DataSys 2024 - ACCSE 2024

VENEZIA MESTRE
14-18 April 2024

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

Andrey Vukolov
Scientific Computing Group
Elettra Sincrotrone Trieste

Erik van Winkle
GOFAIR Foundation Fellow,
DeSci Labs AG

Elizaveta Zhdanova
Faculty of Fine Arts,
Valencia Polytechnic University

George Kourousias
Scientific Computing Group
Elettra Sincrotrone Trieste

<u>Presenter</u>: **Andrey Vukolov**

e-mail: andrey.vukolov@elettra.eu

Presenter

Andrey Vukolov

- Software developer, industrial automation programmer, robotics engineer at Elettra Sincrotrone Trieste.
- Data Stewardship engineer and expert of ExPaNDS project of European Open Science Cloud.
- Andrey works on IPFS, Git and PID problems as side activities beyond the general work at Elettra.
- IPFS and Decentralised Science problems are in Andrey’s scope as outcomes of ExPaNDS.

Working Group: dPID.org

https://dpid.org

- The working group is developing decentralised, deterministically resolvable, sustainable, openly reproducible **persistent identifier** for any kind of digital entity or metadata. Guided and driven by DeSci Labs.
- The work of dPID project is based on well-established, known and documented distributed technologies, such as **IPFS**, **Blockchain**, **WWW Identification standards**.
- dPID leverages the core concepts of **FAIR**, making data and metadata, providing scalable, versioned identification that reduces dependencies from social contracts and institution-driven systems.

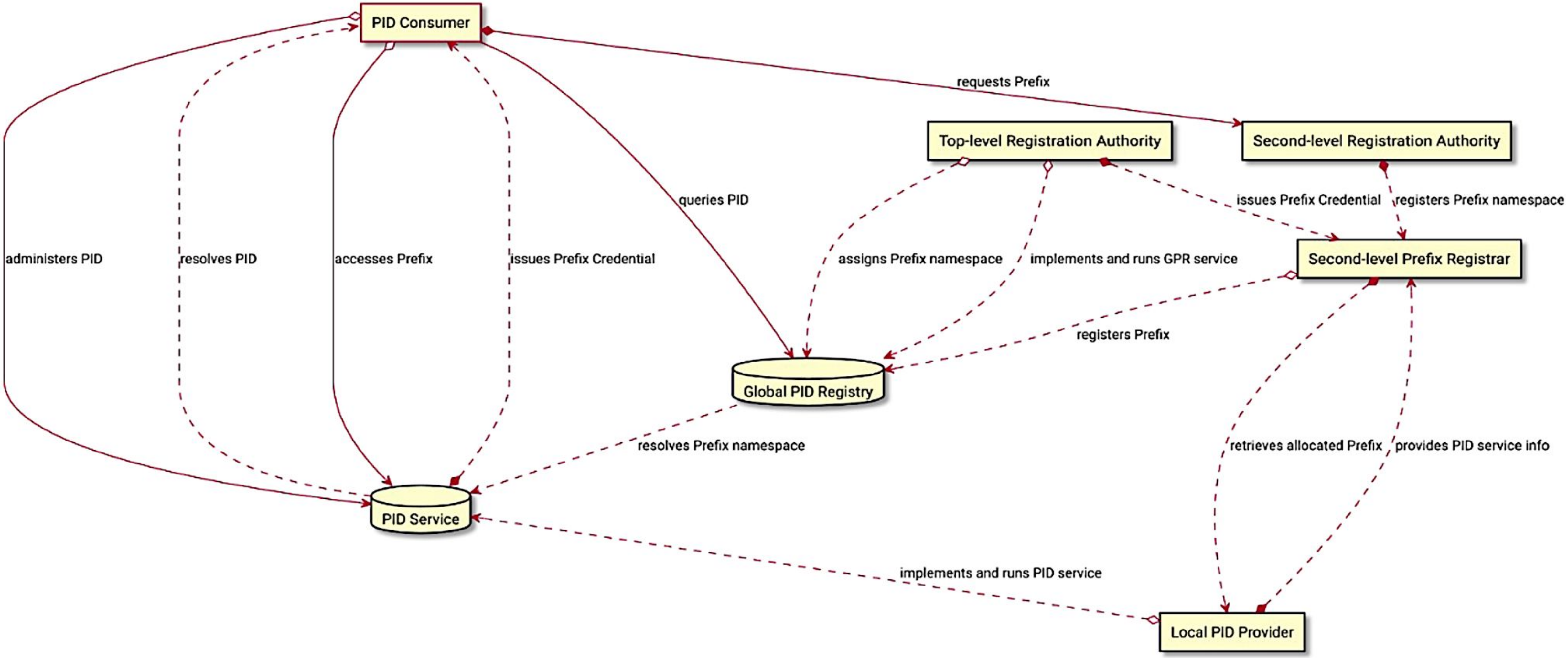
Prerequisites & Problems

- The first of the 15 FAIR Principles, Principle F1, states that "*Data and metadata must have globally unique, persistent, and resolvable identifiers*".
- Before 2030, the quantities of the PIDs needed to be minted, likely increase to trillions.
- The average scientific outcome may contain hundreds of thousands of data entities, including code, data, reviews, sensemaking statements, etc., both created by humans and captured automatically.
- The upcoming world of FAIR science the social aspects of persistence are becoming less important than the technical aspects.
- Growth of quantities of PIDs to mint reduces ability of the social contracts to ensure persistence.

PID Landscape: Providers

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
ORCIDhttps://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/	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

PID Landscape: Centralized Approach


PID Consumer
requests Prefix
Top-level Registration Authority
Second-level Registration Authority
queries PID
issues Prefix Credential
registers Prefix namespace
administers PID
resolves PID
accesses Prefix
issues Prefix Credential
assigns Prefix namespace
implements and runs GPR service
Second-level Prefix Registrar
registers Prefix
Global PID Registry
resolves Prefix namespace
retrieves allocated Prefix
provides PID service info
PID Service
implements and runs PID service
Local PID Provider

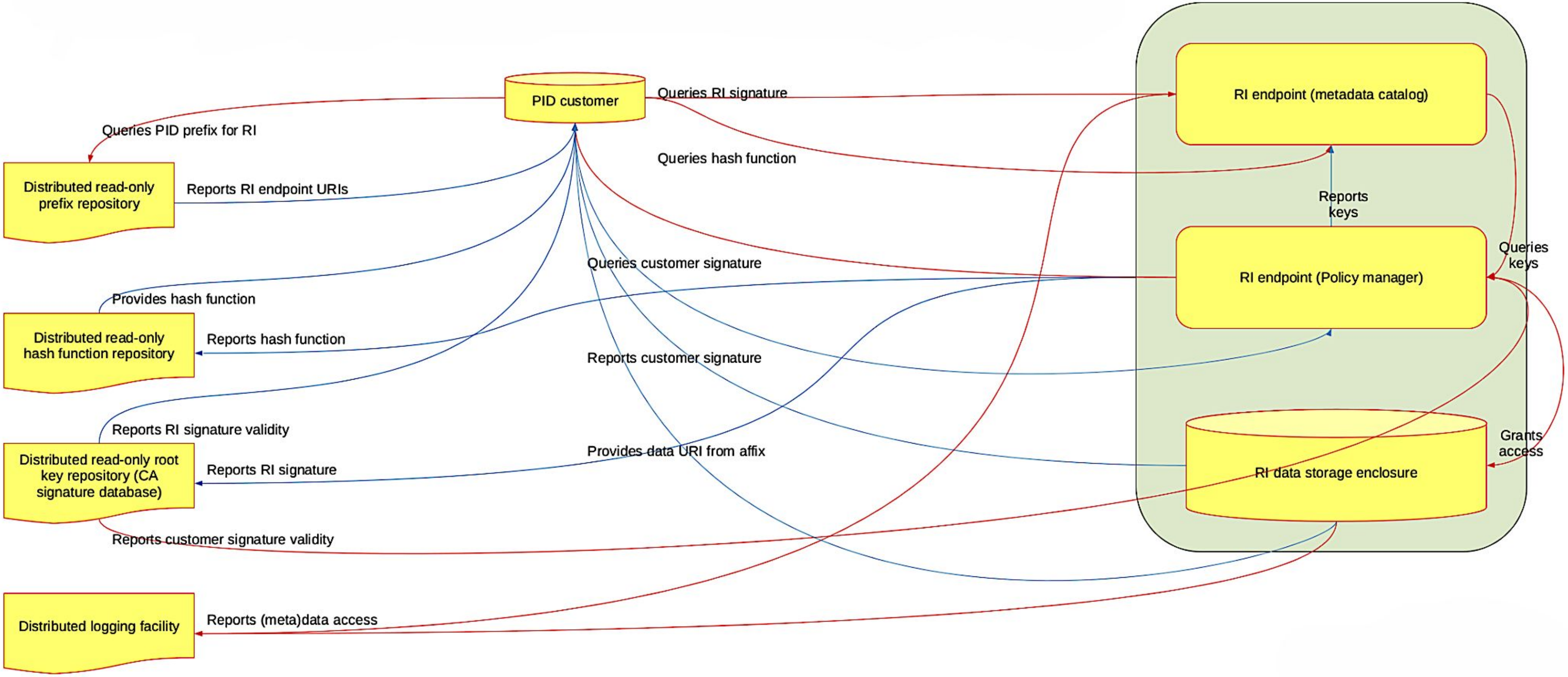
PID Landscape: Centralized Approach

- *At scales of ~1-10M of minted PIDs the centralized approach should be considered fit to its purpose*.
- Having **Global PID Registry** as a basic entity, implements a low number of high-risk singular points of failure.
- Centralized PID governance authorities are the only provenance holders of the underlying record, metadata, and addressed data in the case of **irreproducible PID with closed generation schema**.
- The currently implemented centralised PID systems obtain their persistency from the social contracts, so link rot, inconsistent resolution, content drift, etc. prevail.

PID Landscape: Federated Approach

- Combines the existing centralized governance model with a federated social-driven network of formally independent PID registrars.
- Implements prefix-based resolution model with segmented social network where **every registrar is responsible for the dedicated prefix/segment**.
- Each registrar maintains **its own provenance authority**.
- **Most popular implementation model in the modern world** with non-reproducible PIDs (DOI, Handle).
- Allows **expansion of the computational power** of the entire system, primarily in the aspect of data replication.
- Limited support for interoperability and cross resolution (UNIRESOLVER).

PID Landscape: Decentralized Approach


PID customer
Queries RI signature
Queries PID prefix for RI
Queries hash function
Distributed read-only prefix repository
Reports RI endpoint URIs
RI endpoint (metadata catalog)
Reports keys
Queries keys
RI endpoint (Policy manager)
Queries customer signature
Provides hash function
Distributed read-only hash function repository
Reports hash function
Reports customer signature
Reports RI signature validity
Grants access
RI data storage enclosure
Provides data URI from affix
Distributed read-only root key repository (CA signature database)
Reports RI signature
Reports customer signature validity
Distributed logging facility
Reports (meta)data access

PID Landscape: Decentralised Approach

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.

- Necessitates uniformity in backend technology and storage models across all resolvers and PID generators, likely deployed locally on the end-user's side.
- Replaces asking a single resolution point: "*What is the content stored at this location?*", with asking a swarm P2P network: "*Can you tell me how to find the content to reproduce my PID?*".
- Utilizes special technologies like DHT to obtain mathematically-driven delivery of binary data objects stored "everywhere" in the P2P network.

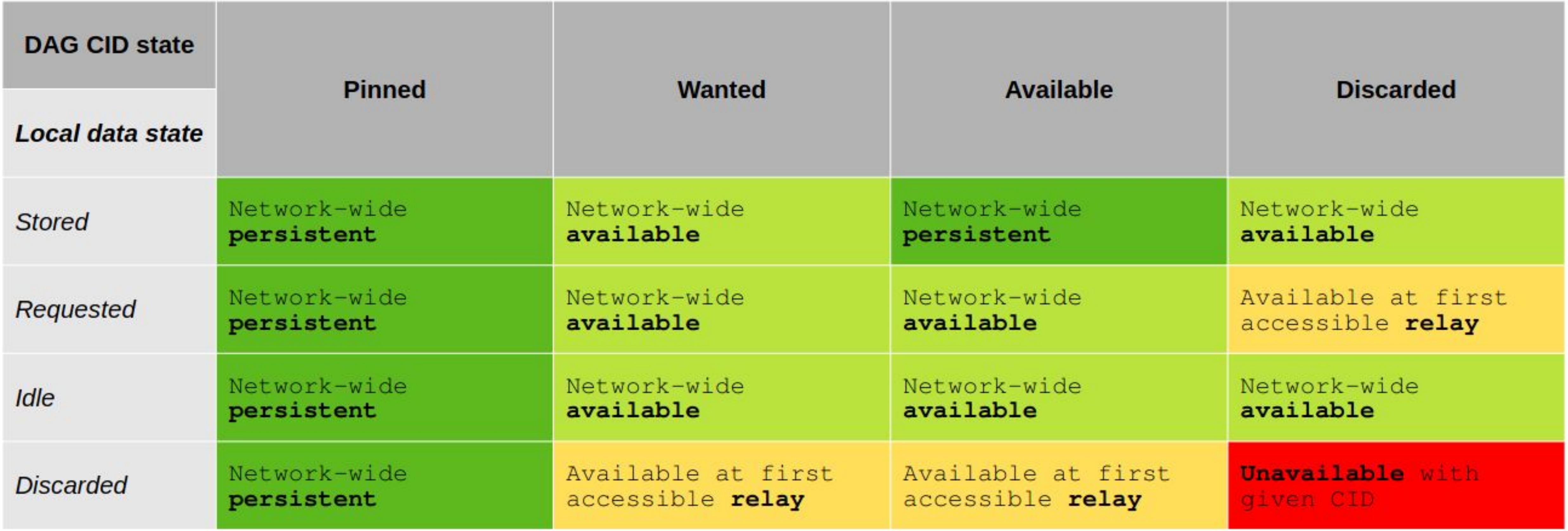
dPID: Decentralized Persistence

Decentralized Persistent Identifier (**dPID**) is a technology that combines distributed software components and technologies into the novel data identification pipeline:

- **InterPlanetary File System** (IPFS) creates immutable, mathematically proven storage layer for any kind of data which could be represented as a bitstream. Content Identifiers (CID) in the IPFS network are immune to content drift and drastically mitigate the effects of link rot.
- **InterPlanetary Linked Data** (IPLD) provides persistent tree structures to interlink and version the data. IPLD ensure the persistence of linkages between data objects ensuring also FAIR compliance status of the identified set of data and metadata.

IPFS Content Identifiers: Persistence

The stored data is as persistent as at least one network participant requests it.



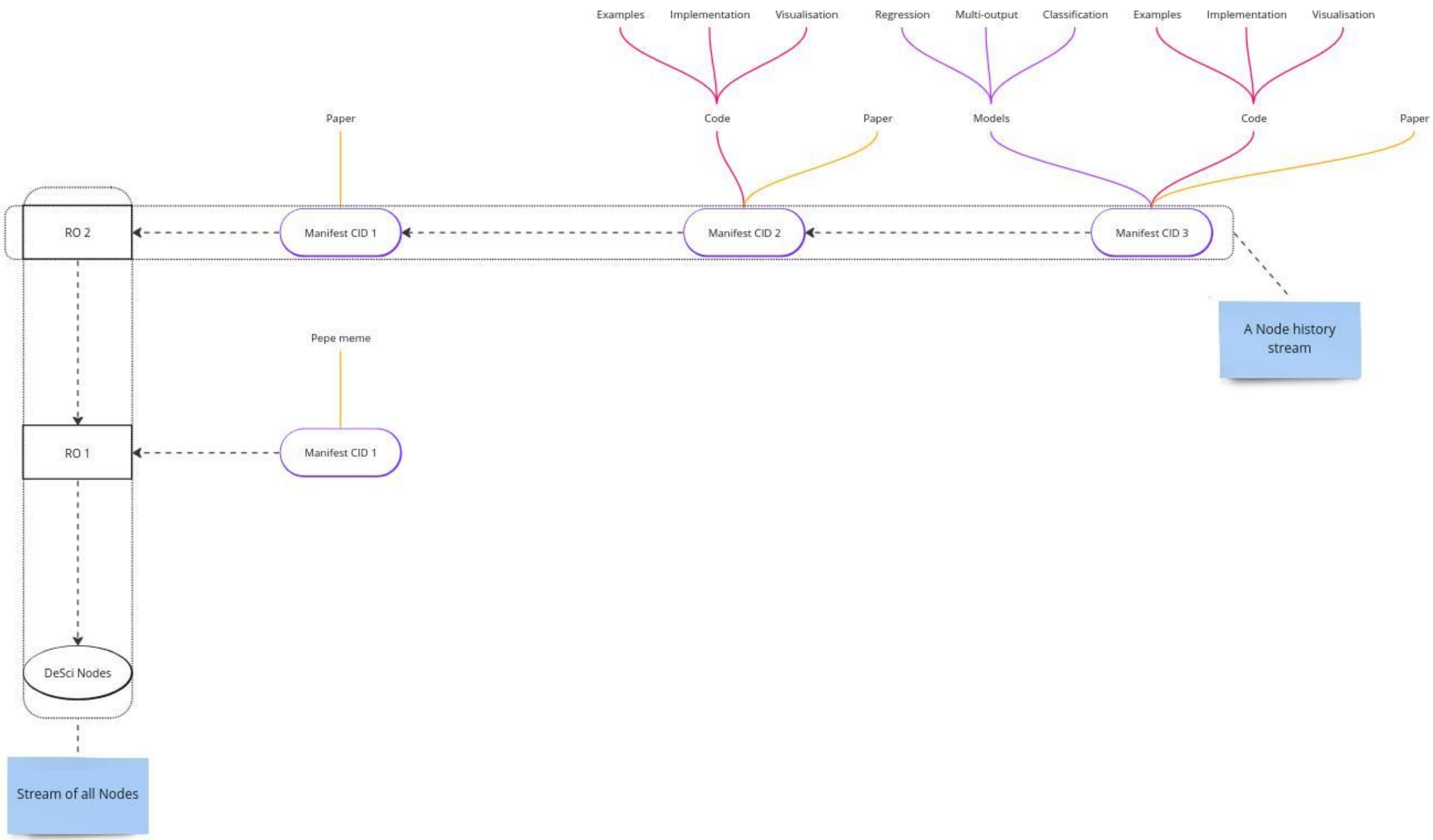
DAG CID state ***Local data state***	**Pinned**	**Wanted**	**Available**	**Discarded**
Stored	Network-wide **persistent**	Network-wide **available**	Network-wide **persistent**	Network-wide **available**
Requested	Network-wide **persistent**	Network-wide **available**	Network-wide **available**	Available 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

dPID: Decentralized Persistence

Decentralized Persistent Identifier (**dPID**) is a technology that combines distributed software components and technologies into the novel data identification pipeline:

- Turing-complete **Sepolia blockchain** implements persistent ledgering. It stores the data structures addressed by CID in a highly persistent, secure and decentralized fashion over the root hash of IPLD Directed Acyclic Graphs.
- Sidetree protocol provides high-level addressing over the blockchain entries. It introduces W3C Decentralized Identification standards to create globally unique, user-controlled identity and manage associated Private Key Infrastructure metadata, all without the need for social contracts.

dPID: Bitstream Storage Model



dPID: Technical Proposal

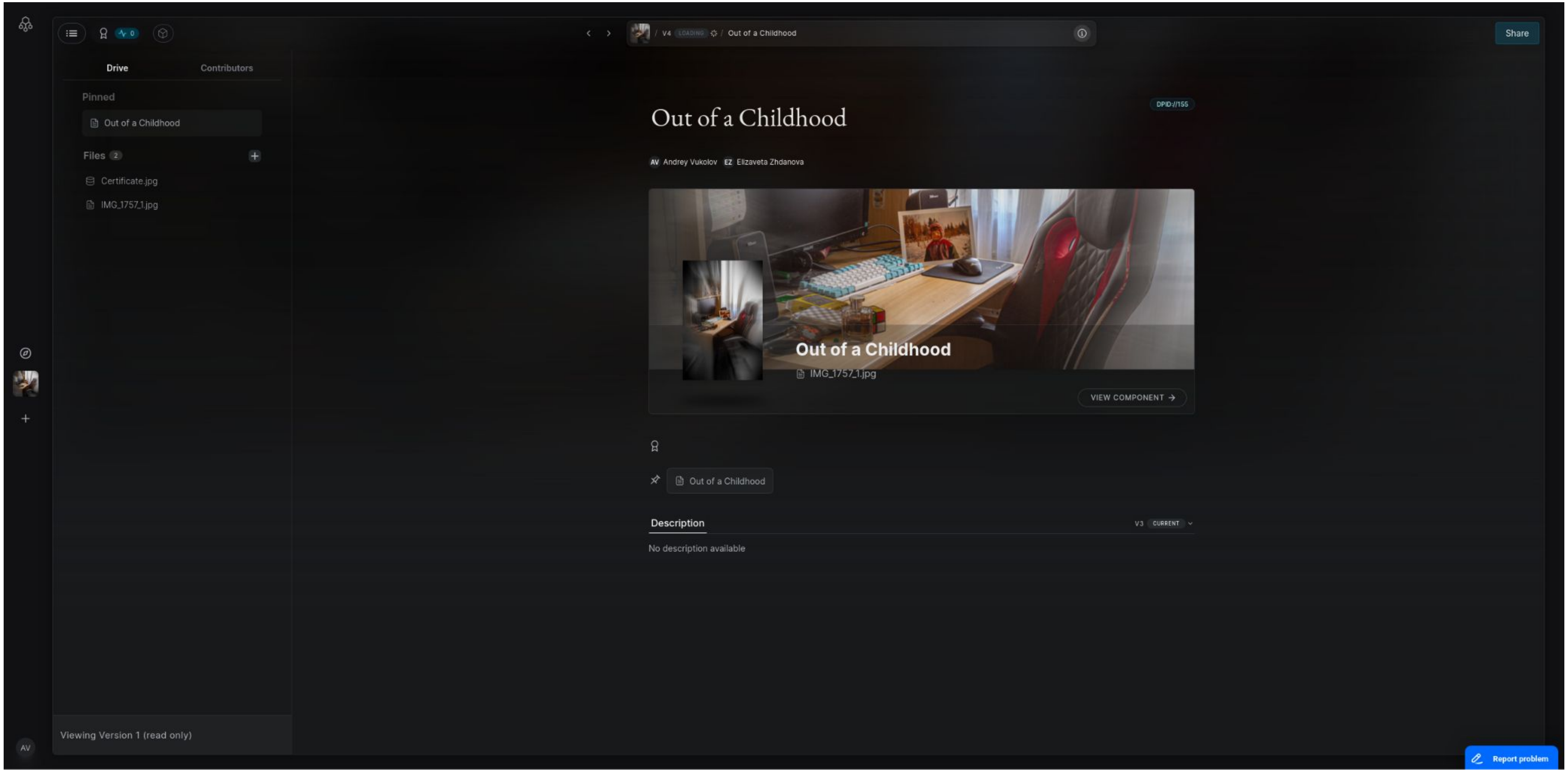
DeSci Nodes - the pilot web solution implementing dPID as the identification, storage and versioning engine for Research Objects (RO-CRATE-compliant).

- JSON-LD-compliant internal data representation for machine-actionability and interoperable HTTP API.
- ORCID-compliant initial user identification (in the pilot implementation) for verifiable ownership and incremental contribution lists.
- Open network participation and metadata redundancy through peer-to-peer nature of IPFS.
- Partial compliance with FAIR Data Object specification.
- "Vendor lock-in" removed in the context of metadata due to the removal of the singular provenance holder of the scientific record.

dPID: DeSci Nodes application

- Open-source Web application written in TypeScript.
- Uses known and well documented IPFS Kubo as an internal API server.
- From the perspective of the end user, can be likened to directories that store research artifacts in a format-agnostic manner.
- Accounts contribution through ORCID in the experimental implementation.
- Formulates a versioned artifact repository for every FAIR Data Object addressed with the given dPID.
- Ensures deterministic resolution of dPID to the internal IPFS CIDs and their associated content through a DAG.
- Employs Ceramic to create a graph-based distributed lookup database. This facilitates addressing using a combination of CID and W3C DID.

dPID: DeSci Nodes user interface



DPID://155

Out of a Childhood

AV Andrey Vukolov EZ Elizaveta Zhdanova

Description

No description available

Viewing Version 1 (read only)

Use case: Artworks

- dPID was introduced at Le Vie delle Foto as an experimental foundational infrastructure for the identification and hybrid (hardcopy + digital) distribution of artworks.
- The initiative facilitates the management of the provenance chain for photographic artworks distributed in hard copies, and in digital formats.
- Extends the regular certification of authenticity issued by social institutions (galleries, traders, etc.) with versioned, immutable digital metadata with immutability ensured via IPFS.
- Uses integrated social attestation mechanisms to build the social web of trust over the decentralized network.

Roadmap

- Provide a **set of convenience libraries** with a comprehensive API and viable examples letting software developers adopt the technology.
- Propose **automated deployment pipeline** feasible for different use cases.
- Develop **comprehensive documentation** for end users, publishers and administrators letting them adopt and use the technology.
- Propose social-oriented mechanisms such as **social attestation**, **conflict moderation**, and **access control**.
- Pass the software for **formal validation of the security and robustness**.
- Explore and adopt multiple data-oriented use cases such as migration, storage sharing, and consortium mechanisms.

Special thanks: **Linda Simeone**, Le Vie delle Foto

Thank you for your attention!

Elettra Sincrotrone Trieste

UNIVERSITAT POLITÈCNICA VALÈNCIA

GO FAIR Foundation