

dPIDs - the Emerging Persistent Identification Technology for FAIR and The Digital Era

Intended length: 4-6 pages

Template: IEEE CS

Process: LaTeX

Deadline: **February 5, 2024**

Authors (please add yourself with ORCID and affiliation):

- Andrey Vukolov 0000-0001-6967-3236, Elettra Sincrotrone Trieste
- Erik Van Winkle, 0000-0002-7567-0311, DeSci Foundation
- Elizaveta Zhdanova, 0009-0005-3701-0266, Universitat Politecnico de Valencia
- Georgios (George) Kourousias, 0000-0002-1243-7168, Elettra Sincrotrone Trieste
- Roberto Pugliese, 0000-0002-1687-4795, Elettra Sincrotrone Trieste

Abstract

This paper presents a comprehensive exploration of an emerging technology focused on the persistent identification and sharing of data and metadata, termed “decentralised Persistent Identifier” (dPID). Utilizing the InterPlanetary FileSystem (IPFS) network, dPID provides reproducible persistent identifiers, ensuring that data can be reliably stored and accessed over time. It is designed to incorporate a distributed ledger, Public Key Infrastructure (PKI), decentralized linking, and lookup databases. It also aims to support version control and provenance tracking based on reproducibility. The distributed approach highlights the potential of dPID to align with the principles and guidelines of Findable, Accessible, Interoperable, and Reusable (FAIR) data, positioning it as a valuable component within the open data ecosystem, mitigating in efficient way the problems, such as link rot and content drift. The proposed technology provides a highly scalable persistent identification and versioning system for shared data that reduces dependencies from social contracts and institution-driven systems. The paper also demonstrates its usability for modern social-oriented identification systems, proposing a use case study for the art industry.

List of Acronyms (removed from LaTeX version)

The Acronyms used in the paper are listed here in the appearance order.

- PID - Persistent Identifier
- dPID - Decentralised Persistent Identifier
- IPFS - InterPlanetary FileSystem
- PKI - Public Key Infrastructure
- FAIR - Findable, Accessible, Interoperable, Reusable

- ExPaNDS - European Open Science Cloud Photon and Neutron Data Services
- IOT - Internet of Things
- OSTP - Office of Scientific and Technological Policies
- EC - European Commission
- NIH - National Institutes of Health
- AI - Artificial Intelligence
- DOI - Digital Object Identifier
- ORCID - Open Researcher and Contributor Identifier
- ROR - Research Organization Registry
- HTTP - HyperText Transfer Protocol
- W3C - World Wide Web Consortium
- DID Decentralised Identifier
- IGSN - International Generalized Sample Number
- SWHID - Software Heritage Identifier
- RAiD - Research Activity Identifier
- DHT - Distributed Hash Table
- CID - Content Identifier
- DAG - Direct Acyclic Graph
- API - Application Programming Interface
- IPLS - InterPlanetary Linked Data
- JSON-LD JavaScript Object Notation Linked Data
- RO-CRATE - Research Object Crate

Introduction

PIDs are not only for documents and data. They can also be used to reference other entities or agents, including people contributing to the research, software, research organizations, physical objects such as samples and instruments, and even abstract concepts such as terms in a controlled vocabulary. As a general rule, whenever something needs to be referenced in a reliable and lasting manner, a persistent identifier should be used.

ExPaNDS Deliverable 2.5¹

The modern world is encountering unprecedented challenges in data handling. Growing amounts of data in the realm of academia, including IOT sensing, multi-disciplinary works and AI form a collective ocean of knowledge, impassible without the collaboration of humans and machines. One point of advocacy for the necessary human-computer symbiosis can be found in the FAIR principles and subsequent literature around machine-actionability²³.

¹ Bunakov, Vasily, Rolf Krahl, Brian Matthews, Noel Vizcaino, Andrey Vukolov. «Advanced Infrastructure for Pids in Photon and Neutron Ris». Zenodo, 3 март 2022 г. <https://doi.org/10.5281/zenodo.5905351>

² Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

³ Bozada T Jr, Borden J, Workman J, Del Cid M, Malinowski J, Luechtefeld T. Sysrev: A FAIR Platform for Data Curation and Systematic Evidence Review. *Front Artif Intell.* 2021 Aug 5;4:685298. doi: <https://doi.org/10.3389/frai.2021.685298>

Policy-based recognition in support of these principles have been released over the past 5 years from organizations such as the OSTP, EC, and NIH⁴⁵⁶.

The first of the 15 FAIR Principles, Principle F1, states that “Data and metadata must have globally unique, persistent, and resolvable identifiers” [2]. In the upcoming world of FAIR Science, the scientific record strives for the permanence of information online. While permanence is impossible, PIDs and underlying technological infrastructures are essential components of continuous navigation and data retrieval. Existing PID providers such as DOI, mitigate the problems of persistence using social and technical mechanisms to guarantee consistent mapping and resolution of a given ID to its target resource. This goal of persistent identification and resolution has a standard set of problems.

Ensuring the scalability of social and technical mechanisms in persistent identification technology can be a challenge. Industry 4.0, IOT and rapidly growing open-source development⁷ drastically increase the amount of data that should be identified and linked in a strictly deterministic manner. The reusability of metadata is required for the efficiency of data access, meaning that the persistence of metadata and its subsequent linkage to data becomes essential too. Social trust as a mechanism to combat the replication crisis requires added persistence requirements. The high diversity of data storage and transportation techniques, protocols, networks, etc. also reveals new data treatment circumstances where every file should be identified, checked for integrity, versioned, and shared over many possible networks with different address resolution mechanisms. The average scientific project or experiment outcome may contain hundreds of thousands of files (research artifacts), including code, data, reviews, sensemaking statements, etc., both created by the people and captured automatically. Global endeavors in a single scientific domain can produce trillions of individual files of their own accord. The pursuit of permanence for scientific knowledge starts with a problem of scaling persistent identification infrastructure.

Additional challenges surrounding current PID infrastructure include but are not limited to the sovereignty and access control over data⁸, the necessity of machine-actionable provenance, client-side reproducibility of existing PIDs due to their generation algorithms being enclosed

⁴ Kuehn BM. NIH Data Sharing. *JAMA*. 2007;298(20):2361. doi:

<https://doi.org/10.1001/jama.298.20.2361-a>

⁵ David, R., Rybina, A., Burel, J. M., Heriche, J. K., Audergon, P., Boiten, J. W., ... & Gribbon, P. (2023). “Be sustainable”: EOSC-Life recommendations for implementation of FAIR principles in life science data handling. *The EMBO Journal*, 42(23), e115008.

<https://doi.org/10.15252/emboj.2023115008>

⁶ Heather Moulaison-Sandy (2023) The Nelson Memo and US Federal Funder Requirements for Public Access: Implications for Technical Services Librarians, *Technical Services Quarterly*, 40:4, 290-297, DOI: <https://doi.org/10.1080/07317131.2023.2271278>

⁷ Plaga, S., Wiedermann, N., Anton, S. D., Tatschner, S., Schotten, H., & Newe, T. (2019). Securing future decentralised industrial IoT infrastructures: Challenges and free open source solutions. *Future Generation Computer Systems*, 93, 596-608. <https://doi.org/10.1016/j.future.2018.11.008>

⁸ Mahalle, P.N., Shinde, G., Shafi, P.M. (2020). Rethinking Decentralised Identifiers and Verifiable Credentials for the Internet of Things. In: Dey, N., Mahalle, P., Shafi, P., Kimabahun, V., Hassanien, A. (eds) *Internet of Things, Smart Computing and Technology: A Roadmap Ahead*. Studies in Systems, Decision and Control, vol 266. Springer, Cham. https://doi.org/10.1007/978-3-030-39047-1_16

within the registrar's infrastructure and an exponentially growing amount of resolution requests due to the formation of PID graphs⁹.

With trillions of PIDs likely needed by 2030, the probability of system failure in current federated but centralized architectures leads to system fragility and a potential loss of valuable knowledge¹⁰.

To stay relevant and reliable under the extreme circumstances of science in the digital age, PID system intended usage in the modern FAIR-compliant data-driven Internet needs to implement a reproducible, fully automated, scalable, globally unique identifier system, within which a minimized reliance on social contracts, leading to fewer centralized points of failure and human interactions. In the following sections, we will explore the range of currently available PID systems, delving into their underlying technologies and unique characteristics. Subsequently, we will present a technical proposal for an emerging decentralized PID system, including a concise analysis and discussion on the prospects of its implementation.

Additionally, this paper proposes a potential application of the dPID technology in the realm of fine arts and photography. This use case focuses on hybrid distribution models that encompass both hardcopy and digital formats, addressing the unique challenges of managing, distributing, and verifying artworks across diverse mediums. By leveraging dPID, stakeholders in the fine arts sector could benefit from enhanced provenance tracking, more secure distribution channels, and improved accessibility to art, all while adhering to the FAIR data principles.

The rest of the paper is structured as follows. In Section 2 we explain the current status of the PID implementation landscape, holding ExPaNDS project outcomes as a basis. In the first part of Section 3, we give a technical proposal based on one of the possible implementations of a decentralised approach to persistent identification. In the second part of this section, we leverage the technology by giving a technical outline of freshly developed open-source software. Then in the Conclusion section, we summarize the problems we have indicated are solved in the paper. This section here is also extended with a use case proposal showing the possibilities of dPID in the area of persistent identification and distribution of the artworks.

⁹ Cousijn, H., Braukmann, R., Fenner, M., Ferguson, C., van Horik, R., Lammey, R., ... & Lambert, S. (2021). Connected research: The potential of the PID graph. *Patterns*, 2(1). <https://doi.org/10.1016/j.patter.2020.100180>

¹⁰ Andrey Vukolov. (2021, июнь 18). Openly reproducible Persistent Identifiers (PIDs) as a factor of FAIRness in data sharing practices. EOSC Symposium 2021. Zenodo. <https://doi.org/10.5281/zenodo.4980522>

Current PID Implementations: Landscape and Centralization Status

This section is partially based on works[1] and [11]¹¹. It explores the implementation details of the most known and popular PID systems in the context of centralization. The mentioned PID systems (DOI, ORCID, Handle, ROR, etc.) implemented worldwide fall between *centralized* and *federated* architectures¹² implementing the centralised resolution model presented in Figure 1. The decentralised architectures are now only emerging, so no one from the existing systems could be named decentralised.

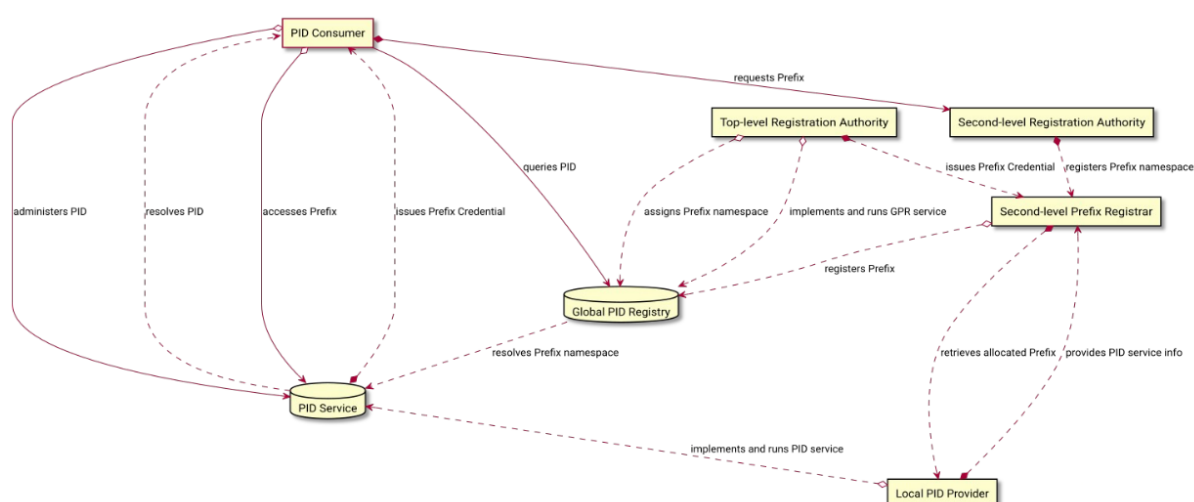


Figure 1. Centralised PID Administration and Implementation Schema

This model implements a low number of high-risk singular points of failure (as it is based on the entity called **Global PID Registry**), despite the existence of both primary and secondary authorities. From the client's point of view, they should be called **centralized**. Every existing point of failure here may lead to the failure of the entire facility or domain. If the PID is not reproducible and the generation schema is closed, the centralised PID governance authorities are the only provenance holders of the underlying record, metadata, and addressed data¹³. This self-assignment trap magnifies when record tracking lags behind the growing number of maintained records. Issues such as link-rot, content drift, artifact fragmentation and inconsistent resolution¹⁴ in these systems can be traced back to human interactions, leading to lacking persistence because they do not provide lookup table

¹¹ Pablo de Castro, Ulrich Herb, Laura Rothfritz, Joachim Schöpfel, Some reflections on the current PID landscape – with an emphasis on risks and trust issues, *Procedia Computer Science*, Volume 211, 2022, Pages 28-35, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2022.10.173>

¹² Brown, Josh. «PID Federation Scoping Study: Final Report». Zenodo, 30 сентябрь 2020 г. <https://doi.org/10.5281/zenodo.4059557>

¹³ Nicolas Soler, Abigail McBirnie, Alejandra Gonzalez-Beltran, Andrey Vukolov, Carlo Minotti, Heike Goerzig, Krisztian Pozsa, Brian Matthews, Majid Ounsy. «Final Recommendations for FAIR Photon and Neutron Data Management». Zenodo, 7 июль 2022 г. <https://doi.org/10.5281/zenodo.6821676>

¹⁴ Klein, M., Balakireva, L. (2020). On the Persistence of Persistent Identifiers of the Scholarly Web. In: Hall, M., Merčun, T., Risse, T., Duchateau, F. (eds) *Digital Libraries for Open Knowledge. TPD 2020. Lecture Notes in Computer Science()*, vol 12246. Springer, Cham. https://doi.org/10.1007/978-3-030-54956-5_8

redundancy¹⁵¹⁶, reproducibility and proper caching. However, while the extreme circumstances of the modern data-driven world do not affect all domains, centralized authorities such as person-oriented ORCID and organization-oriented ROR have proven trustworthy over time, at scales of ~1-10M of PIDs. For their social-driven data flows and scalability cases, they are simple, efficient, and can be considered fit-to-purpose.

As a reply towards the challenges arising, different flavors of **federated** approach have appeared. It combines the existing conservative model of centralized governance authority with a federated social-driven network of formally independent PID registrars. Each registrar manages the underlying system of centralized prefix-based lookup tables with a federated network of resolvers based on HTTP redirection. They act from the governance point of view, as independent registries with independent databases replicated at the discretion of the centralized authority. Each registrar maintains its own provenance authority and ensures the persistence of associated PIDs alone. To control the resources distributed across the federation of infrastructure providers in maintaining persistent registry infrastructure, an external council or legal entity is still needed, leading to a non-profit governance organization overseeing the system. Through this structure, the federated PID infrastructure aims to preserve the administrative efficiency of the centralized approach while reaching redundancy and distributed curation of decentralized approaches. The most popular PID systems in the world, DOI and Handle.net, use the federated approach. It allows expansion of the computational abilities of the system, primarily in the aspect of data replication.

To obtain interoperability and cross-resolution, especially in federated architectures, the third-party aggregation approach, in practice, is the only way. The UNIREsolver initiative is the most known solution for implementing it. It utilizes W3C DID¹⁷¹⁸¹⁹ specification to perform a bidirectional lookup but without internal implementation of metadata versioning and reproducibility tracking, so the records resolved via UNIREsolver could not be considered truly immutable and the associated resolution cannot be considered deterministic on the periods on which the PIDs should persist. However, UNIREsolver demonstrates the possibilities of cross-resolution and request flow balancing in the federated architecture without rebuilding the existing provenance chains.

¹⁵ Klump, J., Lehnert, K., Ulbricht, D., Devaraju, A., Elger, K., Fleischer, D., Ramdeen, S. and Wyborn, L., 2021. Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. *Data Science Journal*, 20(1), p.33. DOI: <https://doi.org/10.5334/dsj-2021-033>

¹⁶ Kunze, J. (2003). Towards Electronic Persistence Using ARK Identifiers. *UC Office of the President*. Retrieved from <https://escholarship.org/uc/item/3bg2w3vs>

¹⁷ Miguel-Angel Sicilia, Elena García-Barriocanal, Salvador Sánchez-Alonso, Juan-José Cuadrado, Decentralized Persistent Identifiers: a basic model for immutable handlers, *Procedia Computer Science*, Volume 146, 2019, Pages 123-130, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2019.01.087>

¹⁸ Decentralized Identifiers (DIDs) v1.0. Core architecture, data model, and representations. W3C Recommendation 19 July 2022. Access date 22.02.2024 <https://w3c.github.io/did-core/>

¹⁹ Bach, N. (2021). Dezentrale Identifikatoren (DIDs): Die nächste PID-Evolution: selbstsouverän, datenschutzfreundlich, dezentral. *o-bib. Das offene Bibliotheksjournal/Herausgeber VDB*, 8(4), 1-20. <https://www.o-bib.de/bib/article/view/5755>

In Table 1 a brief list of currently implemented worldwide PID providers is presented, with their underlying technologies and/or PID type specifications and dedicated entity types. The table is taken from [1] with minor additions.

Table 1. International PID providers, their underlying technologies and entity types

Provider	PID technology	Entities	Centralization
DataCite https://datacite.org/	DOI	General purpose	Federated
Crossref https://www.crossref.org/	DOI	Publications, funders	Federated
ePIC https://www.pidconsortium.net/	Handle	Metadata in all plaintext schemas	Federated
IGSN https://www.igsn.org/	Handle	Experimental samples	Federated
ORCID https://orcid.org	Bespoke (custom)	Persons	Centralized
FigShare PID https://figshare.com/	DOI	Research artifacts	Federated
Zenodo https://zenodo.org	DOI	Publications, digital research artifacts	Federated
EUDAT B2SHARE https://b2share.eudat.eu/	Handle, DOI	Datasets, digital research artifacts	Federated
FAIRshare https://fairsharing.org/	DOI	Datasets, policies, standards	Federated
SWHID https://docs.softwareheritage.org/devguide/swh-model/persistent-identifiers.html	SHA1-based Merkle DAG	Software artifacts, versioned repositories	Federated, decentralized
ROR https://ror.org/	Bespoke (custom)	Research institutions	Centralized
RAiD https://www.raid.org.au/	Handle (custom, prefixed)	Funders, organizations, persons, instruments, datasets	Centralized

As it is easy to see from the entries presented in Table 1, only one of the PID providers implemented worldwide, uses a PID with client-side reproducibility - SWHID. Thus it should be considered the only one that has already implemented the provable immutability and inline versioning, but under the centralized governance. The federated DOI system holds leadership in dissemination; it is used as the underlying provider by most listed PID providers.

dPID: Adding Decentralization to Data Identification Pipelines

The core idea behind decentralized data identification technologies is that the client should be asking a fundamentally different question when resolving the metadata from the PID. Instead of asking a single point: "What is the content stored at this location?", it should be asking a network-based decentralized swarm: "Can you tell me how to find the content with this hash?". This approach is one of the natural continuations of an idea of the PID Graph[9], proposing the technical mechanisms as the primary mode of guaranteeing persistence, replacing the social mechanisms used in current systems. ***In a decentralized PID system, the social persistence of a PID is as strong as the technical prevalence of the network nodes, providing self-describing addressing, resolution and consistent data.*** One of the possible workflow schemas implementing the decentralized approach is presented in Figure 2. As it could be easily deduced from the block names in Figure 2, most of the elements of the PID system indicated are the distributed ledgers and databases where the data are not owned or stored on the side of the single participant or institution: the lingering in this schema requires a mathematically proven integrity of the internal metadata stored in the system. Also, the presented schema should not be considered as a reference: the decentralised approach makes the implementation schema fuzzy and yields its structure in favour of the workflow.

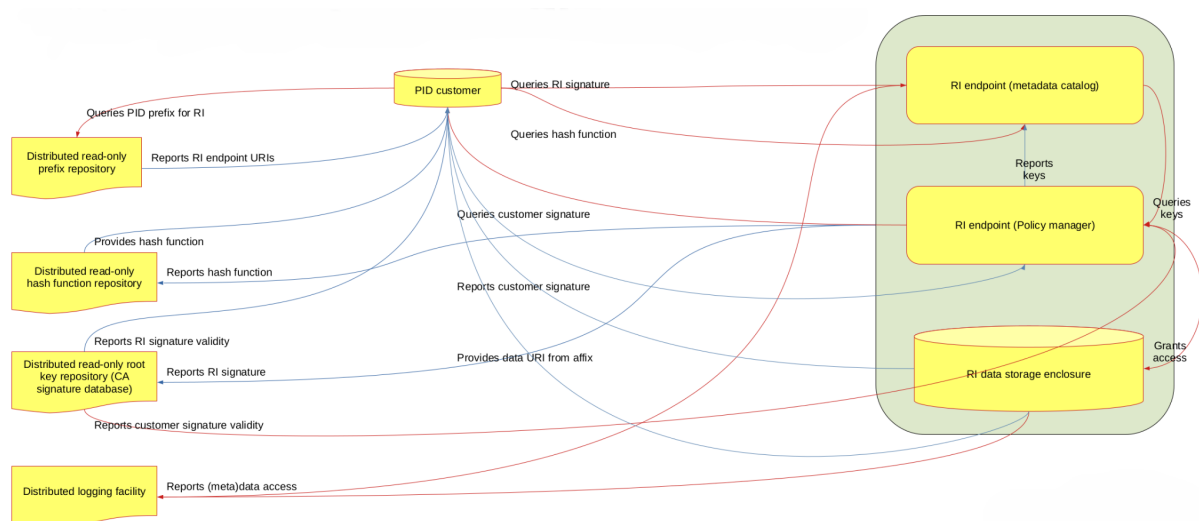


Figure 2. One of the Possible Decentralised PID Workflow Implementation Schemas

The ideal implementation of the PID system necessitates uniformity in backend technology and storage models across all resolvers and PID generators. In such a setup, users would need to install a specific software, possibly on a local machine, and register a public key. This process would grant them access to a resolution endpoint, equip them with a PID generator that supports namespace propagation right from the start, and provide a viewer for the provenance chain. It demands data pipeline technologies that can handle storage, identification, and delivery in a manner that is as type-agnostic and mathematically secure as possible. Fortunately, there exists a category of technologies that fulfill these requirements. Broadly, these can be characterized as decentralized redundant storage systems, with BitTorrent being one of the most well-known examples. Such technologies rely

on hash-based, mathematically-driven content addressing, and utilize DHT for the delivery of binary data objects. DHT, in particular, incorporates identification directly within its addressing stack, making these technologies an apparent choice for the described purposes[17].

The IPFS decentralized storage network²⁰ stands out by allowing data to be stored under a single ID that is immutable and persistent, based on their binary representation. This system employs Content IDs (CIDs), which are derived from the outputs of unidirectional cryptographic hash functions²¹. These IDs incorporate a prefix-based structure, with the specification detailing how they should be resolved, ensuring a comprehensive description within the ID's structure itself. CIDs can be seen as production-ready PIDs that are resistant to content drift and link rot.

The IPFS storage model is designed to support a cache-on-read retrieval approach, where data, along with their metadata, are recorded into DAGs using various storage tree builders. This ensures that data and metadata are automatically cached, with the option for explicit declaration of caching for specific data on a given network node. The persistence levels for data stored in IPFS, as determined by the CID's state declared in the DAG, are outlined in the following table. In this context, a "pinned" state indicates that a specific CID has been explicitly marked for caching on at least one node within the IPFS network.

DAG CID state	Pinned	Wanted	Available	Discarded
Local data state				
Stored	Network-wide persistent	Network-wide available	Network-wide persistent	Network-wide available
Requested	Network-wide persistent	Network-wide available	Network-wide available	Available at first accessible relay
Idle	Network-wide persistent	Network-wide available	Network-wide available	Network-wide available
Discarded	Network-wide persistent	Available at first accessible relay	Available at first accessible relay	Unavailable with given CID

Table 2. Possible availability states of the data entry in the IPFS network

IPFS also incorporates IPLD, utilizing a hash-based addressing model that facilitates the construction of decentralized storage trees²². This model enables a persistent linkage between stored items and their versions, ensuring that each piece of data can be uniquely identified and accessed over time. Given the features that have already been implemented, IPFS emerges as a suitable choice for implementing the PID system, thanks to its robust framework for storing and linking data in a decentralized manner.

²⁰ IPFS Documentation. Access date 21.01.2024 <https://docs.ipfs.tech/>

²¹ CID Specification. Access date 02.02.2024 <https://github.com/multiformats/cid>

²² IPLD Documentation. Access date 02.02.2024 <https://ipld.io/docs/>

dPID: Technical Proposal

To establish a production-grade PID system, integrating IPFS with additional components that facilitate persistent linkage, version control, and contribution tracking is essential²³. This paper introduces dPID, a decentralized data identification and sharing system designed to enhance the management and curation of research artifacts and FAIR Data Objects. dPID represents a pilot project aimed at offering a reliable solution for persistent storage and data curation²⁴.

At its core, dPID leverages a sophisticated integration of the IPFS and IPLD technologies alongside the Sidetree Protocol²⁵. Sidetree acts as a protocol and API layer that can operate atop any data addressing system, enabling users to generate lookup databases and customize identities secured by PKI. Utilizing Sidetree, dPID adopts the standardized W3C DID, which facilitates the creation of universally unique identifiers that are resilient and verifiable²⁶.

dPID provides access to data that is both machine-actionable and human-readable, featuring a web interface built upon the JavaScript JSON-LD specification²⁷ and supported by an open-source API. The software suite responsible for resolving and minting dPIDs serves as a uniform kernel for every installation, ensuring that the system is completely open-source and operable on dedicated systems as-is. This setup not only enables efficient minting and resolution processes but also guarantees a public resolver functionally equivalent to any resolver within the decentralized network.

Enhancing the foundational persistence offered by IPFS, dPID introduces features like high throughput, strong consistency across the network, decentralized indexing, user-friendly URLs, and the incorporation of a Turing-complete blockchain. This blockchain component autonomously records the root CID of an IPLD data structure, ensuring metadata redundancy and immutability. Consequently, dPID promises reliable persistence for stored FAIR data objects and research artifacts, making it a comprehensive solution for decentralized data management and sharing. dPIDs provide:

- Verifiable ownership with ORCID-based person identification and incremental contribution record.
- Open network participation and metadata redundancy through peer-to-peer nature of IPFS.
- Compliance with FAIR principles via FAIR Data Object specification compatibility.
- “Vendor lock-in” removed in the context of data due to the removal of the singular provenance holder of the scientific record.
- Data integrity persistence with DHT, as it was described above.

²³ Anna Niehues, Casper de Visser, Fiona A Hagenbeek, Purva Kulkarni, René Pool, Naama Karu, Alida S D Kindt, Gurnoor Singh, Robert R J M Vermeiren, Dorret I Boomsma, Jenny van Dongen, Peter A C 't Hoen, Alain J van Gool, A multi-omics data analysis workflow packaged as a FAIR Digital Object, *GigaScience*, Volume 13, 2024, giad115, <https://doi.org/10.1093/gigascience/giad115>

²⁴ DeSci Codex. Access date 02.02.2024 <https://codex.desci.com/desci-codex/design-goals>

²⁵ Sidetree Protocol Documentation. Access date 22.01.2024 <https://identity.foundation/sidetree/spec/>

²⁶ Decentralized Identifiers (DIDs) v1.0. Core architecture, data model, and representations. W3C Recommendation 19 July 2022. Access date 22.02.2024 <https://w3c.github.io/did-core/>

²⁷ JSON-LD Documentation. Access date 21.01.2024 <https://json-ld.org/learn.html>

dPID Nodes: Brief Technical Outline

dPID Nodes is an open source²⁸ software suite written in TypeScript and published under MIT license²⁹. It acts over the IPFS HTTP API that should be provided by an API server. The API server currently used is Kubo - the open-source reference implementation of the IPFS node³⁰ that runs in the background provides CID resolution and retrieves the data from the decentralised network. dPID considers the identified data as a collection of versioned IPLD entries following RO-CRATE specification³¹. The simplified example of internal linkage is presented in Figure 3.

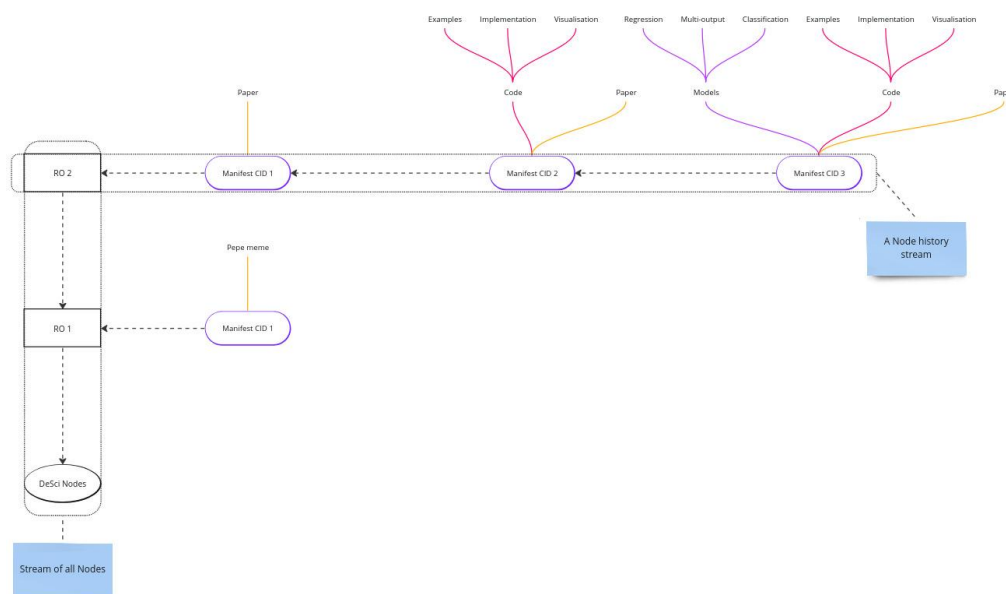


Fig. 3. Simplified internal bitstream storage diagram in dPID

From the perspective of the end user, these systems can be likened to folders that store research artifacts in a format-agnostic manner, with authorship and provenance details facilitated through integration with ORCID. The storage entries are catalogued in a distributed key-value store, each uniquely accessible via a dedicated persistent identifier that leverages underlying IPFS CIDs for addressing. The links between CIDs in Figure 3 illustrate the internal linked data structure that implements versioning and the history of changes. According to IPFS specification, every CID included in this schema is immutable, so the dPID metadata actually formulates a versioned repository for every FAIR Data Object it addresses.

As outlined in the documentation, dPID ensures deterministic resolution of PIDs to the internal CIDs of IPFS and their associated content through a DAG. This process allows the content to be immediately cached on the local database of the node it is accessed from. Building upon the IPFS framework, dPID Nodes additionally employ Ceramic³², a

²⁸ dPID Nodes Code Repository. Access date 23.01.2024 <https://github.com/desci-labs/nodes>

²⁹ MIT License. Access date 23.01.2024 <https://opensource.org/licenses/mit/>

³⁰ IPFS Kubo Code Repository. Access date 23.01.2024 <https://github.com/ipfs/kubo/>

³¹ RO-CRATE. Official Specification. Access date 02.02.2024

<https://www.researchobject.org/ro-crate/1.1/>

³² Ceramic Protocol Documentation. Access date 23.01.2024

<https://developers.ceramic.network/docs/protocol/js-ceramic/overview>

decentralized event streaming protocol, to create a graph-based distributed lookup database. This integration facilitates advanced data addressing using a combination of CID and DID.

When these dPID Nodes are queried through an HTTP API, they return JSON-LD objects. These objects can then be resolved by the web frontend, presenting the data in a human-readable format. This mechanism ensures that end users can easily access and interpret the stored research artifacts, benefiting from a seamless integration of decentralized storage technologies and modern web standards for data representation and access.

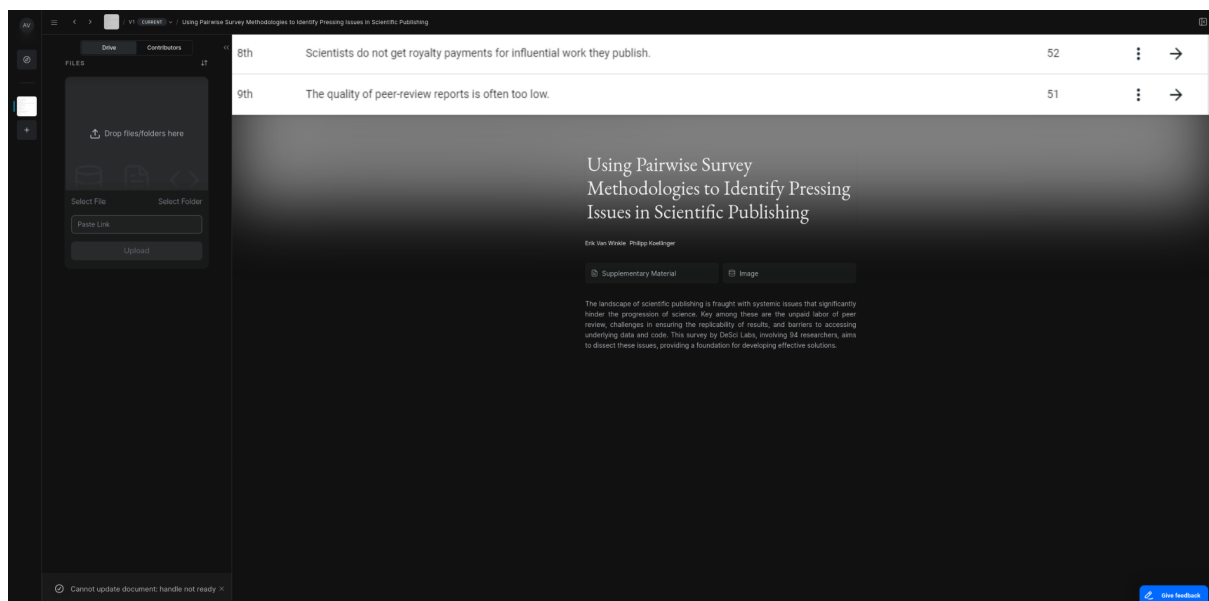


Fig. 4. Web frontend

The system in its current state should be considered as a pre-beta pilot version, work in progress on the deployment process and features list.

Challenges of Security

dPID is designed to be secure. Security is among the top priorities and there is high confidence in the current implementation due to the secure nature of its individual components. Openness and strong cryptography with options for transparent upgradability are already in place considering openly shared source code. It is beyond the scope of this paper to analyse in depth the strengths and weaknesses of the technological choices in the scope of security. Nevertheless, it hints that they are taken into consideration. Those choices of technologies also aim at feasible disaster recovery and the robustness of the system. A recent incident demonstrated the readiness of the system and here it is briefly mentioned. Elettra Sincrotrone Trieste, like similar large facilities, had a security incident of unknown nature but with limited and contained damages. An isolated series of ransomware attacks required swift actions and temporary restriction of experimental and non-essential services to the operation of the facility. Such a service was that of the dPID gateway which albeit not compromised, it had to get offline. The underlying technology of dPID allowed for the rapid migration to a new server, outside the facility and hosted in a different country. The migration

was rapid and easy (around 2 days including overrun and migration of the underlying CIDs and IPLD tracks), allowing for the continuation of the tests. Without focusing on the individual components that allowed this, the takeaway message is the decentralised nature of dPID which is in contrast to all other established, and especially, the centralised PID solutions.

Further Development

The dPID initiative for the future strives to:

- Provide a set of convenience libraries with a comprehensive API and viable examples letting software developers adopt the technology.
- Propose a deployment pipeline feasible for different use cases (OS packages, bundles, automated source builders, Docker scripts, etc.).
- Develop comprehensive documentation for end users, publishers and administrators letting them adopt and use the technology in the most configurable and flexible way.
- Propose social-oriented mechanisms such as social attestation, conflict moderation, and access control.
- Formal validation of the security and robustness models of dPID.
- Explore and adopt multiple data-oriented use cases such as migration, storage sharing, and consortium mechanisms.

Conclusion and Use Case Proposal

A description of the dPID system and the problems it solves is provided above. The Nodes web interface is openly published and used to distribute research artifacts published with open licenses. The experimental resolution endpoint lives on URI <https://nodes.desci.com>. For the installation and usage complexity, there are only qualitative estimations that exist now. However, because every participant of the dPID network should share completely the same software and all elements of the system, including the PID resolution point based on the website, the deployment complexity of the system should be estimated as approximately equal to the complexity of deployment and debugging of the standard organizational website in the cloud. The deployment pipeline uses highly standardised and documented solutions, such as Docker, so the process can be effectively controlled from the developers' side, and volunteer support through Github is also available.

The research introduces a compelling use case for dPID as a foundational infrastructure for the identification and hybrid (hardcopy + digital) distribution of artworks, incorporating provenance tracking and the engagement of social institutions. This initiative aims to facilitate the management of the provenance chain for photographic artworks, among others, that are distributed in hard copies, as Non-Fungible Tokens (NFTs), and in digital formats. The system is designed to link the author's personal PID with the authorized digital copy of an artwork and the CIDs of copies authorized for distribution, accompanied by provenance chain documents validated by social institutions.

This model is particularly suited for electronic use, incorporating linked data that includes technical details to identify unauthorized digital distribution and support implicitly legal promotion mechanisms, such as search engines. Also, it defines the basic identification procedures and unified initial provenance for the hardcopy distribution of the digital artworks,

with an option to extend the practices to the material artworks. Currently, the project is in the stage of defining its workflow model, with the metadata model already established. It is being explored as a potential application for dPID deployment, aiming for social recognition and validation.

Through this approach, dPID could offer a robust solution for artists and institutions to securely manage and distribute artworks. By integrating digital and hardcopy formats with a comprehensive provenance chain, the project seeks to enhance the trust and verifiability of artworks' distribution and ownership, leveraging decentralized technologies for greater transparency and security in the art world.

Due to the integrated attestation mechanisms, for external users, outcomes of the dPID project have approximately the same place as DOIs. Especially for the artworks, dPID extends the usual application area of the authenticity certificates, making them recognisable worldwide using any accessible resolution point held by any dPID network participant. From the author's point of view dPID has an outcome as a versioned, reproducible metadata handler for his authored data (digital photographs, articles, research objects, etc.). Acting locally, it simplifies minting and extends the functionality of systems like DOI, with out-of-the-box immutability.

Reviewer Comments

First Round

Reviewer 1

The paper is interesting and the narrative flows well.

A few suggestions for camera-ready, as they came when reading:

Abstract should be revisited; some part of the current text should go in the INTRODUCTION.

It is very important you use this sequence in the abstract.

- a) What is the problem in a given domain?
- b) Why is the problem interesting? (a few sentences, shortly)
- c) What other did/failed on, and what you are dealing with
- d) What is your solution? (one sentence, very short)
- e) Please, end the abstract with one sentence on what is the main output/conclusion of the paper, with what a reader will grasp once reading it, or any quantitative/qualitative conclusion based on solutions you evaluated or your own proposal.

- including IOT sensing

> ? IOT; spell-out every acronym, even trivial, at its first use

- Policy based recognition

> Policy-based recognition

Revisit Fig. 1, as not legible.

Also, add a text narrative to explain the figure, as not self-explicit.

End the INTRODUCTION with

- Please identify the limitations of the approach
- End the Introduction with a paragraph describing section-by-section, the structure of the paper

- The model in fig. 1 implements
- > The model in Fig. 1 implements

- In Table I a brief list of
- > In Table I, a brief list of

Ref; "This paper introduces dPID, a decentralized data identification"
Yes, I was looking for such a statement.

- by private key infrastructures (PKI)
- > by Private Key Infrastructures (PKI)
- ! same, for all acronyms

Ref: "W3C Decentralized Identifiers (DID),"

At this point the relation with DOI is not clear; you might clarify.

Redo Fig, 3 (low legibility) and Fig 4 (low quality)

In Section IV, you might identify a few inquiries raised during reading:

- what complexity of such tool
- what is the (manual) activity a user should perform (of any)
- who is in charge of conflicts solving

Position the output of dPID in the process of

- the authors
- the publishers

Reviewer 2

The contribution deals with a very relevant topic.

It might be useful to revisit and present a true state-of-the-art section,

- > Indicate what the others did
- > Why the existing solutions are not enough?
- > What is on the wish-list'?

> What is the purpose of your contribution?

Be sure that what you define as a target should be compared with the existing solution in a discussion section.

Indicate any limitations, boundaries and constraints on your proposal.

This will help in identifying the core of the proposal and drive the correct conclusion.

Considering Table 1, what is the process to have dPID standardized and adopted?

Are there any figures and concrete examples (experiments) concerning “Provenance Tracking, and Reproducibility”; details on these items will add value to the proposal.

Provide more details on future work as you mention “Currently, the project is in the stage of defining its workflow model,” Is there are feedback on the current status of the project from those supposed to be helped that can be reported.

Overall, the initiative is laudable, considering digital transformation.

Reviewer 3

The paper formatting needs some adjustments:

!! Correct the title; it is written using different font type and size – see template - Use 14 point Times New Roman in Bold

!! In the lines with the authors names, eliminate the numbers: '1st', '2nd', '3rd'

!! in the affiliation section, put the institution name on line 2, the City, Country information on line 3, and email addresses on line 4 (add the label 'email:' in front of the address)

!! Write the abstract as a compact (single) paragraph

!! Correct the keywords line (see template)

-- Index Terms—PID, Persistent Identifier, IPFS,

Must be:

--> Keywords-PID; Persistent Identifier; IPFS;

!!in Bold and Italic

!! Eliminate the citation from the start of Section 1. Write the text as a regular paragraph and cite the reference at the end.

!! use a uniform figure capture, e.g., "Figure 1. Traditional PID administration and implementation schema."

!! Apply to all

!! In text, use "Figure" at the beginning or inside of a sentence (do not use abbreviation); put a space between "Figure" and the number, such as "Figure 1"; correct all figure notations in text, all across the paper

!! correct the format for all the References (see template):

- some are incomplete; see the required format; for example, add the publisher information in [1], etc.

- list all the authors, if less than 6 [otherwise, use 'et al.' after the first author name] -> correct: [2], [3], [5], etc.

- provide the last access date for ALL on-line references in the form [retrieved: month, year] (close to the camera-ready deadline) -> correct: [3], [5], [11], etc.

!! Correct the format / grammar / punctuation errors. Some examples:

!! expand all the acronyms, even trivial, when firstly used (even in the abstract); examples: dPID, IPFS, IOT, etc.

-- and guidelines of FAIR (Findable, Accessible, Interoperable, and Reusable)

Must be:

--> and guidelines of Findable, Accessible, Interoperable, and Reusable (FAIR)

!! Apply this definition rule for all acronyms.

-- machine-actionability [2], [3].

--> machine-actionability [2][3].

-- from organizations such as OSTP, EC, and
--> from organizations, such as OSTP, EC, and
!! always this way, in similar constructions, across the whole paper

-- protocols, networks, etc. also reveals
--> protocols, networks, etc., also reveals

-- The model in fig. 1 implements a low
--> The model in Figure 1 implements a low

-- table redundancy [15], [16], reproducibility
--> table redundancy [15][16], reproducibility
!! Correct all citations across the paper

-- In Table I a brief list of
--> In Table I, a brief list of

-- Thus it should
be considered
--> Thus, it should
be considered

-- The IPFS (InterPlanetary File System) decentralized
--> The InterPlanetary File System (IPFS) decentralized

!! etc., etc.

=====

Second Round

We will have another verification before uploading with the Publisher.

SO:

1) Carefully follow the required paper format and editorial rules. Please see
<https://www.iaria.org/editorialrules.html>

2) !! Write the abstract as **one** compact (single) paragraph

3) The Abstract is too long. As already indicated by a reviewer, please move some of the details from the Abstract to the Introduction. "Abstract should be revisited; some part of the current text should go in the INTRODUCTION."

4) Do not create a new section of "LIST OF TERMS AND ACRONYMS". That section needs to be removed. Instead, every single acronym needs to be defined the first time it is mentioned in the paper text. For example, the first sentence in the INTRODUCTION should be " Persistent Identifiers (PIDs) are not only for documents and data."

5) "IOT — Internet of Things"-----> this acronym is normally written "IoT" We suggest to use this form in the paper.

6) Remove the entire section "STRUCTURE OF THE PAPER". Instead, at the very end of the INTRODUCTION, add a **very short** paragraph describing the structure of the paper. Something like " The rest of the paper is structured as follows. In Section 2, we explain the current status of the PID implementation landscape. In Section 3, we present..... We conclude our work in Section.... ".

7) In all the figures, especially Figure 3, please make the text inside bigger, for readability. Also, the figures are a bit blurry. Please try to use clear pictures.

8) "The experimental resolution endpoint lives on URI <https://nodes.desci.com>."
-----> do not insert any web links inside the paper text. Here, put this web link in the list of references as a regular reference and refer to it by its corresponding reference number, such as "The experimental resolution endpoint can be found at [...]. "

9) You were asked to "- list all the authors, if less than 6 [otherwise, use 'et al.' after the first author name] -> correct: [2], [3], [5], etc." Please look at Reference [5]. All those authors should **not** be listed, as they are clearly more than 6. Instead, please use this format:

[5] R. David et al., "Be sustainable: EOSC life recommendations for implementation of FAIR principles in life science data handling," The EMBO Journal, vol. 42, no. 23, p. e115008, 2023. [Online]. Available: <https://www.embopress.org/doi/abs/10.15252/emboj.2023115008>

Please check all references and update accordingly.

10) "Programa de Doctorado en Arte Universit`at Politecnico de Val`encia"-----> suggest to replace this by the English translation

11) For the authorship info, please remove "PhD" and other titles of the authors (to comply with the required paper format). Leave the full names of authors.

12) Lastly, we are including again all the review comments received. Please check them all one more time, one by one, to ensure they are all addressed.