DeSci Labs

DeSci Labs and Filecoin: Enabling a Future of Open Science

Brought to you by
Filecoin
Foundation

DeSci Labs

The “Replication Crisis,” the notion that results reported in scientific papers are frequently unable to be reproduced and independently verified by other researchers, has been an **open secret within the scientific community** for several decades now.

Estimates of exactly what percentage of published scientific studies are unable to be replicated vary widely, but recent reports suggest it could be as high as 40 percent. Whatever the figure, this problem obviously **hinders progress and undermines the credibility of the scientific enterprise.**

Concepts such as the Open Science movement and **FAIR Data Principles** have emerged as promising frameworks for addressing this issue by promoting data transparency and reproducibility in scientific research. However, a lack of technical infrastructure and incentives has hindered the broader adoption of these concepts. **DeSci Labs** aims to address these challenges by leveraging decentralized technologies like InterPlanetary File System (IPFS) and Filecoin, which operate “under the hood” of their scientific publication app called **DeSci Nodes**.

Content-addressed storage on IPFS and Filecoin enables content verification and tackles issues such as link rot and content drift.

Decentralized Persistent Identifiers (**dPIDs**) turn scientific publications into APIs, allowing for programmatic operations on these publications, such as data egress, edge compute functions, attestations, and resolution to machine actionable metadata. dPIDs are built from the ground-up to power the FAIR data principles, and unlock new modes of interacting with published research.

Storing data and research objects in a permanent yet cost-effective manner also poses significant challenges in the current environment. **Filecoin provides a needed solution by making storage affordable, preventing vendor lock-in** and fostering a culture of openness. By storing research objects in a decentralized repository, it becomes possible to share the entire research object with data and code, provide versioning, and ensure reproducibility.

By embracing open science principles and leveraging decentralized web technologies such as Filecoin and IPFS, the scientific community can overcome the challenges posed by the lack of transparency and reproducibility. These tools are the foundation of DeSci Labs’ vision for creating an open-access network without paywalls, barriers, and vendor lock-in, where scientific data is stored and accessed in a way that facilitates collaboration, transparency, and the free flow of knowledge.

Science's Open Secret

Science's open secret – its Replication Crisis – is problematic because rigorous testing and replication of studies by other researchers is a foundational component of advancing scientific knowledge – whether it be through repeating the experiment using different data or methods of analysis or by recreating the experiment using the original data and code. **Research cannot be appropriately evaluated and used for future projects if the tools needed to replicate it are lacking.**

Nature reported in 2016 that 70 percent of researchers had tried and failed to reproduce another scientist's experiment. A 2023 **report** suggests that 28 percent of all biomedical **papers** published in the medical field are either "made up or plagiarized" outright, while others **suggest** that number may be much higher. Suffice to say that it is **difficult to discern when scientific studies, even when published in reputable journals, can be fully trusted.**

In a landmark 2012 example, Amgen, one of the world's largest biotech companies, **revealed** that an effort to reproduce results from 53 landmark cancer research papers had largely failed — only six yielded results, meaning nearly **90% were not reproducible**. In 2016, the company announced three more reproduction efforts of high-profile, previously published studies had failed.

In another high-profile example, Science magazine **revealed** in 2022 that the amyloid hypothesis, used for decades as the basis for studying and treating Alzheimer's disease, **was based partly on false data and images.** What was thought to be the most reliable scientific data, leading to countless hours and billions of dollars of follow-up research investment, was largely not reproducible.

Further exacerbating the problem is an **expanding global cottage industry of fake science "paper mills"** - outfits that are paid to produce fabricated scientific submissions. Bernhard Sabel, a researcher at Otto von Guericke University Magdeburg in Germany, detailed the extent of this issue in an interview with the **Financial Times** in May 2023:

"Fake science publishing is possibly the biggest science scam of all time, wasting financial resources, slowing down medical progress and possibly endangering lives."

No Solution in Sight?

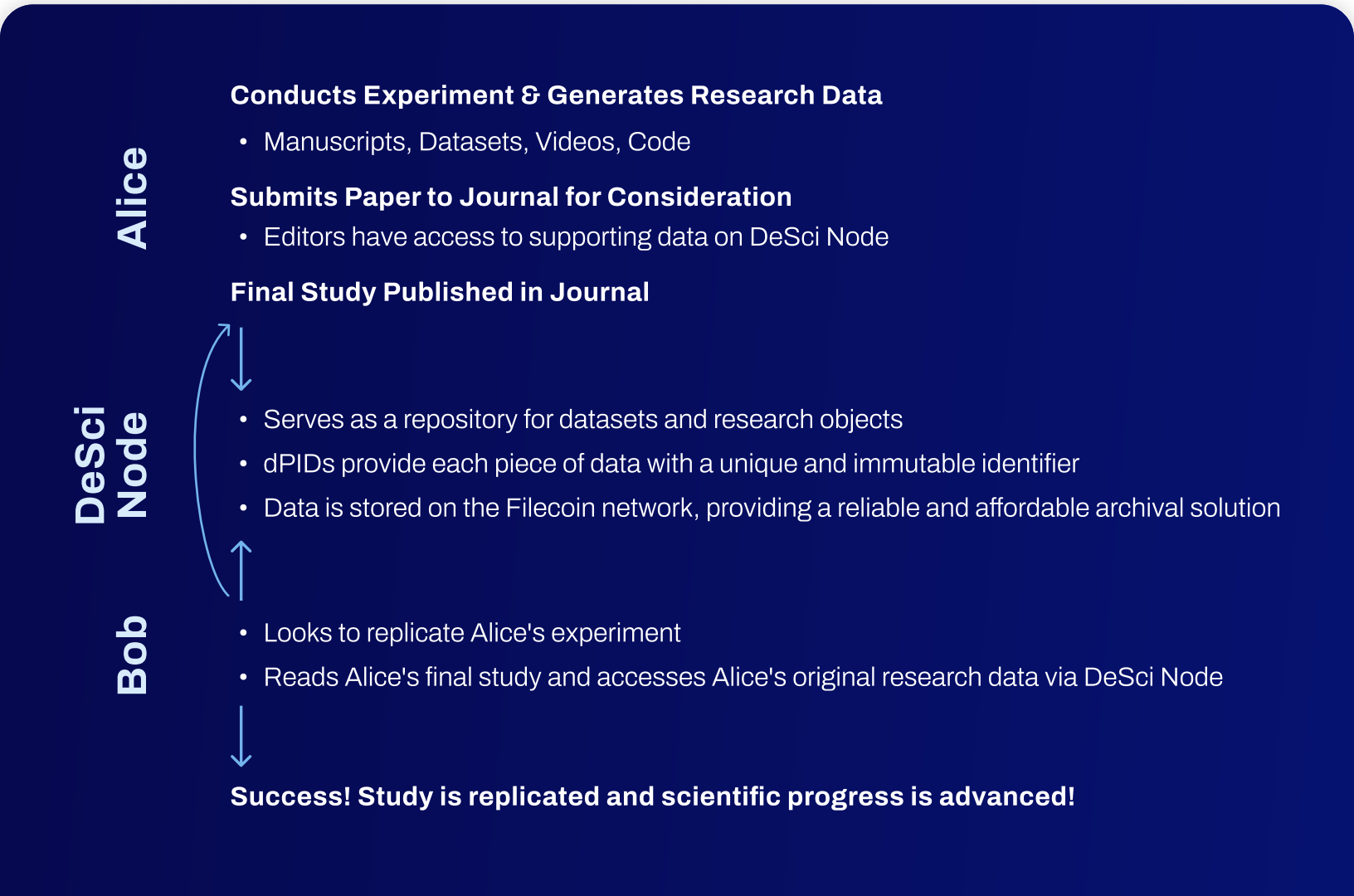
So if this Replication Crisis has been a known issue for several decades, why does the problem persist?

Unfortunately, there exists a lack of incentives for scientists to embrace replicability. The current regime of scientific publishing values the final manuscript, which is published in journals typically without the underlying data, code, and other research artifacts necessary to reproduce the experiment. Journal editors typically look at what researchers found without scrutinizing the data and code that allowed them to arrive at the conclusion.

This is akin to showing only your final answer on an algebra test while withholding the underlying work that shows how you arrived at the answer.

This incentive structure also **fails to provide a relevant assessment of the background work** that went into the research. For example, did the researcher also conduct related experiments that failed to produce significant results? This relevant information is oftentimes left out of the final published study.

Another prevailing issue is that the scientific world lacks a robust technology stack that can enable knowledge sharing for easy reproducibility. While many researchers strive to embrace greater transparency and data sharing, **the infrastructure does not exist currently to do this quickly, easily, and cheaply.**



These figures go beyond hyperbole, as **storing enormous troves of data is commonplace** in the scientific field nowadays. The James Webb telescope alone produced 100TB of data in its first three years. Genetic sequencing for just 800 people yields a 100TB dataset; for context, modern genetic reference panels sequence upwards of 100,000 people.

Another problem with current technical infrastructure is the **lack of persistent identifiers**, or PIDs, which assign unique, long-term references to digital resources. This is a fundamental computer science problem and a pain point inherent to the current architecture of the World Wide Web.

The persistent identifiers used in the current system of scholarly publishing, DOIs, are neither unique nor persistent - they do not consistently resolve to the same, correct resource. Furthermore, they are controlled by central authorities and they are costly to maintain. These are pervasive issues that cause researchers to lose track of and access to critical information. A **study from the Los Alamos National Laboratory** took a sample of science, technology, and medicine papers and found more than 75 percent of the linked URLs had drifted away from its original content or ceased to be available. All of this creates significant concerns about published scientific research's long-term reliability and integrity.

A New Protocol for Securing Knowledge

The basis of the work by **DeSci Labs** is to **build tools and infrastructure that make reproducible science with low barriers to access a reality.** DeSci Labs is a team of scientists and engineers bringing a modern approach to the future of scientific research, with the perspective of having used siloed data and centralized systems for years without significant progress. The vision is a world without vendor lock-in, paywalls, proprietary analytics, link rot and fake science.

The **key enabling technologies for this is the InterPlanetary File System (IPFS)** - which enables content-addressing of data and research objects, and the Filecoin network for reliably storing these objects.

By providing a unique and immutable identifier for each piece of data - and creating a new record for content each time it's updated, **IPFS mitigates link rot and eliminates content drift entirely.**

DeSci Nodes is an interface built by DeSci Labs on top of tools developed by Protocol Labs, that **allows scientists to create digital research objects** that contain manuscripts, datasets, code, presentations, videos or other relevant research artifacts. DeSci Nodes functions similarly to a pre-print server where researchers can post their work, having the possibility to get recognition not only for the final manuscript, but also for all related items such as data and code.

Embracing Open Science and FAIR Data Principles

The momentum for Open Science and FAIR Data Principles is growing in response to the loss of valuable datasets and research code that make science verifiable, and the fact that the vast majority of published manuscripts are hidden behind paywalls. Even though many of these studies are publicly-funded, the final work is published and copyrighted by the journal, and then frequently rendered inaccessible to potential readers (and sometimes the researchers themselves) via paywall.

FAIR scientific data is not just desirable from a philosophical perspective; **there are economic efficiencies to be realized as well.** The European Commission **estimates** that the financial cost of not having FAIR scientific data is more than 10 billion euros annually. Recognizing this, governments are increasingly requiring the adoption of FAIR principles in grant and funding applications to combat the waste of valuable resources.

While the idea of Open Science serves as a guiding North Star for making manuscripts, data, and other information readily accessible and available for accurate replicability and reproducibility, the **FAIR Data Principles** lay out a practical framework for how to actually do that in practice. FAIR stands for:

- **Findable:** To use data, scientists must first be able to find them. Machine-readable metadata is necessary to discover datasets and services, so they should be easy to find for both humans and computers.
- **Accessible:** After the data is found, researchers need to be able to access it and know whether any authentication or authorization processes are necessary to view the datasets.
- **Interoperable:** Different datasets must be able to work together. They must also integrate with applications and workflows for analysis, storage, and processing.
- **Reusable:** The overarching goal of FAIR is data reusability. For this to happen, data must include enough detail that analysts can replicate and combine different datasets under other conditions.

There are still major infrastructure barriers to implementing FAIR data in practice, however. For one, a massive amount of affordable storage is required. While many providers for housing scientific data currently exist, this ecosystem is highly fragmented and **storage and retrieval costs can be prohibitive for scientists and organizations with lots of data and metadata storage needs.** For example, to store a 100TB dataset and access it approximately five times a day costs more than $1 million annually on Amazon Web Services.

Scientists who use Nodes can also publish their work in the journals that are most relevant to them. Indeed, Nodes help scientists to get their work published in the venues needed to advance their careers by allowing referees and editors to easily access and check the origins of the results they are reporting in their manuscripts. The goal is to **augment current research and publishing infrastructure by providing a reliable and accessible repository for research artifacts** that publishers are not equipped to handle.

DeSci Nodes provide **free, universal access** to research and offer authors free uploads for content, up to a specific limit, while charging a fee for storing very large datasets or using compute abilities.

The DeSci Nodes interface is co-developed with an open protocol for assigning interoperable persistent identifiers called dPID (short for decentralized persistent identifiers). dPIDs take hashes generated by IPFS and turn them into shorter, human-friendly PIDs. These PIDs not only protect from content drift and minimize the threat of link rot, but **turn every Node into an interoperable research object in a way that makes them actionable for machines.** With dPIDs, researchers can import a dataset directly into their IDEs, or send a compute job where the data is stored. dPIDs provide consistent resolution to their machine-actionable form which is bound to become extremely important in the age of AI-augmented research. And finally, they serve as addresses for attestations to curate the content of Nodes.

Another powerful feature that Nodes leverage is edge computing, such as the compute-over-data functionality being developed by Expansio (formerly known as Bacalhau). **Edge computing allows scientists to send containerized code to be executed directly where a particular dataset is physically stored on the Filecoin network.** This is important because data has gravity: egressing large datasets is not only costly, but wasteful for the environment. Edge computing comes with another benefit: all computations become traceable, and provide a new type of metric to assess the impact of particular datasets.

Edge computing creates a more efficient environment for collaboration and dramatically cuts the cost of transferring large datasets; currently, one pebibyte (PiB) of data can take a month to download. Using CoD networks like Bacalhau, another Filecoin ecosystem partner, stakeholders can view datasets, run compute jobs over the data, and retrieve these results seamlessly.

The Filecoin network serves as the archival layer of the dPID protocol, and its utility as an affordable and reliable storage solution addresses several hurdles facing researchers wishing to share their data. Emerging possibilities offered by the FVM (Filecoin Virtual Machine) will further bolster the capabilities of dPIDs: imagine if every decentralized persistent identifier minted could provably deliver long-term storage guarantees for the underlying data payload it indexes? This could dramatically simplify data cataloging and archiving to meet funder requirements.

The dPID protocol also addresses the issue of data provenance, which is becoming increasingly critical with the emergence of generative artificial intelligence.

Currently, observers often don't have the necessary context to determine the validity of a particular research object. Today, certification relies on the existence of the publication's metadata forwarded by a publisher to an indexing service such as the **Web of Science**. In contrast, dPIDs serve as targets for verifiable attestations, which can be issued by entities curating the scientific record such as journals or scientific societies. These attestations, pioneered by the **Center for Open Science** and used by the **ACM**, act as badges of quality and help researchers engaging in open science practices to gain recognition. Today, there exists no good solution to securely attach these attestations to PIDs and make them legible by indexing services. dPIDs aim to change this by providing a secure substrate for these markers of quality and act as a connection point for indexing services.

Ultimately, every research object published on the dPID protocol lives in a common, or "open state" repository. This means that access to this repository architecturally enforces openness. Anyone can fetch the hash associated with a dPID by reading the stats on the dPID contracts, and retrieve the Merkle-DAGs containing the data and metadata from the IPFS network. This creates a type of credibly neutral infrastructure on which others can build securely, without fears of vendor lock-in.

Importantly, this architecture will accommodate restricted access data: a particular PID can be marked as requiring authentication to access its underlying data payload – redirecting the user to the enterprise grade storage provider that complies with regulatory restrictions on data access and provides a secure authentication procedure.

Reclaiming the Scientific Record

The Filecoin archival network and IPFS are key building blocks to realize DeSci Labs vision of creating a more interoperable, open, and reproducible scientific record. This foundation enables reliable and scalable access to research objects secured by the dPID protocol, which will in turn allow the FAIR Data Principles to manifest their promise of interoperability and reusability of research artifacts for all humans and machines.

The vision of DeSci Labs is to allow scientific publishing communities to upgrade their unit of knowledge, from the solitary PDFs sitting behind the paywalls, into reproducible and interoperable research objects stored in a common repository built from the ground-up to power the FAIR principles. In this story of old against new, **DeSci Labs is building the catapult to propel the scientific record into the age of reproducible research.**

DeSci Labs

DeSci Labs and Filecoin: Enabling a Future of Open Science

To learn more about how to get involved with DeSci, please visit **desci.com**

Learn more about the Filecoin Ecosystem

fil.org
@FilFoundation

ipfs.tech
@IPFS

filecoin.io
@filecoin

Brought to you by
Filecoin
Foundation