovation that automates and enforces agreements between parties without the need for trusted intermediaries. First conceptualized by Nick Szabo in the late 1990s, they were envisioned as self-executing contractual arrangements in which terms are translated into code and enforced automatically by the system itself[?]. Szabo illustrated this concept using the analogy of a vending machine, where the mechanism guarantees the delivery of goods upon receiving correct payment, thereby minimizing the necessity for trust between parties[?].

2.3.1 Operations

From a technical standpoint, smart contracts are executable programs that encapsulate business logic and operate within virtual machines embedded in

blockchain or distributed ledger infrastructures. Their implementation typically follows a structured workflow: stakeholders define functional and operational requirements; conditional logic, such as payment authorization or delivery confirmation is formalized; developers translate this logic into code using domain-specific languages or frameworks; and the resulting contract undergoes rigorous testing and auditing prior to deployment. Once active, the contract listens for event data from trusted external inputs known as oracles. Upon detecting fulfillment of predefined criteria, the contract autonomously triggers execution[?].

This automation not only removes the need for intermediaries but also introduces transparency, auditability, and deterministic behavior in transaction processing. Every execution is permanently recorded on the blockchain, establishing an immutable trail that reinforces accountability.

2.3.2 Decentralized Applications

Within the decentralized application (dApp) ecosystem, smart contracts function as core computational units, encoding autonomous logic that operates without centralized oversight. This architecture facilitates trustless interactions among participants, allowing for complex operations such as asset transfers, service exchanges, or collaborative workflows, all without requiring mutual trust. Governance mechanisms embedded in these contracts ensure that operational rules are transparent and verifiable to all network participants. Moreover, their capacity to execute multi-party agreements algorithmically supports decentralized decision-making structures and streamlines value exchange across distributed systems[?].

2.3.3 Smart Contracts for Reproducibility

The application of smart contracts in scientific research introduces an innovative mechanism for addressing reproducibility challenges. By encoding experimental protocols, data acquisition processes, and analytical workflows into smart contracts, researchers can formalize and preserve their methodologies in a tamper-evident manner[?]. These contracts can autonomously coordinate tasks such as:

- Tracking data provenance.
- Recording experimental parameters.

- Executing statistical analyses.
- Publishing results with cryptographic verification.

This procedural formalization ensures that research outputs are not only transparent but also replicable by other investigators, thus aligning with the broader objectives of Open Science.

Smart contracts extend the capabilities of distributed ledger technologies by introducing programmable, self-executing logic into transactional workflows. Their role in enabling decentralized governance, automating complex processes, and preserving procedural integrity positions them as a foundational element in next-generation digital infrastructures. In the context of scientific research, smart contracts can transform how reproducibility is achieved, shifting it from a manual, error-prone process to one grounded in automation, transparency, and verifiability. As the ecosystem matures, continued research is essential to address technical, ethical, and interoperability challenges, thereby unlocking their full potential across domains.

2.3.4 Hyperledger Implementations

The Hyperledger ecosystem is an open-source collaborative effort hosted by the Linux Foundation Decentralized Trust, designed to advance cross-industry blockchain technologies. Rather than a single blockchain platform, Hyperledger serves as an umbrella project encompassing a suite of modular frameworks, tools, and libraries intended for developing enterprise-grade, permissioned distributed ledger solutions. It brings together a diverse community of developers, enterprises, and academic institutions to foster innovation in areas such as supply chain, finance, healthcare, and, increasingly, scientific research. By offering a range of specialized blockchain framework, each optimized for different technical and organizational needs, Hyperledger supports the creation of scalable, secure, and interoperable distributed systems.

The Hyperledger ecosystem comprises several smart contract frameworks, each with its own architecture and set of capabilities.

- **Hyperledger Fabric:** Hyperledger Fabric has progressively incorporated support for Ethereum Virtual Machine (EVM) bytecode, permitting the development of smart contracts in languages such as Solidity and Vyper. Introduced in version 1.3, this functionality facilitates the migration and creation of decentralized applications within permissioned networks, thereby broadening Fabric's applicability[?].

- **Hyperledger Burrow:** Hyperledger Burrow is an Ethereum-compatible, permissioned blockchain designed to execute smart contract code on a dedicated virtual machine. Powered by the Tendermint consensus engine, Burrow achieves transaction finality and high throughput suited to proof-of-stake environments. Its role within the Hyperledger suite includes a connection to the Tendermint/Cosmos ecosystem, an Ethereum side-chain compatible with expressive smart contract languages, and a modular, customizable EVM/Solidity execution library[?]. As a complete blockchain platform, Burrow integrates a Unix-style permissioning scheme within its EVM, supporting fine-grained control over actions such as token transfers, contract deployment, and validator participation[?].
- **Hyperledger Iroha:** Hyperledger Iroha adopts a unique strategy for implementing smart contracts. While version 1 offered a constrained command set without Turing completeness, version 2 introduces Iroha Special Instructions (ISI), which enable Turing-complete computation while preserving simplicity for routine operations. Contrary to Ethereum's model of deploying and interacting with contract instances, Iroha is structured around a core model of domains, accounts, and assets. These entities evolve through defined interactions. ISI behaves more like database triggers than autonomous smart contracts, activating in response to transactions that satisfy specific conditions[?].

2.4 The InterPlanetary File System

The InterPlanetary File System (IPFS) is a decentralized, peer-to-peer distributed file system designed to interconnect computing devices through a unified storage network [?]. By employing content-addressable storage and cryptographic hashing, IPFS enables the efficient sharing and retrieval of immutable data objects. Conceptually, it operates as a single, large-scale BitTorrent swarm exchanging objects within a Git-like repository. The system leverages a high-throughput, content-addressed block storage model with content-addressed hyperlinks, forming a generalized Merkle Directed Acyclic Graph (DAG). This architecture underpins various functionalities, including versioned file systems, blockchain data structures, and a more persistent and decentralized web. By integrating a Distributed Hash Table (DHT), an incentivized block exchange protocol, and a self-certifying namespace, IPFS eliminates single points of failure and minimizes the necessity for inter-node trust.

2.4.1 Background and Architectural Principles

IPFS is inspired by and synthesizes multiple established peer-to-peer technologies, including DHTs, BitTorrent, Git, and Self-Certifying Filesystems (SFS) [?]. While traditional web protocols such as HTTP provide an efficient means of requesting and transmitting data, they do not inherently incorporate modern file distribution strategies, resulting in limitations concerning redundancy, security, and scalability. By contrast, IPFS addresses these challenges by integrating and refining established distributed system techniques into a unified framework, ensuring efficient content retrieval and data persistence across geographically distributed nodes.

2.4.2 Key Architectural Components

IPFS is structured as a modular protocol stack, with distinct subsystems contributing to its overall functionality [?]:

- **Identities:** Manages node identity generation and verification, typically employing static cryptographic puzzles derived from S/Kademlia.
- **Network:** Governs peer-to-peer communication through various transport mechanisms, including WebRTC DataChannels and uTP.
- **Routing:** Facilitates object discovery and peer lookup through a DHT, defaulting to a structure based on S/Kademlia and Coral.
- **Exchange:** Utilizes BitSwap, a block exchange protocol that promotes efficient data replication through an incentive-driven market mechanism.
- **Objects:** Implements a Merkle DAG to store immutable, content-addressed objects, forming the foundation for versioned data structures.
- **Files:** Introduces a versioned file system hierarchy inspired by Git, supporting efficient data storage and retrieval.
- **Naming:** Incorporates a self-certifying mutable name system, enabling dynamic reference updates for evolving datasets.

These components collectively ensure a robust and scalable approach to distributed data management, making IPFS a compelling alternative to centralized storage solutions.

2.4.3 Content Addressing and Data Integrity

A fundamental principle of IPFS is its reliance on content addressing, wherein data objects are identified by the cryptographic hash of their contents [?]. This methodology not only facilitates efficient deduplication but also guarantees data integrity, as any modification to an object results in a new, distinct hash. The Merkle DAG structure further enhances the efficiency of storage and retrieval processes, allowing for seamless handling of large datasets and enabling robust versioning capabilities. By leveraging cryptographic hashing, IPFS inherently verifies data authenticity and prevents unauthorized modifications.

2.4.4 Routing and Peer Discovery

IPFS employs a DHT-based routing mechanism to locate peers and retrieve specific data objects [?]. This approach enables highly scalable and decentralized object discovery, as network participants contribute to maintaining a distributed index of content addresses. Small data values (typically under 1KB) can be stored directly within the DHT, whereas larger datasets are distributed across multiple nodes, optimizing both redundancy and retrieval speed.

2.4.5 Block Exchange and Incentive Mechanisms

The BitSwap protocol governs IPFS's block exchange mechanism, ensuring efficient data distribution and replication [?]. BitSwap operates as a marketplace where nodes trade blocks with one another, prioritizing the replication of rarer blocks to optimize availability. This economic model incentivizes participation and data persistence, ensuring that frequently accessed content remains accessible without relying on a centralized infrastructure.

2.4.6 Security and Node Authentication

Security within IPFS is underpinned by a cryptographic identity system, where nodes are identified by the hash of their public key [?]. This ensures that nodes can authenticate one another upon connection, exchanging public keys to verify identities. Additionally, IPFS adopts a flexible multi-hash format for cryptographic digests, allowing for algorithmic adaptability as cryptographic standards evolve.

2.4.7 Mathematical Foundations

The InterPlanetary File System (IPFS) is a peer-to-peer distributed file system that redefines content addressing and data distribution on the internet. At its core, IPFS leverages mathematical foundations from cryptographic hashing, directed acyclic graphs (DAGs), and content-based addressing to ensure integrity, permanence, and deduplication of data. Understanding these underlying principles is essential to appreciating how IPFS provides a robust and tamper-evident mechanism for storing and retrieving research data in a decentralized network. This section explores the key mathematical constructs that enable IPFS to achieve its goals of immutable, verifiable, and scalable content distribution.

Cryptographic Hashing Functions

The security and functionality of the InterPlanetary File System (IPFS) rely heavily on cryptographic hashing. IPFS typically employs the SHA-256 algorithm [?], which generates a fixed-size (256-bit) output from a variable-sized input. The mathematical properties of cryptographic hash functions that make them suitable for IPFS include:

- **Determinism:** The same input always produces the same output hash. Formally, for any input x , $\text{hash}(x) = h$, where h is the fixed-length output hash.
- **Pre-image Resistance:** Given a hash value h , it is computationally infeasible to find any input x such that $\text{hash}(x) = h$ [?]. Mathematically, for a given h , finding an x that satisfies $\text{hash}(x) = h$ is infeasible.
- **Second Pre-image Resistance:** Given an input x_1 , it is computationally infeasible to find a different input x_2 such that $\text{hash}(x_1) = \text{hash}(x_2)$ [?]. Formally, for any x_1 , finding $x_2 \neq x_1$ such that $\text{hash}(x_1) = \text{hash}(x_2)$ is infeasible.
- **Collision Resistance:** It is computationally infeasible to find any two different inputs x_1 and x_2 such that $\text{hash}(x_1) = \text{hash}(x_2)$ [?]. This property ensures that each piece of data in the IPFS system has a unique identifier and prevents malicious actors from substituting data.

Merkle Trees and Content Addressing

In IPFS, Merkle Trees play a critical role in efficiently managing and verifying large datasets. A Merkle Tree is a binary tree in which each leaf node is a hash of a data block, and each non-leaf node is a hash of its children. The root of the Merkle Tree provides a compact fingerprint of the entire dataset. This tree structure ensures efficient verification of data, allowing nodes to verify the integrity of large files with minimal data transfer [?].

Mathematically, a Merkle Tree can be described as a hash tree where:

$$H_{\text{root}} = \text{hash}(H_{\text{left}}, H_{\text{right}})$$

where H_{left} and H_{right} are the hashes of the left and right subtrees, respectively. The root hash H_{root} serves as the unique identifier for the entire dataset in IPFS.

In the context of IPFS, this mechanism ensures that content can be efficiently addressed and verified, even when data is distributed across multiple nodes in a decentralized system. Each piece of content in IPFS is associated with a unique cryptographic hash, known as the Content Identifier (CID), which is derived from the content itself using the SHA-256 hashing algorithm [?].

Implications for Data Integrity and Security

The reliance on cryptographic hashing, combined with Merkle Trees, allows IPFS to guarantee data integrity and security. By hashing each piece of content, IPFS ensures that data cannot be tampered with or modified without altering its corresponding hash. This provides the basis for trust in a decentralized network, where data retrieval can be verified against the content's cryptographic fingerprint.

The collision resistance property of the hash function also ensures that two different pieces of data will not have the same identifier, thereby eliminating the risk of data substitution or forgery. This is especially crucial in a distributed system where trust must be established without the need for central authority [?].

Efficiency and Scalability

While the cryptographic operations in IPFS ensure data integrity and security, they also contribute to the system's efficiency. The use of SHA-256 and Merkle Trees allows IPFS to quickly verify large datasets with minimal overhead. Furthermore, as each piece of content is independently addressable

through its hash, the system is highly scalable, allowing for decentralized and efficient content distribution [?].

However, the computational complexity of cryptographic operations and the overhead associated with maintaining Merkle Trees could pose challenges for large-scale systems, especially as the volume of data increases [?]. These challenges must be considered when evaluating the scalability of IPFS in future applications.

2.5 Immutable Records and Scalable Storage

Blockchains are highly effective in establishing distributed consensus and maintaining immutable transaction records. However, they exhibit significant limitations when used for storing large and heterogeneous datasets. Large files often vary widely in both size and format, and storing such data directly on-chain results in excessive volume growth, placing unsustainable demands on storage resources [? ?]. These limitations arise because blockchain architectures are not inherently designed to support the data-intensive operations typical of many decentralized applications.

By providing a decentralized storage layer, IPFS addresses this challenge while complementing blockchain infrastructure by offering a decentralized storage layer that complements blockchain infrastructure. Rather than placing entire files on-chain, only their content identifiers (CIDs), cryptographic hashes of the file contents, are recorded on the blockchain. This approach substantially reduces the storage burden while preserving verifiable links to the original data [? ?]. Through content addressing, IPFS ensures data integrity and decouples bulk storage from the blockchain itself, thereby enhancing scalability and operational efficiency [?].

2.5.1 Limitations of Ledger and Storage Systems

Despite their individual strengths, both technologies face inherent limitations when used in isolation. Blockchain's scalability is constrained by the requirement that all nodes replicate the complete on-chain state, which significantly limits its capacity to handle large-scale data [?]. Moreover, transaction fees, particularly on public blockchains, can further exacerbate these limitations, creating cost barriers to widespread adoption [?].

Conversely, IPFS provides a scalable and cost-effective mechanism for decentralized storage [?], but it does not offer guarantees for long-term data

persistence. Files stored on IPFS may become inaccessible if no node continues to host them. Additionally, while IPFS supports efficient content distribution, it lacks built-in mechanisms for immutability and fine-grained access control [?], both of which are essential in contexts such as research data management or digital provenance.

2.5.2 Integrating Ledger with Distributed Storage

The integration of IPFS with blockchain infrastructure offers a synergistic solution that addresses the individual limitations of each technology. Instead of storing full datasets on-chain, only the IPFS-generated CIDs are recorded on the blockchain [?]. This hybrid approach minimizes storage costs while maintaining the ability to verify data authenticity and integrity.

For instance, research datasets can be stored on IPFS, while their CIDs are immutably anchored to the blockchain to guarantee provenance. Although a CID uniquely identifies a file, modifying the content yields a new CID without inherently linking it to its previous version [?]. Blockchain resolves this by offering an immutable and timestamped record of each CID, thereby supporting tamper-proof verification. Even minor alterations in a file generate a new CID, allowing changes to be immediately detected through comparison with the recorded CID.

Moreover, blockchains can be employed to manage metadata in conjunction with IPFS. While the IPFS network handles the actual file content, the blockchain stores associated metadata, such as ownership credentials, timestamps, version history, and descriptive annotations. This metadata framework enhances the transparency, traceability, and auditability of the data stored on IPFS, reinforcing its usability in Open Science and other provenance-critical domains.

2.6 Open Science and Reproducibility

To understand the proposed d-OSP, this chapter discusses fundamental concepts and synthesizes relevant scholarly work, situating the research within ongoing debates and developments. It reviews prior initiatives that have leveraged decentralized technologies to address the reproducibility crisis in scientific research, highlighting both their contributions and limitations. Through this analysis, the chapter establishes the conceptual grounding for the proposed model, identifying the key theoretical and technological foundations that inform the design and implementation of the d-OSP.

The core objectives of Open Science are to enhance the transparency, accessibility, and reproducibility of research, while fostering collaboration and accelerating scientific discovery [?].

The reproducibility crisis in scientific research has prompted widespread interest in Open Science as a corrective framework. The Center for Open Science articulates five foundational principles: Open Data, Open Materials, Open Access, Preregistration, and Open Analysis, each contributing to the transparency, verifiability, and reliability of research outcomes. This section presents each principle in turn, emphasizing its significance for reproducibility.

- **Open Data:** This principle advocates for unrestricted access to the raw datasets underlying scientific publications. Open Data allows independent researchers to validate findings, reanalyze results using alternative methods, and repurpose data for novel investigations. Effective implementation depends on the use of domain-specific repositories that support standardized metadata formats, persistent identifiers (e.g., DOIs), and long-term availability. Provenance metadata, which details the origin, transformation, and contextual history of datasets, plays a crucial role in enabling traceability and ensuring reproducibility.
- **Open Materials:** Reproducibility requires access to the tools, instruments, and procedures used in the original study. Open Materials involves sharing experimental protocols, questionnaires, software, and computational tools needed to replicate findings. When properly archived in interoperable repositories with sufficient metadata and licensing information, these materials facilitate independent verification and extension of prior work, reducing ambiguity and enhancing methodological transparency.
- **Open Access:** This principle aims to remove paywalls and legal barriers to accessing scholarly publications. By ensuring that articles are freely available to all readers, Open Access increases the visibility, accessibility, and impact of research. For reproducibility, Open Access is critical: it ensures that the methods, analyses, and interpretations reported in a study can be scrutinized by a broader community, including researchers without institutional subscriptions. This accessibility is essential for enabling transparent post-publication peer review and validation.
- **Preregistration:** Preregistration entails publicly documenting a study's research questions, hypotheses, design, and planned analyses before

data collection begins. This practice mitigates risks associated with selective reporting, outcome switching, and post hoc hypothesis formulation. Preregistration records are typically stored in time-stamped registries and serve as an immutable reference that distinguishes confirmatory research from exploratory investigations, thereby enhancing both transparency and epistemic rigor.

- **Open Analysis:** Open Analysis refers to the practice of sharing statistical code, computational workflows, and digital environments used to produce published results. When accompanied by proper documentation and executed within reproducible environments (e.g., containers or notebooks), open analytical pipelines allow others to verify computational steps, reproduce results, and explore alternative methodologies. This level of transparency is particularly vital in data-intensive fields, where minor changes in code or parameters can substantially influence outcomes.

Collectively, these five principles form the basis of a robust and reproducible scientific workflow. The artifact presented in this dissertation aligns with these principles by embedding technical mechanisms, such as blockchain-based audit trails, decentralized storage, and smart contracts, that enforce transparency, traceability, and verifiability across the research lifecycle.

2.6.1 Benefits, Pros, and Cons of Open Science

The implementation of Open Science practices offers several potential benefits. These include:

- **Increased Collaboration:** Open Science fosters a more collaborative environment, enabling researchers to build on each other's work and accelerating scientific progress [?].
- **Greater Transparency and Trust:** Open access to data and methodologies improves the transparency of research processes, fostering public trust in scientific findings [?].
- **Enhanced Reproducibility:** Open Science enhances the reproducibility of research by making raw data, code, and methods publicly available for verification [?].

- **Public Engagement:** Open Science facilitates public engagement with research, empowering individuals to access scientific findings and contribute to the discourse [?].

However, there are also several challenges and drawbacks:

- **Funding and Resource Allocation:** Open Science initiatives often require significant investment in infrastructure, software, and training, which may not always be supported by traditional funding models [?].
- **Intellectual Property Concerns:** Researchers may be hesitant to share data or methodologies due to concerns about intellectual property protection or the potential misuse of their work [?].
- **Quality Control Issues:** Open peer review and open data sharing may raise concerns about the quality and integrity of research, as the absence of rigorous vetting processes can lead to the publication of incomplete or erroneous findings [?].

2.7 Summary

This chapter established the foundational concepts necessary for the development of the Decentralized Open Science Platform. It began by distinguishing key characteristics of decentralized and distributed systems, followed by an introduction to blockchain technologies and their relevance to data integrity and transparency. Smart contracts were then examined as programmable and enforceable agreements embedded within blockchain infrastructures. The chapter also addressed the InterPlanetary File System as a peer-to-peer storage solution and analyzed its complementary relationship with blockchain. Finally, the principles of the Open Science movement were explored, highlighting its emphasis on transparency, accessibility, and reproducibility in scientific research. Collectively, these topics provide the conceptual and technical groundwork for the design and implementation of the platform presented in the following chapters.

Chapter 3

Literature and Area Review

*That's not right.
That's not even wrong.*

*Wolfgang Pauli, Austrian physicist & pioneer of quantum mechanics
(1900–1958)*

To contextualize the proposed Decentralized Open Science Platform, this chapter surveys the current landscape of reproducibility and Open Science. It synthesizes recent developments, initiatives, and scholarly contributions that address reproducibility concerns across disciplines. Section ?? introduces the reproducibility crisis in scientific research, outlining key drivers and implications. Section ?? explores the persistent challenges that undermine reproducibility efforts. Section ?? presents influential initiatives in Open Science and research data management, underscoring institutional and technological responses. Section ?? reviews related work that employs decentralized technologies to enhance transparency, traceability, and accountability in research workflows. These sections collectively position the present study within ongoing conversations about reproducibility, offering a critical foundation for understanding the rationale and design of the d-OSP. A summary of key insights is provided in Section ??.

3.1 Reproducibility Crisis in Science Research

Science, as a systematic and empirical endeavor, fundamentally depends on the ability of researchers to verify and extend the findings of others. Central to this process is the concept of reproducibility, which encompasses both the capacity to obtain consistent results using the same data and methods, and the ability to reach comparable outcomes when similar studies are conducted with new data under the same experimental design [? ?].

Concerns about the difficulty of reproducing published research findings have gained prominence across a wide range of disciplines. This phenomenon, commonly termed the "reproducibility crisis", has cast doubt on the reliability of scientific outputs and eroded confidence in the research enterprise [?]. Alarmingly, studies suggest that as many as 50% of published experiments may fail to replicate, pointing to a systemic issue that extends beyond isolated methodological flaws or instances of misconduct [?].

In the life sciences, the financial implications of irreproducible research are substantial. It is estimated that over \$28 billion is lost annually in the United States due to studies that do not meet reproducibility standards [?]. These findings underscore the urgency of adopting systemic reforms aimed at improving transparency, methodological rigor, and long-term accessibility to scientific materials and data.

3.1.1 Challenges to Scientific Integrity

The consequences of the reproducibility crisis extends beyond the academia to affect public trust in science, slow down the translation of research into practical applications, and potentially lead to the misallocation of substantial resources and the implementation of misinformed policies based on unreliable findings. The inability to reproduce preclinical research, for example, can significantly delay the development of therapies that are life saving, increase the pressure on already strained research budgets, and drive up the costs associated with drug development. The societal impacts are also significant, with misdirected effort, funding, and policies potentially being implemented based on research that cannot be validated [?].

Several interconnected factors contribute to this crisis, spanning issues within the publication system to the prevalence of questionable research practices and the inherent complexities encountered in certain scientific disciplines. Journals often exhibit a publication bias, preferentially publishing

novel and positive results while overlooking negative findings or replication studies [?]. This creates a skewed representation of the scientific landscape and can lead to the neglect of important information about what does not work [?]. Furthermore, researchers may engage in questionable research practices, such as p-hacking (manipulating data to achieve statistical significance) and HARKing (hypothesizing after results are known), which can distort results and make replication exceedingly challenging. Inadequate statistical methods, including the use of suboptimal analyses, can also lead to erroneous conclusions, further hindering the replication process. A significant contributing factor is the lack of data sharing among researchers; when data and methods are not openly accessible, the ability of others to verify and replicate the work is severely limited [?].

The intense pressure to publish, often described by the expression "publish or perish," can incentivize researchers to prioritize the quantity of publications over their quality, potentially leading to rushed and less rigorous research. Incentive structures within universities may inadvertently reward the mere act of publication in prestigious journals, sometimes at the expense of methodological rigor and the pursuit of accurate and reproducible findings. This competitive environment can implicitly or explicitly encourage the use of questionable research practices to achieve publication, such as selectively reporting parts of datasets or trying different analytical approaches until the desired outcome is obtained [?].

The reproducibility crisis in science also reveals a strong connection between data management practices and the ability to replicate experimental results. Transparent and accessible data are essential for verifying findings and ensuring their reliability across disciplines. Insufficient metadata, unavailability of raw data, and incomplete methodological reporting are major contributors to irreproducibility. Without proper documentation and sharing protocols, researchers face significant barriers in reusing or validating published results [?].

3.1.2 A Paradigm Shift

In response to concerns about the reproducibility and reliability of scientific production, a movement emerged advocating for a fundamental transformation in how knowledge is generated and disseminated, emphasizing transparency, accessibility, and collaboration within the scientific community and with the broader public. Although the ideals of openness and sharing have long been embedded in scientific practice, the Open Science movement gained momentum with the advent of the internet and the more interactive capabilities made available by the Web 2.0 [?].

3.2 Challenges in Scientific Reproducibility

One of the primary objectives of the Open Science movement is to address the persistent concern surrounding research reproducibility. As a cornerstone of scientific integrity, reproducibility ensures that findings can be independently verified and validated [?]. Despite its importance, a considerable number of scientific studies, particularly in complex domains such as biomedical research, remain difficult or impossible to replicate [?]. Several interrelated factors contribute to these reproducibility challenges:

- **Insufficient Methodological Reporting:** Many studies lack the methodological detail necessary for replication, impeding efforts to validate results independently [?].
- **Limited Access to Data and Code:** The unavailability of raw data and source code prevents independent analysis and verification of published findings [?].
- **Academic Incentive Structures:** The prevailing "publish or perish" culture emphasizes speed and quantity over transparency and rigor, often compromising reproducibility [?].

Although numerous initiatives have emerged to strengthen reproducibility, such as the establishment of reporting standards, the promotion of data sharing platforms, and the adoption of open source tools, substantial obstacles persist in fostering a transparent and reproducible scientific ecosystem [?].

Despite these challenges, Open Science offers a compelling framework for transforming how research is conducted, disseminated, and validated. Achieving its full potential, however, will depend on continued efforts to address barriers related to funding models, intellectual property, and systemic academic practices. Nonetheless, the movement remains a critical pathway toward enhancing the transparency, collaboration, and reliability of scientific research.

3.3 Key Initiatives

The growing emphasis on transparency, reproducibility, and collaboration in scientific research has led to the emergence of several influential

initiatives that support the implementation of Open Science and effective Research Data Management (RDM). These initiatives provide frameworks, tools, and community-driven guidelines that help researchers and institutions manage data more responsibly, ensuring that research outputs are not only preserved but also accessible and reusable. By fostering interoperability, encouraging FAIR (Findable, Accessible, Interoperable, and Reusable) data practices, and promoting a culture of openness, these efforts contribute to a more trustworthy and efficient research ecosystem. This section discusses a selection of leading initiatives spanning international collaborations, policy frameworks, and infrastructural developments that collectively shape the evolving landscape of Open Science and RDM.

Leveraging European Research Data

The LEARN Toolkit (Leveraging European Research Data) was developed to assist research institutions in implementing effective Research Data Management (RDM) policies and practices. Grounded in the recommendations of the LERU (League of European Research Universities) Roadmap for Research Data, the Toolkit offers guidance on institutional policy development, advocacy, training, infrastructure, and best practices. It emphasizes the strategic role of data management planning and encourages institutions to embed RDM into the research lifecycle. By providing a series of model policies, case studies, and checklists, LEARN promotes a culture of data stewardship aligned with the principles of FAIR data (Findable, Accessible, Interoperable, and Reusable), contributing to the broader objectives of Open Science [?].

FAIR Guiding Principles

Represents a cornerstone of responsible data stewardship in the context of Open Science. These principles aim to improve the infrastructure supporting the reuse of scholarly data. By encouraging data producers to make their outputs Findable, Accessible, Interoperable, and Reusable, FAIR fosters machine-readability, long-term preservation, and seamless data integration across platforms and disciplines. Although not inherently open, FAIR complements Open Science by providing the technical and semantic standards necessary for data sharing and reuse. Adoption of FAIR principles by research funders, repositories, and institutions has significantly influenced data policies across scientific communities and reinforced efforts toward more transparent and collaborative research practices [?].

GO FAIR

This initiative builds on the momentum of the FAIR principles, functioning as a bottom-up, stakeholder-driven movement to implement FAIR

data stewardship globally. It encourages the development of implementation networks and collaborative groups that share expertise and develop domain-specific solutions for achieving FAIR data practices. GO FAIR's focus extends to governance, education, and infrastructure, aiming to create a distributed ecosystem that facilitates the reuse of scientific data. By promoting interoperability standards and cultural change across the scientific community, GO FAIR advances Open Science by ensuring that data outputs can be seamlessly discovered, accessed, and reused across institutional and national boundaries [?].

Research Data Alliance

The Research Data Alliance (RDA) is a global community-driven initiative that brings together data practitioners, technologists, and policymakers to build the social and technical infrastructure necessary for open data sharing across disciplines. Founded in 2013, RDA operates through working groups and interest groups that develop recommendations, standards, and best practices for data interoperability and stewardship. The RDA fosters international cooperation and bridges disciplinary gaps by aligning data governance, metadata standards, and infrastructure development. Its outputs support the implementation of Open Science by ensuring that research data is not only preserved but also rendered useful and actionable across diverse research contexts [?].

Committee on Data of the International Science Council

Is an international organization committed to advancing data science and improving the quality and accessibility of research data. It supports the development of data policies, fostering international collaboration, and providing strategic guidance on data governance. CODATA actively contributes to the advancement of the FAIR principles and supports initiatives that aim to make research data a reusable, sustainable, and equitable public good. Through its coordination efforts and engagement with global stakeholders, CODATA helps shape the infrastructures and norms that underpin responsible data sharing and Open Science [?].

Open Access Infrastructure for Research in Europe

Represents a pan-European initiative designed to support the open dissemination and reuse of research outputs. Originating as a response to the European Commission's Open Access policies, OpenAIRE has developed into a robust infrastructure that aggregates metadata and full-text content from a wide array of data providers, including institutional repositories, data

archives, and scholarly journals. By facilitating interlinking between publications, datasets, software, and project information, OpenAIRE enhances the discoverability and interoperability of research products across disciplines. Its suite of services, such as the OpenAIRE Graph and Research Community Dashboards, provides tools for compliance monitoring, impact assessment, and reproducibility tracking. Furthermore, OpenAIRE actively contributes to policy development and technical alignment in the global Open Science ecosystem, advocating for standardized metadata schemas and persistent identifiers. Through its alignment with FAIR principles and support for the European Open Science Cloud (EOSC), OpenAIRE plays a foundational role in shaping a transparent, interconnected, and researcher-centric data landscape [?].

DataCite

Is a global non-profit organization that plays a foundational role in the research data ecosystem by providing persistent identifiers, most notably Digital Object Identifiers (DOIs), for datasets and other research outputs. Founded to support data citation practices, DataCite promotes the discoverability, accessibility, and reuse of research data by ensuring that data can be persistently linked to scholarly publications and contributors. It collaborates with data centers, publishers, and repositories to establish metadata standards that facilitate interoperability across infrastructures. Through services such as DOI registration, metadata management, and citation tracking, DataCite actively contributes to the implementation of the FAIR principles and strengthens the overall architecture of Open Science and Research Data Management worldwide [?].

Nelson Memo - Office of Science and Technology Policy (OSTP)

The White House Office of Science and Technology Policy (OSTP) issued a directive known as the “Nelson Memo,” which requires that all federally funded research publications and associated data be made freely available to the public. This policy marks a pivotal shift in U.S. open access strategy by eliminating embargo periods and strengthening mandates for data transparency. Building on previous open science policies, the Nelson Memo seeks to ensure equitable access to publicly funded knowledge, drive reproducibility, and accelerate scientific progress through a national commitment to openness and accountability [?].

3.4 Related Work

The limitations of traditional research data management systems have sparked growing interest in alternative models. Decentralized technologies,

particularly blockchain, have gained increasing recognition for their ability to enhance transparency, accountability, and trust across various domains, including scientific research. Their potential to address long-standing inefficiencies and structural shortcomings within the research ecosystem has attracted significant attention from the academic community.

Blockchain has evolved into a broader paradigm of distributed ledger technology, collectively maintained by a network of nodes. Through immutability and consensus-based validation, it ensures the integrity of recorded data. These foundational features offer a technological infrastructure for verifying the authenticity, provenance, and persistence of digital records, features that align closely with Open Science objectives and the FAIR principles (Findable, Accessible, Interoperable, and Reusable). In an era of data-intensive research and multi-stakeholder collaboration, such assurances are critical for enabling reproducibility, facilitating the auditability of research processes, and ensuring reliable attribution of intellectual contributions.

Decentralized solutions introduce a novel approach to scientific data governance. This paradigm shift directly supports key principles of Open Science such as openness, inclusivity, reproducibility, and collaboration by embedding accountability and traceability into the technical fabric of research infrastructures.

In this context, decentralized applications function as strategic enablers of both cultural and procedural transformation in science. They offer pathways to reconfigure incentive structures, reduce access barriers, and reinforce the reproducibility and credibility of scientific outputs. The remainder of this section explores the current state of such applications, their underlying architectures, and the roles they play in advancing Open Science and addressing reproducibility challenges.

Blockchain technologies have increasingly drawn scholarly attention for their potential to transform academic publishing and research dissemination. An early synthesis of this trajectory is provided by [?], who captured discussions from the 2018 NFAIS conference on blockchain in scholarly communication. The article emphasized blockchain's capacity to reshape research workflows, peer review, and intellectual property management, spotlighting pilot initiatives such as ARTiFACTS and Knowbella Tech. These efforts highlighted the technology's promise for enhancing provenance tracking and decentralizing research funding. However, despite visible enthusiasm for secure, transparent, and decentralized scholarly infrastructures, the article concluded that broader adoption hinges on increasing awareness and achieving a more nuanced understanding of blockchain's capabilities.

Echoing this vision of decentralization, [?] explores how blockchain might disrupt traditional academic publishing workflows by unbundling them and empowering content creators with more equitable systems of recognition and compensation. Through token-based platforms like Steem, BAT, and LBRY, the article illustrates alternatives to centralized content discovery systems that could provide direct value to researchers. It argues that designing systems centered on discoverability and monetization may help build more effective and transparent publishing ecosystems. Nevertheless, it acknowledges that realizing these benefits would require a fundamental rethinking of current revenue models.

In a similar fashion, [?] frames blockchain as a vehicle for fulfilling Eugene Garfield's vision of recognizing a broader spectrum of scholarly contributions. Using ARTiFACTS as a case study, the article describes how blockchain can secure provenance and attribution even during the early stages of research. This capability opens the door to formal recognition for pre-publication outputs and thereby alters existing academic incentive structures. It presents a long-term outlook in which blockchain infrastructure supports not only open science principles but also new academic reputation systems.

A broader perspective is provided by [?], who examines the potential of blockchain to address inefficiencies and credibility deficits in science. The article discusses its applications in peer review, reproducibility, and research funding via cryptocurrency-based mechanisms. Concepts such as decentralized repositories, micro payments, and digital rights management are introduced as possible innovations for academic publishing. Although the potential to support transparency and equity is clear, the article cautions that overcoming entrenched institutional inertia remains a major hurdle to implementation.

The technical pathways to integrate blockchain with existing systems are examined in [?], which proposes a blockchain-based cloud middleware to improve manuscript submission and peer review. Rather than disrupting current infrastructures, the framework emphasizes compatibility, with the aim of enhancing anonymity, reducing bias, and increasing decentralization. Built with Java Spring and Ethereum, and tested on simulated data, the proposed system demonstrates viability but also exposes real-world challenges related to scalability and integration.

Offering a broader field overview, [?] provides a systematic review of 60 blockchain-based projects, mapping their relevance to open science. The

article identifies decentralization, immutability, and transparency as key alignments between blockchain and open science principles. The projects span areas such as reproducibility, secure sharing, and intellectual property management. However, despite the enthusiasm, the article points out barriers such as the lack of standardization, technical risks from smart contracts, and the difficulty of creating sustainable incentive models. It concludes that while blockchain is promising, a supportive ecosystem is required for its broader acceptance and functionality within scientific communities.

Practical experimentation with blockchain for peer review incentives is addressed by [?], which introduces Ants-Review, a protocol utilizing Ethereum smart contracts to reward high-quality anonymous reviews. The system uses ANTS tokens as reward for peer reviews and integrates the AZTEC protocol to maintain anonymity. Through a gamified evaluation process and plans for Decentralized Finance (DeFi) integration and DAO-based governance, the authors propose a robust and inclusive mechanism to promote more fair and efficient scientific publication.

The potential of blockchain potential applications in academia is examined [?], including its role in open data sharing, governance, and transparent evaluation of academic output through immutable tracking of citations and usage metrics. Although the outlook is optimistic, the study also notes ongoing challenges such as the complexity of user interfaces, unresolved legal questions, and the friction between decentralized technologies and established academic conventions.

Exploring newer technologies, [?] proposes non-fungible tokens (NFTs) as a means to restore ownership and value in scholarly communication. The article suggests pathways for integrating NFTs into university presses and submission platforms, under conditions of zero cost, autonomy, interoperability, and ease of use. Complementing this vision is the development of Open Lab [?], a browser-based experiment platform that integrates with the Open Science Framework to simplify participant management and promote open collaboration, thus further supporting the ethos of transparent and reproducible research.

A more security-focused solution is proposed by [?], who introduce Open-Pub, a private Ethereum-based publishing system. Addressing the dual demands of privacy and transparency, it features a threshold identity-based group signature scheme for anonymous peer review and uses IPFS for decentralized data storage. A token-based incentive structure for reviewers aims to strengthen community engagement. Although incentives for authors are less explored, the architecture reflects the growing demand for systems that reconcile openness with secure control in academic publishing.

Finally, [?] critiques the limitations of current academic recognition models, which prioritize publications and citations while overlooking essential but undervalued contributions. Proposing a blockchain-based token economy, the article advocates rewarding a wider range of academic activities, including peer review and committee work with non-tradable, non-transferable tokens, offering a pathway to more holistic and inclusive systems of academic recognition.

3.5 Summary

This chapter surveyed the evolving landscape of reproducibility and Open Science to contextualize the proposed Decentralized Open Science Platform (d-OSP). It introduced the reproducibility crisis and identified systemic and methodological factors that contribute to irreproducible research. Persistent challenges, such as fragmented incentives, insufficient infrastructure, and lack of standardisation, were examined to highlight barriers to reform. A review of key initiatives showcased institutional and technical efforts aimed at promoting transparency and data stewardship. The chapter also explored how decentralized technologies, particularly blockchain, are being employed to support traceability, accountability, and openness in research workflows. Together, these insights establish the foundation for the d-OSP by demonstrating both the urgency of reproducibility issues and the potential for decentralized approaches to contribute meaningful solutions.

Chapter 4

Proposed Decentralized Model

*Nothing in life is to be feared,
it is only to be understood.
Now is the time to understand more,
so that we may fear less.*

*Marie Curie, Polish and naturalised-French physicist and chemist
(1867–1934)*

The d-OSP is conceived as a decentralized infrastructure for research data management. This chapter outlines the guiding design criteria and rationale and details its key technical components. It introduces the main features and of the platform in Section ??, followed by an overview of the underlying technology stack in Section ?. The operational workflow is described in Section ?. The internal data representation is explained in Section ?, and its integration with the Hyperledger Iroha v1 data model is discussed in Section ?. Section ? explores the role of ontologies in supporting semantic interoperability, while Section ? focuses on the role of metadata in ensuring traceability and linkage between users, projects, and storage layers. The chapter concludes with a summary in Section ?.

4.1 The Decentralized Open Science Platform

Traditional centralized research repositories often exhibit data silos, limited verifiability, and susceptibility to manipulation, impeding the openness and reliability of scientific practices. The decentralized model introduced in this work is designed to mitigate these challenges by enabling efficient data sharing, fostering collaboration, and enhancing the validation of research outputs, thereby strengthening reproducibility and transparency.

This chapter details the design and implementation of the d-OSP, a infrastructure that integrates blockchain, IPFS, and smart contracts to improve research reproducibility. By leveraging immutable records and decentralized storage, the platform ensures transparent and verifiable research data management. Additionally, extended services are incorporated to facilitate file indexing, metadata extraction, and search functionality. The proposed platform aligns with Open Science principles by providing verifiable and persistently traceable access to research data. Figure ?? presents the high level building blocks of the platform.

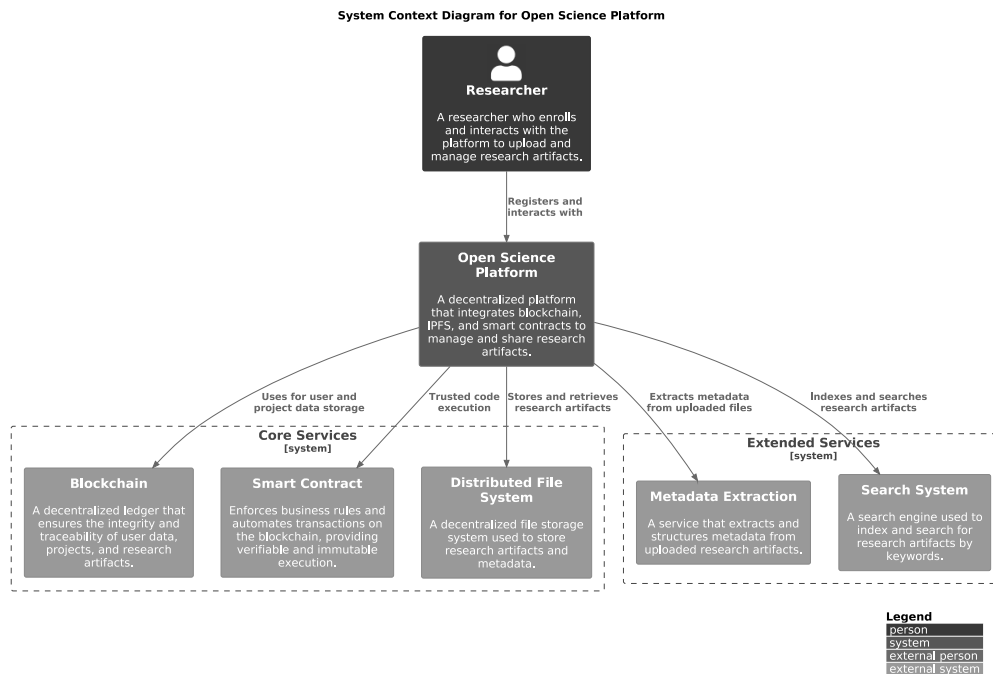


Figure 4.1: System context diagram.

4.1.1 Distributed Trust Architecture

The d-OSP integrates decentralized and distributed technologies to enhance reproducibility in scientific research. Traditional research infrastructures often suffer from data silos, paywalled access, and risks of data loss or manipulation. By leveraging blockchain and IPFS, the platform ensures that research data remains tamper-proof, permanently accessible, and verifiable.

Researchers can upload experimental protocols, datasets, and publications to the IPFS decentralized network, preventing single points of failure and enabling unrestricted access to research outputs. Blockchain serves as a provenance-tracking mechanism by recording immutable hashes (CIDs) of research data, ensuring the integrity and authenticity of published findings.

Through the integration of these technologies, the d-OSP mitigates the risks associated with centralized control in research dissemination. Traditional repositories may impose restrictions on data access, suffer from institutional biases, or become unavailable over time. In contrast, a decentralized infrastructure empowers researchers to share knowledge freely, ensuring that scientific progress remains transparent and universally accessible.

Decentralization and distributed systems redefine how data integrity, accessibility, and transparency are maintained across various domains. Blockchain and IPFS provide complementary solutions that enhance security, immutability, and scalability. In the context of Open Science, these technologies eliminate reliance on centralized institutions, ensuring that research data remain verifiable and permanently accessible. By leveraging decentralization, the d-OSP fosters an ecosystem of trustless collaboration, where scientific knowledge can be openly shared and validated by the global research community. To fully grasp the impact of these technologies, it is essential to examine their core components and underlying mechanisms. The following sections explore blockchain fundamentals, decentralized applications (dApps), and IPFS, detailing how each contributes to building a resilient and transparent digital infrastructure to enhance science reproducibility.

4.2 Technology Stack

The d-OSP is built on a hybrid architecture that strategically integrates decentralized and centralized components to balance security, traceability, and efficiency in data management. Decentralized technologies, such as

blockchain and IPFS, ensure data integrity and tamper resistance, while centralized components facilitate indexing, search, and user interactions. Figure ?? presents a high-level breakdown of the platform's core building blocks.

4.2.1 Core Services

The core services of the d-OSP provide the fundamental infrastructure for secure and verifiable research data management.

- **Hyperledger Iroha v1 Blockchain:** Acts as the immutable ledger 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 files and metadata, ensuring persistent and verifiable access to shared data.

4.2.2 Extended Services

The extended services enhance the platform's features by improving file and metadata processing.

- **Apache Tika:** Extracts metadata from uploaded files, enhancing research data organization and search.
- **Whoosh:** Facilitates efficient indexing and keyword-based search for stored files.

4.2.3 User Interface, integration and execution

- **Jupyter Notebooks (Python):** Powers the front-end interface, facilitating the automation and display of the execution steps. Blockchain interactions are managed via the Iroha v1 Python library, while communication with the IPFS network is handled through the HTTPS client library.

4.2.4 System Components and Interactions

The d-OSP consists of multiple interconnected components, each serving a distinct role in ensuring secure, verifiable, and reproducible research data management. The primary components include Jupyter Server, the blockchain Hyperledger Iroha v1 and the InterPlanetary File System (IPFS). Each of these elements are encapsulated within a Docker container to provide modularity, ease of deployment and reproducibility. The implementation level architecture is presented in Figure ??, the network topology is depicted in Figure ??

4.2.4.1 Jupyter Server

The Jupyter Server serves as the primary interface for users engaging with the platform through notebooks. It provides a Python kernel as the execution environment, integrating the Iroha v1 library, an IPFS HTTPS client, Apache Tika for metadata processing, and the Woosh indexer and search system. This environment allows users to:

- Execute Python scripts to submit transactions and perform queries on the blockchain using smart contracts.
- Upload and retrieve files and metadata (as JSON objects) via IPFS.
- Process and index research data with Apache Tika and Woosh to improve searchability.
- Access and visualize metadata anchored on the blockchain in support of Open Science practices.

4.2.4.2 Blockchain

The blockchain runs based on a Hyperledger Iroha v1 network and acts as a distributed ledger for recording transactions. It ensures immutability, transparency, and verifiability of stored research metadata. This component:

- Receives transactions from the Jupyter Server via a gRPC API.
- Stores metadata references, ensuring that uploaded research files can be authenticated.

- Interacts with PostgreSQL for structured storage of blockchain metadata.
- Supports smart contracts through the integration of Hyperledger Burrow, which provides a modular blockchain client with a permissioned smart contract interpreter partially developed to the specification of the Ethereum Virtual Machine (EVM).

4.2.4.3 Storage

The InterPlanetary File System (IPFS) is a decentralized storage solution that manages the research outputs. This component:

- Stores digital research files in a content-addressed manner.
- Allows the Jupyter Server to upload and retrieve files via an HTTPS API.
- Ensures long-term availability of scientific data through distributed storage principles.

4.2.4.4 Relational Database (PostgreSQL)

The PostgreSQL database provides structured storage for blockchain-related data. It is used exclusively and managed by Iroha v1 to:

- Maintain an efficient and queryable record of transactions.
- Ensure that research metadata stored on the blockchain can be retrieved and verified.
- Support blockchain operations requiring fast access to structured data.

4.2.4.5 Component Interactions

The components interact in a seamless and decentralized manner:

- The user submits transactions, uploads files, and queries research data through Jupyter Notebooks.

- The Jupyter Server communicates with the Iroha blockchain using the gRPC API to store and retrieve metadata.
- Iroha manages its internal PostgreSQL database to persist metadata and account records efficiently.
- Research files are uploaded to IPFS, with their content identifiers (CIDs) stored on the blockchain for traceability.
- Users retrieve files from IPFS using their respective CIDs, ensuring content integrity and reproducibility.

This architecture guarantees trustworthy and reproducible scientific research by leveraging blockchain for integrity, IPFS for decentralized storage, and Jupyter as an accessible research environment.

4.3 Operations

The platform supports a set of core operations that regulate user interactions with projects and research artifacts.

4.3.1 User Enrollment and Project Registration

The d-OSP enables user enrollment and project registration, ensuring transparent and verifiable account management on the blockchain. Users self-enroll by providing cryptographic credentials and identity details, which are securely stored using a combination of blockchain attributes and decentralized storage through IPFS. Similarly, projects are registered with essential metadata, establishing a distinct blockchain account for each. To maintain traceability and facilitate efficient project management, the system links user and project accounts bidirectionally, allowing for streamlined queries and provenance tracking. These processes are depicted in Figure ??.

- **User Self-Enrollment** – A user self-enrolls on the platform by providing a private key that complies with the ED25519 or SHA-3 standards and identity information, including full name, institution, email, ORCID, and role. An account is created for the user in the blockchain. All data provided in the enrollment is structured in key/value pairs into a JSON object and uploaded to IPFS, with the corresponding Content Identifier (CID) permanently linking the user's metadata to their blockchain account.

- **Project Registration** – Users can register a project by specifying a descriptive name, an abstract, relevant keywords, start and end dates, funding agency, and location. Upon registration, a blockchain account is created. This data is structured in key/value pairs into a JSON object and uploaded to IPFS, with the related Content Identifier (CID) ensuring that project-related metadata remains immutable and verifiable within the blockchain.
- **User and Project Accounts Linkage** – Once both user and project accounts are created, the system updates their attributes to establish a bidirectional association. This ensures that querying a user account reveals linked project accounts, and vice versa, facilitating traceability and efficient project management.

4.3.2 Scientific Data Management

The d-OSP provides a structured approach to managing research data, ensuring their integrity, traceability, and accessibility. Users can upload various types of research files, including papers, datasets, and images, which are securely stored in a decentralized manner using IPFS. Each uploaded file is assigned a unique Content Identifier (CID), which is recorded on the blockchain, creating a tamper-proof reference. To enhance discoverability, the file metadata is extracted, structured, and stored on IPFS, with its CID also registered on the blockchain. The system further supports indexing and full-text search capabilities, enabling efficient retrieval of research files.

The file upload and metadata management workflow are illustrated in Figure ?? . A user may upload research files, such as papers, datasets, and images, which are stored on IPFS. Each file is assigned a unique CID, ensuring traceability and integrity, and this CID is recorded on the blockchain attributes of the corresponding project account. After upload, metadata is extracted, structured in key/value pairs, and uploaded to IPFS, with its CID also recorded on the blockchain to preserve provenance. To enhance searchability, the system indexes metadata, including full-text indexing for text-based files.

The platform also enables users to search for research data files using keyword-based queries. As depicted in Figure ?? , the search engine looks up keywords in the indexed metadata and returns relevant results. Each result includes metadata details such as descriptions, subjects, and authorship, allowing users to identify relevant files efficiently.

Once a file has been located, the platform performs a validation step to ensure its integrity and authenticity. The CID stored on IPFS is compared against the CID recorded on the blockchain. If they match, the file is considered valid; otherwise, the system flags it as potentially tampered with or corrupted. This validation mechanism safeguards research data against unauthorized modifications. The file validation and retrieval process is depicted in Figure ?? . A validated file can then be retrieved and downloaded from IPFS to the user’s local system for further use.

4.4 Proposed Data Model

The entity-relationship model for the d-OSP defines the logical structure of users and research projects, capturing the associations between these entities. The primary entities in this model are `User` and `Project`, which are connected through a linked relationship, Figure ?? presents the model.

4.4.1 User Entity

The `User` entity represents an individual interacting with the platform. Each user is uniquely identified by an account ID and has attributes that describe personal and institutional information. The attributes of the `User` entity are listed in Table ?? .

Attribute	Description
<code>account_id</code>	A unique identifier assigned to the user.
<code>full_name</code>	The complete name of the user.
<code>email</code>	The email address used for communication.
<code>institution</code>	The organization to which the user is affiliated.
<code>orcid</code>	The Open Researcher and Contributor ID.
<code>role</code>	The role of the user within the research project.
<code>linked_project</code>	The research project the user is assigned to.

Table 4.1: *User Entity Attributes.*

4.4.2 Project Entity

The `Project` entity represents a research project registered in the platform. It contains essential metadata to describe the project and facilitate

discovery and collaboration. The attributes of the `Project` entity are listed in Table ??.

Attribute	Description
<code>project_id</code>	A unique identifier assigned to the project.
<code>name</code>	The official name of the project.
<code>abstract</code>	A brief summary outlining the research objectives.
<code>keywords</code>	A list of relevant keywords associated with the project.
<code>startDate</code>	The date when the project officially begins.
<code>endDate</code>	The date when the project was concluded or is expected to conclude.
<code>city</code>	The city where the project is primarily conducted.
<code>country</code>	The country associated with the research project.
<code>linked_user</code>	The user linked to the project.

Table 4.2: *Project Entity Attributes.*

4.4.3 Linked Entities

A `User` is linked to one or more `Project` entities, establishing a one-to-many relationship. This means that a single user can be associated with multiple projects. This model ensures a structured representation of research projects and their linked users, supporting an organized approach to data management in the d-OSP.

4.5 Hyperledger Iroha Data Model

The entity-relationship (ER) model of Hyperledger Iroha defines the core entities, attributes, and relationships that facilitate role-based access control, asset management, and multi-signature security. While Iroha v1 includes a broader set of entities, this research focuses solely on the account and domain related classes and attributes, as presented in Figure ??

4.5.1 Account Entity

The `account` entity represents a user or system account registered on the blockchain. Table ?? lists its attributes.

Attribute	Description
account_id	Unique identifier of the account
domain_id	Links the account to a specific domain
quorum	Required number of signatories for multi-signature transactions
data	Stores additional metadata in JSON format

Table 4.3: *Attributes of the `account` entity.*

4.5.2 Domain Entity

The `domain` entity organizes accounts within logical boundaries. A `domain` can contain multiple `accounts`, as illustrated in Table ??.

Attribute	Description
domain_id	Unique identifier for the domain
default_role	Default role assigned to accounts created within the domain

Table 4.4: *Attributes of the `domain` entity.*

This ER model follows Hyperledger Iroha’s permissioned blockchain structure. It ensures fine-grained access control, multi-signature security, and domain-based account management.

4.6 Relationship Between Models

The d-OSP ER model leverages the entity structure of the Iroha v1 ER model, particularly the `account` entity, to represent both the `User` and `Project` entities. In this approach, instead of introducing separate entities for users and projects, the `account` entity in the Iroha v1 ER model serves as a general-purpose representation, encapsulating all necessary attributes in a structured format.

The attributes specific to users and projects, which are not natively present in the Iroha v1 `account` entity, are stored as JSON objects within the `data` field of the `account` entity. This design provides a flexible and scalable means of extending the entity’s attributes without modifying the core schema of the Iroha blockchain.

From a relational perspective, the `account` entity maintains its standard associations with roles, permissions, and assets as defined in the Iroha v1 ER model. This ensures that user accounts and project accounts can both participate in the blockchain's permissioning system, asset ownership model, and role-based access control without requiring modifications to the underlying structure.

By reusing the `account` entity, the d-OSP ER model ensures compatibility with Iroha's existing mechanisms for identity management, cryptographic signing, and permission delegation. Additionally, this approach aligns with the decentralized and immutable nature of blockchain, ensuring that both user and project entities benefit from the security and transparency features inherent to the Iroha v1 framework. Figure ?? provides a comparison between models and the rationale of use.

4.7 Metadata and Ontologies

Metadata plays a crucial role in both the `Account` and `Project` classes within the d-OSP. It is used to capture and represent essential attributes about the user and the research project, providing context and structure to their respective data. The metadata is stored as JSON objects, following established semantic web 2.0 standards and leveraging ontologies to enhance data interoperability and accessibility.

4.7.1 Ontologies

An ontology is a formal representation of knowledge as a set of concepts within a domain and the relationships between those concepts. Ontologies help structure data in a way that promotes interoperability, consistency, and clarity. The use of ontologies such as FOAF, Schema.org, and Dublin Core ensures that data is standardized and can be easily shared and understood across different systems. These ontologies were selected because of their widespread adoption, their ability to standardize data across different systems, and their support for rich, machine-readable representations.

As shown in table ??, by aligning with these ontologies, the platform ensures that its metadata is compatible with other Open Science initiatives and services, facilitating seamless integration and data exchange.

4.8 Storing User and Project Attributes

The two core entities, user and project, are tightly integrated with the blockchain and IPFS through metadata structures. This coupling enables per-

Ontology	Description
FOAF (Friend of a Friend)	A vocabulary used to describe people, their activities, and their relationships to other people and objects. It is used to describe the <code>User</code> entity, including attributes like name, email, and organization.
Schema.org	A collaborative initiative that provides a structured vocabulary for data markup on the web. It is used for describing both <code>User</code> and <code>Project</code> metadata, ensuring compatibility with web standards and promoting data discoverability.
Dublin Core (DC)	A metadata standard used for describing a wide range of resources, for describing the abstract, keywords, and other descriptive elements of the <code>Project</code> entity.

Table 4.5: *Ontologies.*

sistent identification, traceability, and contextualization of scientific assets across the platform.

4.8.1 User Metadata

The metadata for the `Account` class describes the attributes associated with a user on the platform. This metadata is structured using multiple ontologies, primarily `FOAF` (Friend of a Friend) and `Schema.org`, to provide detailed and interoperable information about the user. The key attributes in the `Account` metadata include the user's name, email, organizational affiliation, unique identifier (ORCID), role, public key, and linked project.

As presented in Table ??, this structured metadata helps ensure the user information is standardized and interoperable across different systems and platforms.

4.8.2 Metadata representation

JSON for Linked Data (JSON-LD) is a lightweight Linked Data format designed to structure and interconnect data on the web using standard JSON. It extends JSON by incorporating semantic web principles, making data more

Attribute	Description
<code>foaf:name</code>	The name of the user.
<code>foaf:mbox</code>	The email address of the user.
<code>foaf:organization</code>	The organization the user is affiliated with, described as an instance of the <code>foaf:Organization</code> class.
<code>schema:identifier</code>	A unique identifier for the user, such as an ORCID identifier.
<code>foaf:holdsAccount</code>	The user's account details, including their role and public key.
<code>schema:linked_project</code>	The project associated with the user.

Table 4.6: *Account Metadata Attributes.*

discoverable, reusable, and machine-readable. JSON-LD achieves this by including a `@context` element, which maps terms to well-defined ontologies, and a `@graph` element, which structures entities and their relationships in a linked data format.

A key advantage of JSON-LD is its compatibility with existing JSON-based systems while enabling seamless integration with the semantic web. By leveraging vocabularies such as Schema.org and Dublin Core, JSON-LD ensures interoperability across diverse platforms and datasets. This makes it particularly useful for Open Science applications, where structured metadata enhances research reproducibility and data sharing.

In the context of the d-OSP, JSON-LD is used to encode metadata for users and research projects, ensuring alignment with widely accepted ontologies. The structured representation enables automatic indexing, metadata enrichment, and semantic search capabilities, facilitating better knowledge discovery and integration within the scientific community.

4.8.3 The User Metadata

The user metadata is structured using two primary ontologies: Friend of a Friend (FOAF) and Schema.org.

The FOAF ontology is used to describe personal and organizational attributes of users within the platform. It provides well-defined properties such

Listing 1 User Metadata

Listing 2 Project Metadata

as `foaf:name` for the user's full name, `foaf:mbox` for email addresses, and `foaf:organization` for institutional affiliations. By leveraging FOAF, the platform ensures standardized representation of user identities and their associations, facilitating integration with other systems that utilize FOAF-based user profiles.

Schema.org complements FOAF by enriching the user metadata with structured properties that enhance discoverability and machine readability. The `schema:identifier` property, for instance, is used to store unique user identifiers such as ORCID, ensuring compatibility with global researcher identification systems. Additionally, `schema:roleName` captures the user's role within the platform (e.g., reviewer, publisher), while `schema:publicKey` stores cryptographic keys associated with the user's account. The `schema:linked_project` property establishes connections between users and their associated research projects, enabling efficient metadata retrieval and knowledge graph construction as exhibited in Figure ??, the JSON-LD structure represents the project metadata in the d-OSP.

By combining FOAF and Schema.org, the d-OSP ensures that user metadata is both human-readable and machine-actionable, promoting seamless integration with external research infrastructures and fostering an interoperable ecosystem for Open Science.

4.8.4 The Project Metadata

The metadata for the `Project` entity provides essential details about the research project hosted on the platform. Similar to the `User` metadata, the project metadata is structured using Schema.org and Dublin Core (dc) ontologies. This structure allows for a comprehensive description of the project, including its name, abstract, keywords, timeline, funding details, and location.

The following JSON structure describes the metadata for a `Project` in the d-OSP as shown in Figure ??.

This metadata not only captures the essential details of the project but also ensures that these details are linked to the user's profile, making it easier to track the relationship between users and their associated research efforts.

Attribute	Description
<code>schema:name</code>	The name of the research project.
<code>dc:abstract</code>	A brief abstract describing the project's objectives and focus.
<code>schema:keywords</code>	Keywords associated with the project, such as "precision agriculture" and "global supply chains."
<code>schema:startDate</code>	The start date of the project.
<code>schema:endDate</code>	The end date of the project.
<code>schema:funding</code>	The funding organization for the project, described as an instance of the <code>schema:Organization</code> class.
<code>schema:location</code>	The physical location where the project is based, described as an instance of the <code>schema:Place</code> class.
<code>schema:metadataCID</code>	A unique identifier for the metadata of the project.
<code>schema:linked_user</code>	The user associated with the project.

Table 4.7: Project Metadata Attributes.

4.8.5 Basic Metadata Workflow

The d-OSP implements a structured approach to metadata handling to ensure that all metadata is semantically interoperable, verifiable, and tamper-evident. Metadata may describe entities such as users, projects, or files and is serialized using the JSON-LD standard, aligning with established ontologies like Dublin Core (DCT), FOAF, and PROV-O.

Once generated, the metadata object is stored on the IPFS, which returns a CID derived from the cryptographic hash of the content. This CID is then written into the account details of the corresponding entity on the blockchain, ensuring that metadata references are immutable, timestamped, and auditable.

This blockchain-anchored CID acts as a provenance record, enabling stakeholders to verify the authenticity and trace the history of metadata. Indexed metadata is then made discoverable through the platform's index and search system, promoting transparency, accessibility, and reproducibility across research projects.

Figure ?? illustrates the sequence of operations in the basic metadata workflow.

4.8.6 Extended Metadata Workflow

The extended metadata workflow in d-OSP expands the basic model to support provenance tracking between related research artifacts. This feature is essential for ensuring that derivative files can be cryptographically linked to their source, while preserving the integrity and discoverability of both.

4.8.6.1 Original Artifact Workflow (Researcher A)

Researcher A begins by uploading an original artifact, such as a dataset or image. Metadata describing the artifact, such as `dct:creator`, `dct:created`, project affiliation, and descriptive annotations, is either manually provided or automatically extracted and serialized as JSON-LD.

The metadata object is submitted to IPFS, which returns a CID uniquely identifying the artifact. This CID is anchored on the blockchain via a transaction tied to the relevant account, ensuring that the record is immutable and verifiable. Once processed, the metadata is indexed and becomes accessible through the platform's search interface.

4.8.6.2 Derived Artifact Workflow (Researcher B)

At a later stage, Researcher B discovers the original artifact and downloads it for reuse. After modifying the content, Researcher B uploads a derived artifact to their own project. A new JSON-LD metadata object is generated and submitted to IPFS, resulting in a new CID.

To preserve lineage, the system automatically includes a `prov:wasDerivedFrom` assertion in the new metadata, linking the derived CID to the source CID. This provenance link, along with metadata such as `dct:creator` and `dct:created`, is recorded on the blockchain. No transformation details are stored, but the link between artifacts is cryptographically verifiable, system-enforced, and queryable.

This extended model ensures that all derivative relationships between artifacts are transparently and immutably captured, enabling robust provenance tracking across decentralized research workflows.

Figure ?? illustrates the extended metadata workflow with lineage tracking.

Listing 3 Blockchain Representation of User Account

4.8.7 Blockchain Representation

In the d-OSP, metadata for users, projects, and files are stored on the blockchain. This ensures the integrity and provenance of the metadata while leveraging decentralized technologies. The following subsections describe the structure of blockchain representations for both user and project data, as well as the files associated with these projects.

4.8.8 User Account

The representation of a user's account on the blockchain contains the standard Iroha v1 attributes for the account entity, such as the unique account identifier, domain information, and quorum for consensus. Additionally, the `json_data` attribute references both the project to which the user is linked and the user's metadata CID (Content Identifier) stored on IPFS. This blockchain-based approach ensures that the user's information remains immutable and traceable, which is critical for maintaining the integrity of research data.

Figure ?? presents the JSON structure for account details for a user in the Iroha v1 blockchain.

4.8.9 Project Account

The project account representation similarly uses a blockchain-based structure to store project-related metadata. Each project is identified by a unique account ID, along with the project's domain and quorum. The project metadata is linked to the user and includes important information about files associated with the project, including their CID references on IPFS. This ensures that the project data is linked to the user's account and that all files and metadata related to the project are securely stored on the blockchain for provenance tracking.

The JSON structure describes the account details for a project in the Iroha v1 blockchain as shown in Figure ??.

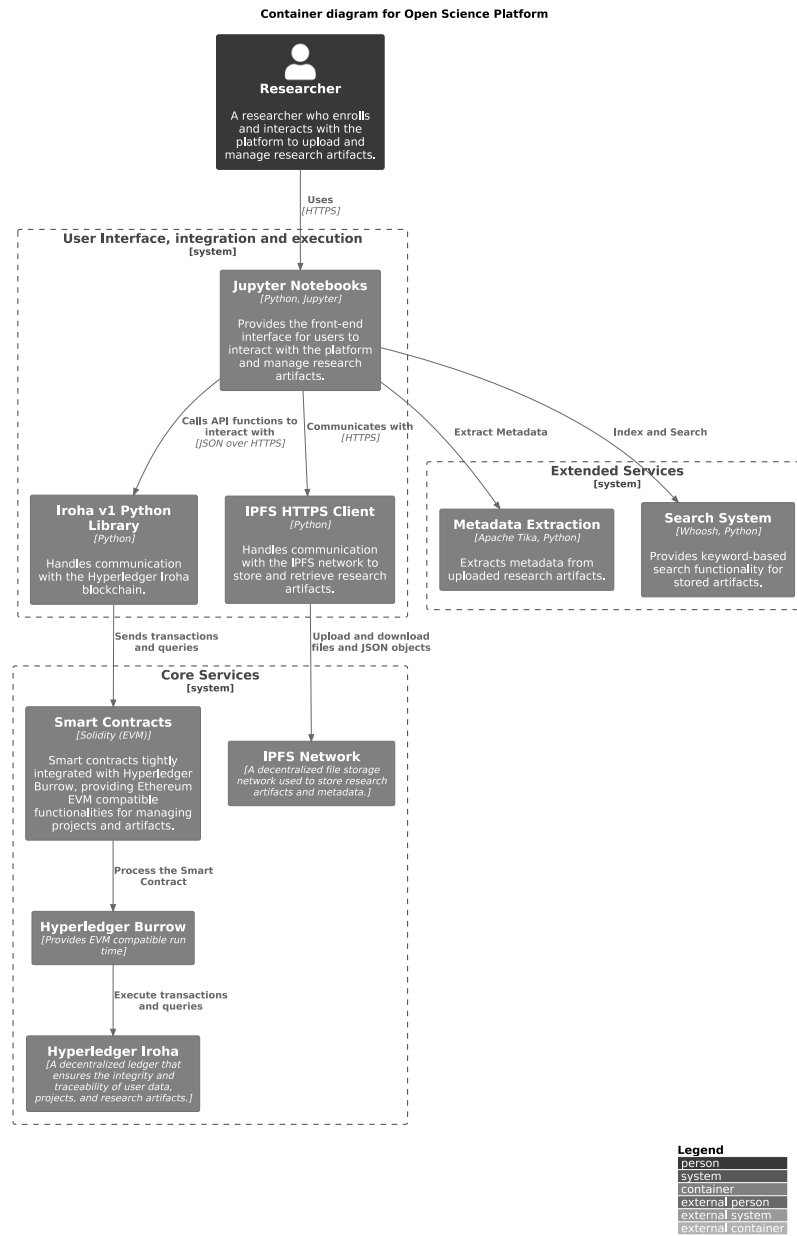


Figure 4.2: Container diagram.

Listing 4 Blockchain Representation of Project Account

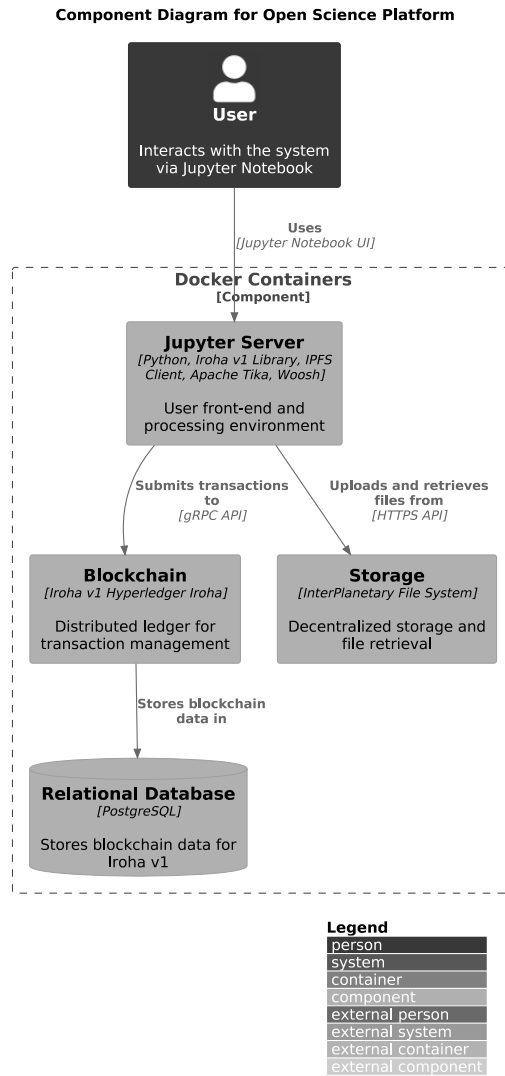


Figure 4.3: Component diagram.

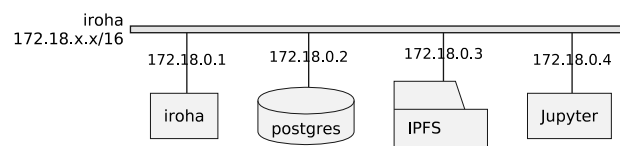


Figure 4.4: Docker network topology.

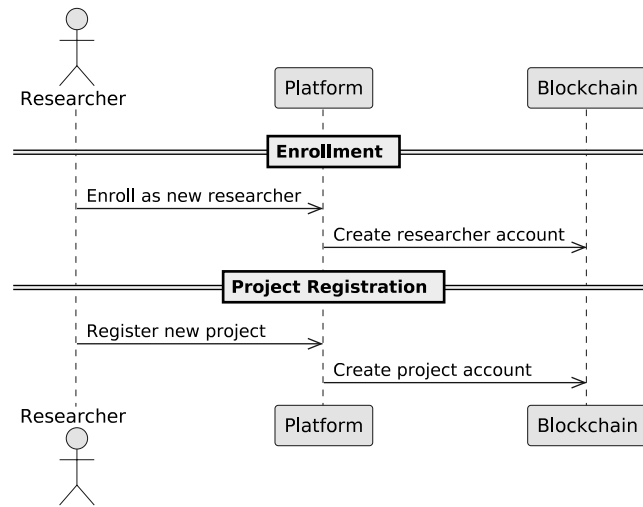


Figure 4.5: User enrollment and project registering.

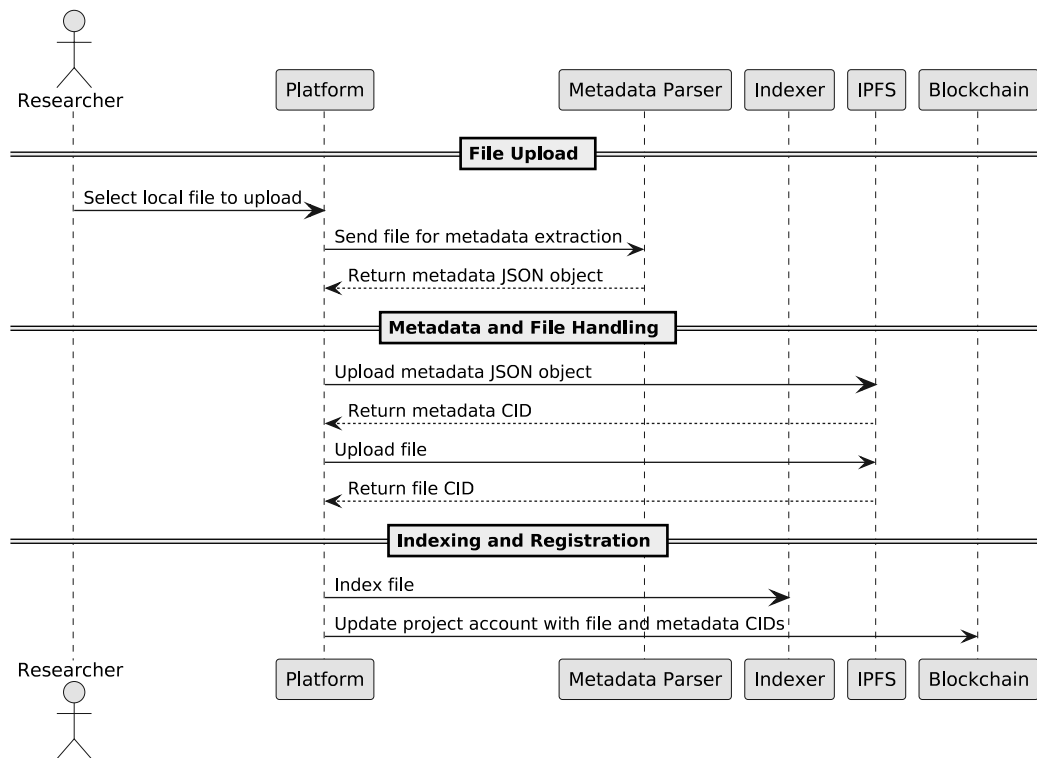


Figure 4.6: File operations diagram.

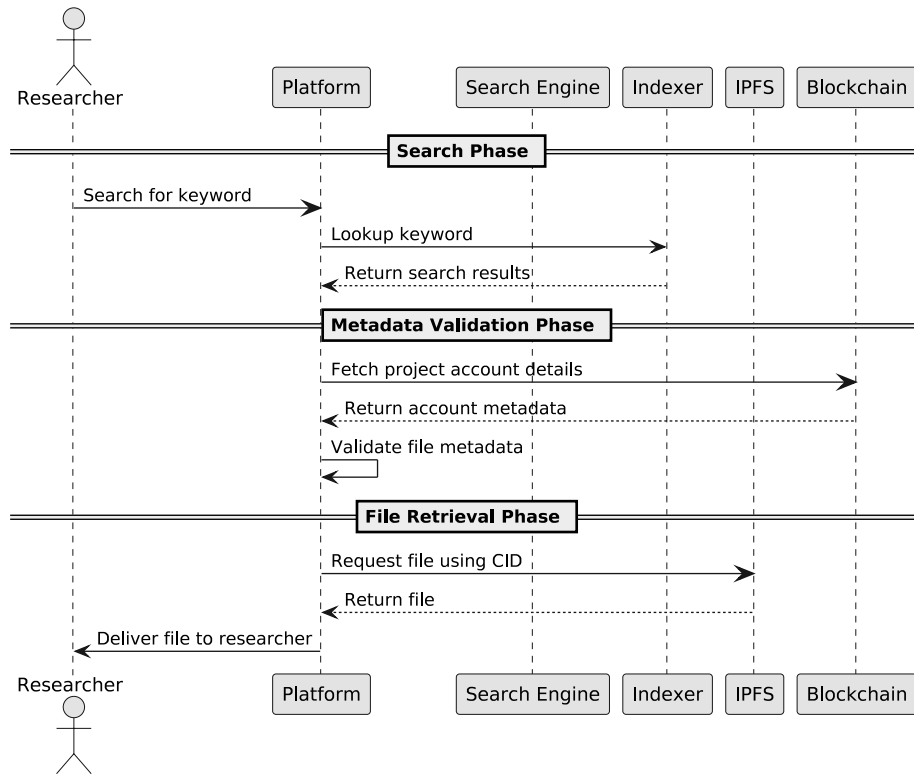


Figure 4.7: Keyword search, file validation, and download.

4.8.10 File Representation

Within the project account, each file associated with the project is represented by a CID pair. The first CID refers to the file stored on IPFS, while the second CID references the metadata associated with that file. This structure ensures that the file's content and its metadata are both stored and tracked independently, but are still linked to the blockchain for integrity and provenance.

- The **first CID** (QmTLZSqzPexwEdniZXLPN6fUfmEXX6MXS3b4QjKURgxc9y) corresponds to the **file**.
- The **second CID** (Qmchg7At5whR1T4xP8TwTMd8ntQqJXbbSicJRtGGaW1Z2P) corresponds to the **metadata of the file**, ensuring that all relevant details are retrievable.

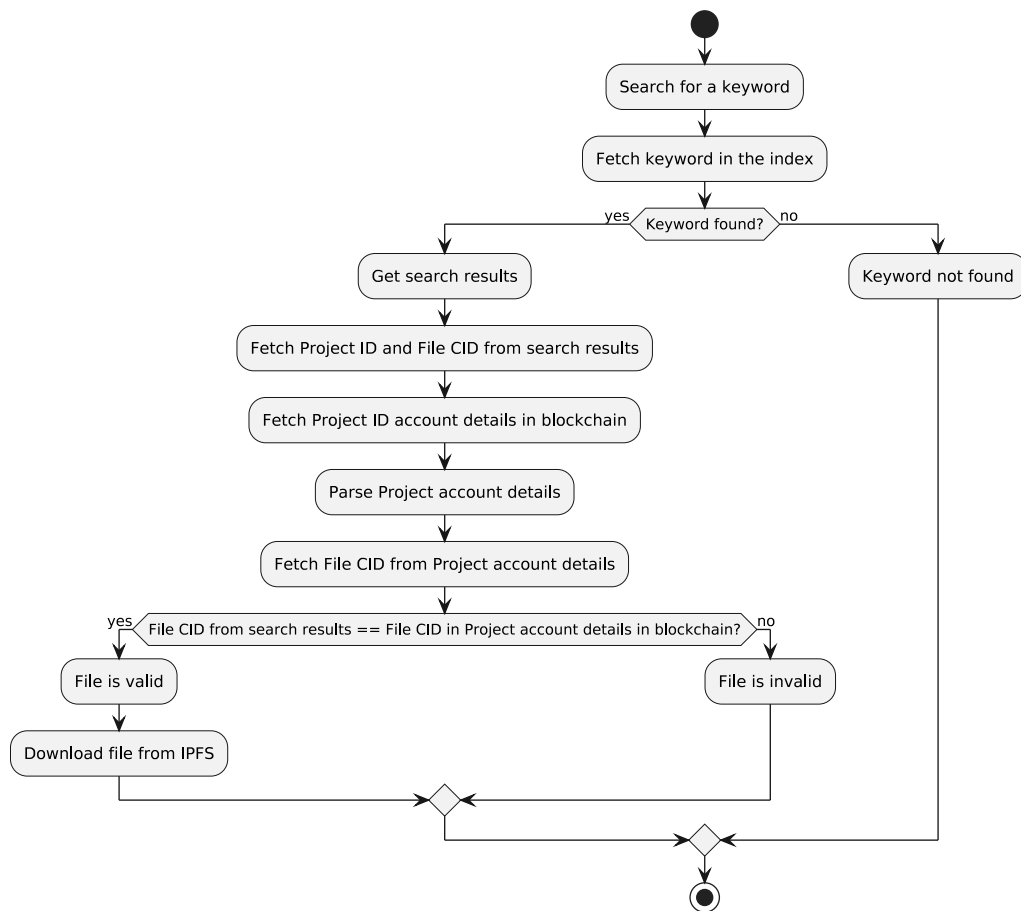
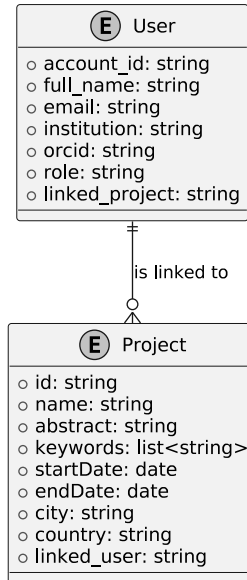
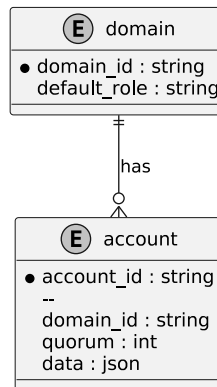


Figure 4.8: *File validation and download.*

This structure allows for the efficient tracking and retrieval of research project data while maintaining provenance and integrity through blockchain storage.

4.9 Provenance

The provenance system takes a two-fold approach, with both methods being native features of their respective systems. The first approach leverages Iroha v1's transaction logging capabilities, where each transaction is recorded with a hexadecimal hash and timestamp. This provides a reliable mechanism for tracking the evolution of account states over time. The hash acts as a snapshot, allowing for the retrieval of any past state of an account based on the corresponding transaction hash, as depicted in in Figure ??.

Open Science Platform Entity-relationship model**Figure 4.9:** *Entity-relationship model.***Hyperledger Iroha v1 Entity-relationship model(subset)****Figure 4.10:** *Subset of the Iroha v1 Entity-relationship model.*

The second approach makes use of IPFS's native feature of Content Identifiers (CIDs) to track metadata associated with accounts, projects, and files. Each piece of metadata is linked to a unique CID, which allows for decentralized storage and immutability. A mismatch of the CID indicates that the metadata or file has been modified, ensuring the integrity of the information over time.

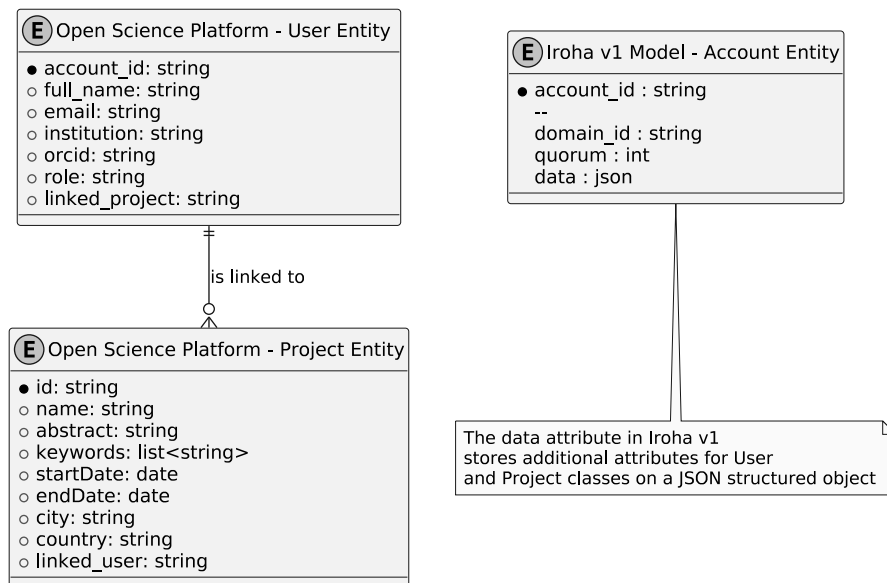


Figure 4.11: Comparison of the Entity-relationship models.

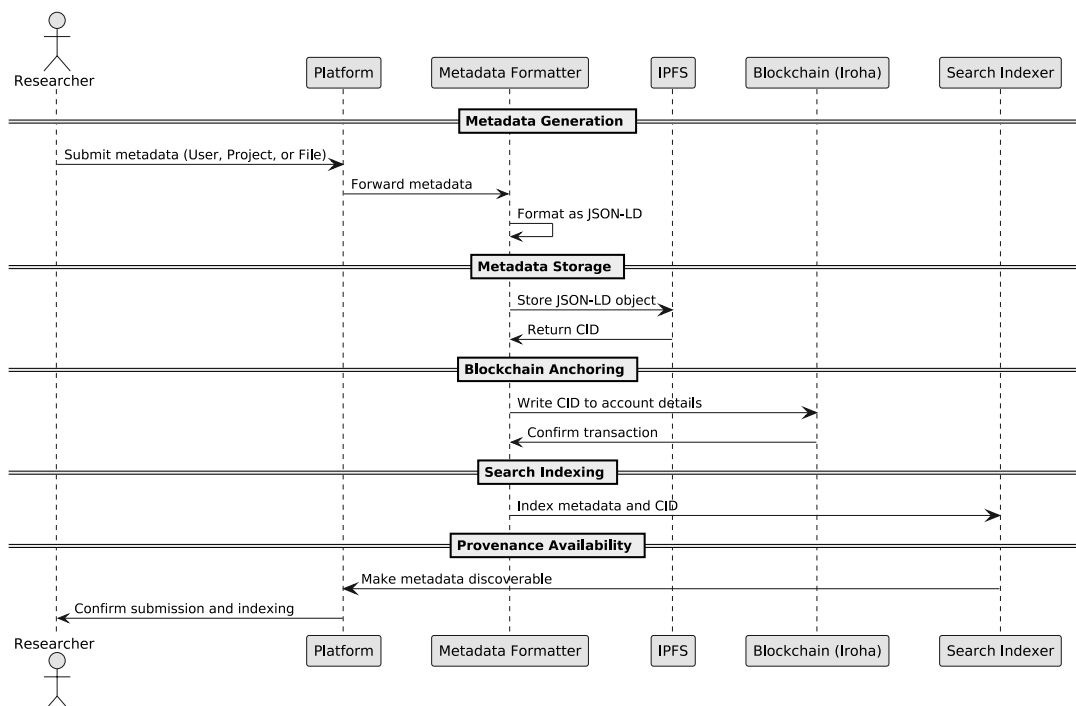


Figure 4.12: Basic metadata workflow.

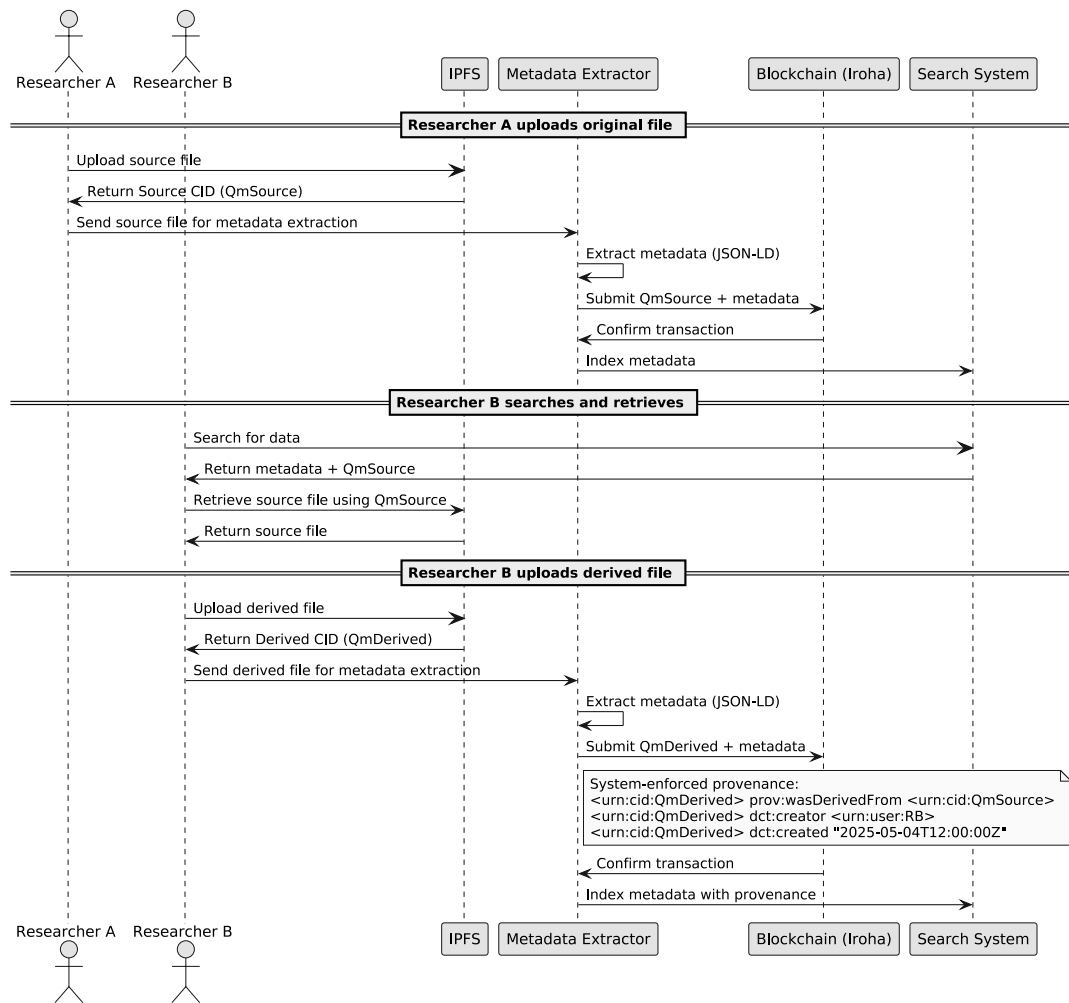


Figure 4.13: *Extended metadata workflow.*

Together, these two approaches, transaction logging through Iroha v1's blockchain and metadata tracking through IPFS CIDs, provide a robust and transparent provenance system, ensuring both the transaction history and the integrity of metadata are verifiably recorded and traceable.

4.10 Summary

The d-OSP integrates a cohesive technology stack to promote reproducibility and transparency in research. The Iroha v1 blockchain manages user and project accounts, while smart contracts executed via Hyperledger Burrow automate operations and enforce trust. IPFS provides decentralized,

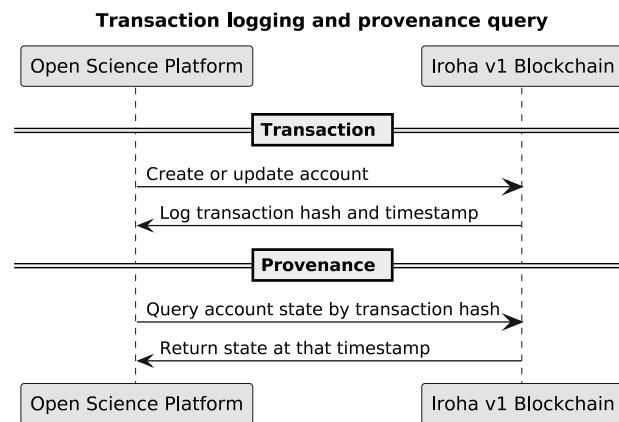


Figure 4.14: *Transaction logging and provenance query.*

persistent storage for research data. Jupyter Notebooks serve as the primary user interface, supporting interactive workflows. Apache Tika extracts metadata from diverse formats, and Woosh enables advanced search and indexing. Users and projects are registered as blockchain accounts, supporting secure role management and traceability. Metadata is structured using FOAF, Schema.org, and Dublin Core ontologies to ensure standardization and interoperability. An ER model defines the platform's data structure, while all metadata transactions are transparently recorded on-chain. This enables robust provenance tracking and reinforces the integrity of scientific outputs.

Chapter 5

Conclusions and Future Work

*Science is one of the very few human activities,
perhaps the only one,
where errors are systematically criticized and fairly often, in time, corrected.*

Karl Popper, British Philosopher (1902–1994)

The ongoing reproducibility crisis has underscored the structural limitations of contemporary scientific practices, revealing systemic issues in transparency, accountability, and data accessibility. While Open Science has emerged as a response to these challenges, its implementation often remains constrained by centralized infrastructures that inadequately support verifiability and long-term data integrity. In contrast, the foundational tenets of decentralized technologies, immutability, distributed trust, provenance tracking, and censorship resistance, offer a compelling architectural framework to address these concerns. By integrating blockchain, smart contracts, and IPFS into a cohesive platform, it becomes possible to enforce transparent, tamper-evident records of research processes, enable automated compliance with reproducibility protocols, and ensure persistent access to underlying data and metadata.

This dissertation has demonstrated how such an integration can support a more trustworthy and verifiable scientific ecosystem, grounded in the principles of Open Science. The decentralized Open Science Platform (d-OSP) presented here exemplifies this potential, offering a technically grounded and ontologically structured response to reproducibility challenges. Although barriers remain, ranging from legal and infrastructural to usability and interoperability, the progress outlined in this work opens promising pathways forward. As this research draws to a close, the next phase will focus on

advancing the platform's capabilities, ensuring broader adoption, and deepening its alignment with the evolving Open Science landscape. In doing so, this work aims to contribute meaningfully to reshaping the foundations of scientific practice in the digital age.

This work has introduced a decentralized Open Science Platform (d-OSP) designed to address persistent challenges in scientific reproducibility and open research collaboration. By building on the integrated use of blockchain, IPFS, and smart contracts, the platform offers a cohesive technological foundation to support more transparent, verifiable, and collaborative research practices.

Reproducibility crises in science have often stemmed from a lack of traceable provenance, inaccessible data, opaque methodologies, and unreliable long-term availability of research artifacts. Centralized infrastructures tend to reinforce these issues by restricting control, limiting visibility, and introducing single points of failure. In contrast, decentralized technologies offer architectural characteristics that directly counter these systemic weaknesses.

Blockchain contributes immutability, auditability, and decentralized governance, ensuring that research records, once registered, are tamper-proof and independently verifiable. This integrity is fundamental to preserving the trustworthiness of experimental logs, provenance metadata, and contributions over time. Moreover, the use of smart contracts facilitates automated enforcement of verification processes and access control, providing reproducible workflows without dependence on centralized intermediaries.

IPFS complements this by enabling distributed and content-addressed storage, making research outputs, datasets, notebooks, protocols available across a peer-to-peer network. This ensures persistent accessibility and efficient dissemination, both crucial for reproducibility and knowledge transfer.

Together, these technologies do not merely offer individual improvements, but form an integrated and synergistic approach to fulfilling the core propositions of open science: transparency, accessibility, verifiability, and collaboration. By enabling trustless interactions, distributed storage, and automated protocol enforcement, the d-OSP positions itself as a natural technical response to the reproducibility problem.

Looking ahead, the platform presents ample opportunities for continued development and refinement. Future work will focus on expanding artifact management capabilities, enhancing metadata extraction and search

functionalities through deeper integration with knowledge graphs, and enabling support for complex, multi-source research data workflows. Research into domain-specific ontologies for scientific data will play a key role, alongside efforts to decentralize the deployment of the metadata parser and search engine. To foster a more engaged and collaborative research community, blockchain-based incentive mechanisms will be introduced to reward valuable contributions such as data sharing, result verification, and metadata enrichment. Other priorities include improving the overall user experience and implementing smart contract support for domain-specific logic, both of which are critical to the broader adoption and long-term impact of the platform.

In summary, the integration of decentralized technologies into scientific infrastructure is not an arbitrary technical upgrade, but a strategically aligned response to deeply rooted problems in reproducibility and openness. The d-OSP lays the groundwork for a future in which scientific knowledge is not only more accessible, but inherently more trustworthy and resilient.

Bibliography

- [] M. Alharby and A. van Moorsel. *Blockchain-based smart contracts: A systematic mapping study*. *Computer Science & Information Technology*, pages 125–140, 2017.
- [] A. Alkhateeb, C. Catal, G. Kar, and A. Mishra. *Hybrid blockchain platforms for the internet of things (iot): A systematic literature review*. *Sensors*, 22(4), 2022.
- [] M. Baker. *1,500 scientists lift the lid on reproducibility*. *Nature*, 533 (7604):452–454, 2016.
- [] D. Bayer, S. Haber, and W. S. Stornetta. *Improving the efficiency and reliability of digital time-stamping*. *Sequences II: Methods in Communication, Security, and Computer Science*, pages 329–334, 1993.
- [] J. Benet. *Ipfs - content addressed, versioned, p2p file system*. *arXiv preprint*, arXiv:1407.3561, 2014.
- [] J. Benet. *Ipfs - the interplanetary file system*, 2015. URL: <https://ipfs.io>.
- [] J. Benet. *IPFS - Content Addressed, Versioned, P2P File System (DRAFT 3)*, November 2022. URL: <https://docs.ipfs.tech/concepts/further-reading/academic-papers/#ipfs-content-addressed-versioned-p2p-file-system>. Last accessed on 30 Nov 2022.
- [] F. Berman and M. Crosas. *The Research Data Alliance: Benefits and Challenges of Building a Community Organization*. *Harvard Data Science Review*, January 2020.
- [] D. Bernstein, T. Lange, and R. Niederhagen. *High-speed high-security signatures*. *Journal of cryptographic engineering*, 2(2):77–89, 2012.

- [] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, and E. W. Felten. [Sok: Research perspectives and challenges for bitcoin and cryptocurrencies](#). *Proceedings - IEEE Symposium on Security and Privacy*, pages 104–121, 2015.
- [] C. L. Borgman. *The conundrum of sharing data*. *Journal of the American Society for Information Science and Technology*, 63(1):105–118, 2012.
- [] G. Boulton, L. Allen, and M. H. Brooks. *Open science: A new era for the future of research*. *Science*, 348(6237):29–32, 2015.
- [] M. N. Branch. [The “Reproducibility Crisis:” Might the Methods Used Frequently in Behavior-Analysis Research Help?](#) *Perspect Behav Sci*, 42(1):77–89, March 2019.
- [] J. Brase. [DataCite - A Global Registration Agency for Research Data](#). In *2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, pages 257–261, Beijing, China, November 2009. IEEE.
- [] D. Brown. *Standards for efficient cryptography (sec) 1: Elliptic curve cryptography*. Technical report, Certicom Research, 2010.
- [] M. Budiu and B. J. Schwartz. *Designing distributed systems: A service-oriented approach*. Prentice Hall, 2007.
- [] C. Cachin. [Architecture of the hyperledger blockchain fabric](#). In *Workshop on Distributed Cryptocurrencies and Consensus Ledgers (DCCL)*, 2016.
- [] C. Cachin. [Blockchains: The magic of cryptographic protocols](#). *IEEE Security & Privacy*, 14(2):56–62, 2016.
- [] M. Castro and B. Liskov. [Practical byzantine fault tolerance](#). In *Proceedings of the Third Symposium on Operating Systems Design and Implementation (OSDI)*, pages 173–186, 1999.
- [] F. S. Collins and L. A. Tabak. [Policy: NIH plans to enhance reproducibility](#). *Nature*, 505(7485):612–613, January 2014.
- [] Committee on Reproducibility and Replicability in Science, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, Division of Behavioral and Social Sciences and Education,

- Nuclear and Radiation Studies Board, Division on Earth and Life Studies, Board on Mathematical Sciences and Analytics, Committee on Applied and Theoretical Statistics, Division on Engineering and Physical Sciences, Board on Research Data and Information, Committee on Science, Engineering, Medicine, and Public Policy, Policy and Global Affairs, and National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press, Washington, D.C., September 2019.
- [] G. Coulouris, J. Dollimore, T. Kindberg, and J. Blair. *Distributed systems: Concepts and design*. Addison-Wesley, 2011.
 - [] David Robert Grimes, D. R. Grimes, Chris T. Bauch, C. T. Bauch, John P. A. Ioannidis, and J. P. A. Ioannidis. *Modelling science trustworthiness under publish or perish pressure*. *Royal Society Open Science*, 5 (1):171511–171511, January 2018.
 - [] D. Easley, M. O’Hara, and S. Basu. *From mining to markets: The evolution of bitcoin transaction fees*. *Journal of Financial Economics*, 134: 91–109, 2019.
 - [] C. Fan, Y. Chen, R. Chen, H. Zhou, and L. Chen. *Analysis of elliptic curve cryptography for blockchain*. *Future Generation Computer Systems*, 89:735–743, 2018.
 - [] E. D. Foster and A. Deardorff. *Open Science Framework (OSF)*. *jmla*, 105(2), April 2017.
 - [] H. Foundation. *Hyperledger fabric now supports ethereum*, 2018. URL: <https://www.hyperledger.org/blog/2018/10/26/hyperledger-fabric-now-supports-ethereum>. Accessed: 2025-03-17.
 - [] H. Foundation. *Burrow - the boring blockchain*, 2019. URL: <https://www.hyperledger.org/blog/2019/10/08/burrow-the-boring-blockchain>. Accessed: 2025-03-17.
 - [] H. Foundation. *Iroha special instructions dsl*, 2020. URL: <https://lf-hyperledger.atlassian.net/wiki/spaces/iroha/pages/21012073/Iroha+Special+Instructions+DSL>. Accessed: 2025-03-17.
 - [] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe. *The Economics of Reproducibility in Preclinical Research*. *PLOS Biology*, 13(6):e1002165, June 2015.

- [] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe. [The economics of reproducibility in preclinical research](#). *PLoS Biology*, 13(6):e1002165, 2015.
- [] A. Gazis, G. Anagnostakis, S. Kourmpetis, and E. Katsiri. [A Blockchain Cloud Computing Middleware for Academic Manuscript Submission](#). *Wseas Transactions On Business And Economics*, 2022.
- [] S. Haber and W. S. Stornetta. [How to time-stamp a digital document](#). *Journal of Cryptology*, 3(2):99–111, 1991.
- [] D. Hankerson, A. Menezes, and S. Vanstone. *Guide to elliptic curve cryptography*. Springer Science & Business Media, 2006.
- [] P. C. Henning, C. J. S. Ribeiro, L. O. B. Da Silva Santos, and P. X. Dos Santos. [GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta](#). *EQ*, pages 389–412, April 2019.
- [] A. R. Hevner, S. T. March, J. Park, and S. Ram. *Design science in information systems research*. *MIS Quarterly*, 28(1):75–105, 2004.
- [] M. Holmen. [Blockchain and scholarly publishing could be best friends](#). *ISU*, 38(3):131–140, October 2018.
- [] Hyperledger Foundation. [Hyperledger iroha documentation](#), 2025. Accessed: 2025-03-17.
- [] J. P. A. Ioannidis. [Why most published research findings are false](#). *PLoS Medicine*, 2(8):e124, 2005.
- [] D. Johnson, A. Menezes, and S. Vanstone. *The elliptic curve digital signature algorithm (ecdsa)*. *International Journal of Information Security*, 1(1):36–63, 2001.
- [] J. Katz and Y. Lindell. *Introduction to modern cryptography*. Chapman and Hall/CRC, edition 3rd, 2020.
- [] A. Kiayias, A. Russell, B. David, and R. Oliynykov. [Ouroboros: A provably secure proof-of-stake blockchain protocol](#). In J. Katz and H. Shacham, editors, *ADVANCES IN CRYPTOLOGY - CRYPTO 2017, PT I*, volume 10401 of *Lecture Notes in Computer Science*, pages 357–388. Int Assoc Cryptol Res; NSF; Fujitsu; Western Digital; Mozilla, 2017. 37th Annual International Cryptology Conference (Crypto), Univ Calif, Santa Barbara, CA, AUG 20-24, 2017.

- [] N. Koblitz. *Elliptic curve cryptosystems*. *Mathematics of computation*, 48(177):203–209, 1987.
- [] D. Kochalko. *Making the unconventional conventional: How blockchain contributes to reshaping scholarly communications*. *ISU*, 39(3):199–204, December 2019.
- [] A. Kosmarski. *Blockchain Adoption in Academia: Promises and Challenges*. *Journal of Open Innovation: Technology, Market, and Complexity*, 2020.
- [] L. Lamport. *Time, clocks, and the ordering of events in a distributed system*. *Commun. ACM*, 21(7):558–565, July 1978.
- [] L. Lamport, R. Shostak, and M. Pease. *The byzantine generals problem*. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.
- [] S. C. Landis, S. G. Amara, K. Asadullah, C. P. Austin, R. Blumenstein, E. W. Bradley, R. G. Crystal, R. B. Darnell, R. J. Ferrante, H. Fillit, R. Finkelstein, M. Fisher, H. E. Gendelman, R. M. Golub, J. L. Goudreau, R. A. Gross, A. K. Gubitz, S. E. Hesterlee, D. W. Howells, J. Huguenard, K. Kelner, W. Koroshetz, D. Krainc, S. E. Lazic, M. S. Levine, M. R. Macleod, J. M. McCall, R. T. Moxley III, K. Narasimhan, L. J. Noble, S. Perrin, J. D. Porter, O. Steward, E. Unger, U. Utz, and S. D. Silberberg. *A call for transparent reporting to optimize the predictive value of preclinical research*. *Nature*, 490(7419):187–191, October 2012.
- [] A. Langley, M. Hamburg, and M. Naehrig. *Curve25519: new diffie-hellman speed records*. *International Association for Cryptologic Research*, 2016.
- [] B. Lawlor. *An overview of the NFAIS conference: Blockchain for scholarly publishing*. *ISU*, 38(3):111–130, October 2018.
- [] L. A. R. N. (LEARN). *LEARN Toolkit of Best Practice for Research Data Management*. Technical report, LEARN, April 2017.
- [] J. Lee, M. Moroso, and T. K. Mackey. *Unblocking recognition: A token system for acknowledging academic contribution*. *Frontiers in blockchain*, 2023.
- [] S. Leible, S. Schlager, M. Schubotz, and B. Gipp. *A Review on Blockchain Technology and Blockchain Projects Fostering Open Science*. *Frontiers in blockchain*, 2019.

- [] S. Leonelli. *The practices and politics of open science*. In *Springer Handbook of Science and Technology Studies*, pages 1–19. Springer, 2016.
- [] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone. *Handbook of applied cryptography*. CRC Press, Boca Raton, 1996.
- [] R. Merkle. [A certified digital signature](#). *Proceedings of the 2nd ACM Conference on Computer and Communications Security*, pages 24–34, 1989.
- [] R. C. Merkle. [A digital signature based on a conventional encryption function](#). In *Advances in Cryptology — CRYPTO '87*, volume 293, pages 369–378. Springer Berlin Heidelberg, Berlin, Heidelberg, 1988.
- [] R. Merkle. *A digital signature based on a conventional encryption function*. *Advances in cryptology—CRYPTO'87*, pages 369–378, 1988.
- [] A. Miller, D. S. Rowe, and A. Papageorgiou. [Scaling blockchain storage and its applications](#). *Journal of Computer Science and Technology*, 31(4):873–886, 2016.
- [] V. Miller. *Use of elliptic curves in cryptography*. *Advances in cryptology—CRYPTO'85*, pages 417–426, 1986.
- [] Monax. [Hyperledger burrow](#), 2020. URL: <https://www.npmjs.com/package/@monax/burrow>. Accessed: 2025-03-17.
- [] H. Moulaison-Sandy. [The Nelson Memo and US Federal Funder Requirements for Public Access: Implications for Technical Services Librarians](#). *Technical Services Quarterly*, 40(4):290–297, October 2023.
- [] M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. P. d. Sert, U. Simonsohn, E. J. Wagenmakers, J. Ware, and J. P. A. Ioannidis. [A manifesto for reproducible science](#). *Nature Human Behaviour*, 2017.
- [] M. R. Munafò, B. A. Nosek, D. V. M. Bishop, K. S. Button, C. D. Chambers, N. Sert, and J. P. A. Ioannidis. [A manifesto for reproducible science](#). *Nature Human Behaviour*, 1:0021, 2017.
- [] F. Muratov, A. Lebedev, N. Iushkevich, B. Nasrulin, and M. Takemiya. [YAC: BFT Consensus Algorithm for Blockchain](#), September 2018. URL: <https://arxiv.org/abs/1809.00554>. arXiv:1809.00554 [cs].

- [] S. Nakamoto. *Bitcoin: A peer-to-peer electronic cash system*, 2008. URL: <https://bitcoin.org/bitcoin.pdf>.
- [] S. Nakamoto. *Bitcoin: A peer-to-peer electronic cash system*, 2009. URL: <http://www.bitcoin.org/bitcoin.pdf>.
- [] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder. *Bitcoin and cryptocurrency technologies: a comprehensive introduction*. Princeton university press, Princeton (N.J.), 2016.
- [] B. A. Nosek, M. E. Ebersole, K. M. DeHaven, and D. L. Mellor. *The preregistration revolution*. *Proceedings of the National Academy of Sciences*, 112(5):1–8, 2015.
- [] N. I. of Standards and Technology. *Fips pub 180-4: Secure hash standard (shs)*, 2012. URL: <https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf>. Accessed: 2025-03-17.
- [] E. Oliveira, R. Z. Frantz, C. Molina-Jiménez, T. Heck, S. Sawicki, and F. Roos-Frantz. *On the use of blockchain technology to improve the reproducibility of preclinical research experiments*. In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023)*, pages 173–179. INSTICC, SciTePress, 2023.
- [] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher. *The prisma 2020 statement: an updated guideline for reporting systematic reviews*. *BMJ*, 372, 2021.
- [] E. D. Pellizzari, K. N. Lohr, A. R. Blatecky, and D. R. Creel. *Reproducibility: A Primer on Semantics and Implications for Research*. RTI Press, 2017.
- [] R. D. Peng. *Reproducible research in computational science*. *Science*, 334(6060):1226–1227, 2011.
- [] N. Percie du Sert, A. Ahluwalia, S. Alam, M. T. Avey, M. Baker, W. J. Browne, and others. *Reporting animal research: Explanation and elaboration for the arrive guidelines 2.0*. *PLOS Biology*, 18(7):e3000411, 2020.

- [] S. Pilehchiha. *Improving reproducibility in smart contract research*. Master's thesis, Concordia University, 2022. Accessed: 2025-03-17.
- [] H. A. Piwowar. *Who shares? who doesn't?* *PLoS ONE*, 6(7):e18657, 2011.
- [] M. Putnings. *Non-Fungible Token (NFT) in the Academic and Open Access Publishing Environment: Considerations Towards Science-Friendly Scenarios*. *Journal of Electronic Publishing*, 2022.
- [] N. Rettberg and B. Schmidt. *OpenAIRE - Building a collaborative Open Access infrastructure for European researchers*. *LIBER*, 22(3):160–175, November 2012.
- [] R. Rivest, A. Shamir, and L. Adleman. *A method for obtaining digital signatures and public-key cryptosystems*. *Communications of the ACM*, 21(2):120–126, 1978.
- [] T. Roughgarden. *Twenty-one lectures on algorithmic game theory*. Cambridge University Press, 2016.
- [] S. Samuel and B. König-Ries. *Understanding experiments and research practices for reproducibility: an exploratory study*. *PeerJ*, 9: e11140, April 2021.
- [] J. Sedlmeir, H. U. Buhl, G. Fridgen, and R. Keller. *The Energy Consumption of Blockchain Technology: Beyond Myth*. *Bus Inf Syst Eng*, 62(6):599–608, December 2020.
- [] D. Selematsela, U. Mujoomunshi, and G. Yadav. *Data for development and the role of codata of the international science council (isc)*. *ANNALS OF LIBRARY AND INFORMATION STUDIES*, 71(4):506–511, DEC 2024.
- [] Y. Shevchenko. *Open Lab: A web application for running and sharing online experiments*. *Behav Res*, 54(6):3118–3125, March 2022.
- [] Simplilearn. *What is a smart contract in blockchain and how does it work?*, 2022. URL: <https://www.simplilearn.com/tutorials/blockchain-tutorial/what-is-smart-contract>. Accessed: 2025-03-17.
- [] P. B. Stark. *Before reproducibility must come preproducibility*. *Nature*, 557(7705):613, 2018.

- [] M. Steichen, B. Fiz, R. Norvill, W. Shbair, and R. State. *Blockchain-based, decentralized access control for ipfs*. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 1499–1506, 2018.
- [] N. Szabo. *Smart contracts*, January 1994. URL: <https://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart.contracts.html>. Last accessed on 28 Jan 2021.
- [] N. Szabo. *Smart contracts: Building blocks for digital markets*. *EXTROPY: The Journal of Transhumanist Thought*, 16, 1996.
- [] N. Szabo. *Formalizing and securing relationships on public networks*. *First Monday*, 2(9), 1997.
- [] D. Tapscott and A. Tapscott. *Blockchain revolution: How the technology behind bitcoin and other cryptocurrencies is changing the world*. Penguin Random House, 2016.
- [] R. T. Thibault, O. B. Amaral, F. Argolo, A. E. Bandrowski, A. R. Davidson, and N. I. Drude. *Open Science 2.0: Towards a truly collaborative research ecosystem*. *PLoS Biol*, 21(10):e3002362, October 2023.
- [] B. Trovò and N. Massari. *Ants-Review: A Privacy-Oriented Protocol for Incentivized Open Peer Reviews on Ethereum*. *Lecture Notes in Computer Science*, 2021.
- [] J. Van Rossum. *The blockchain and its potential for science and academic publishing*. *ISU*, 38(1-2):95–98, October 2018.
- [] N. A. Vasilevsky, M. H. Brush, H. Paddock, L. Ponting, S. J. Tripathy, G. M. LaRocca, and M. A. Haendel. *On the reproducibility of science: unique identification of research resources in the biomedical literature*. *PeerJ*, 1:e148, September 2013.
- [] M. Vukolić. *Rethinking permissioned blockchains*. In *Proceedings of the ACM Workshop on Blockchain, Cryptocurrencies and Contracts, BCC '17*, page 3–7, New York, NY, USA, 2017. Association for Computing Machinery.

- [] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. Da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 'T Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. Van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. Van Der Lei, E. Van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. [The FAIR Guiding Principles for scientific data management and stewardship](#). *Sci Data*, 3(1):160018, March 2016.
- [] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, and others. [The fair guiding principles for scientific data management and stewardship](#). *Scientific Data*, 3:1–9, 2016.
- [] G. Wood. [Ethereum: A secure decentralised generalised transaction ledger](#). *Ethereum White Paper*, 2014.
- [] Y. Wu, X. Zheng, J. Ren, K. Zhang, and Y. Shen. *Blockchain-based secure key management scheme with high availability*. *IEEE Access*, 6: 27766–27774, 2018.
- [] X. Xu, L. Yang, and L. Zhao. [A survey of blockchain data storage issues and solutions](#). *Journal of Computer Networks and Communications*, 2018:1–17, 2018.
- [] X. Zhang, H. Li, and B. Xu. [Decentralized data storage systems: Blockchain and beyond](#). *Journal of Blockchain Research*, 3(2):45–56, 2020.
- [] Y. Zhou, Z. Wan, and Z. Guan. [Open-pub: A transparent yet privacy-preserving academic publication system based on blockchain](#). In *2021 International Conference on Computer Communications and Networks (ICCCN)*, pages 1–11, 2021.

*This document was typeset on May 14, 2025 using class $\text{\texttt{RG-BOK}}$ $\alpha 2.14$ for $\text{\texttt{LATEX}}_2\epsilon$. As of the time of writing this document, this class is not publicly available. Only members of *The Distributed Group (TDG)* and the *Applied Computing Research Group (GCA)* are allowed to typeset their documents using this class.*