Elettra Sincrotrone Trieste

ExPaNDS

European Open Science Cloud Photon and Neutron Data Services

Openly reproducible Persistent Identifiers (PIDs) as a factor of FAIRness in data sharing practices

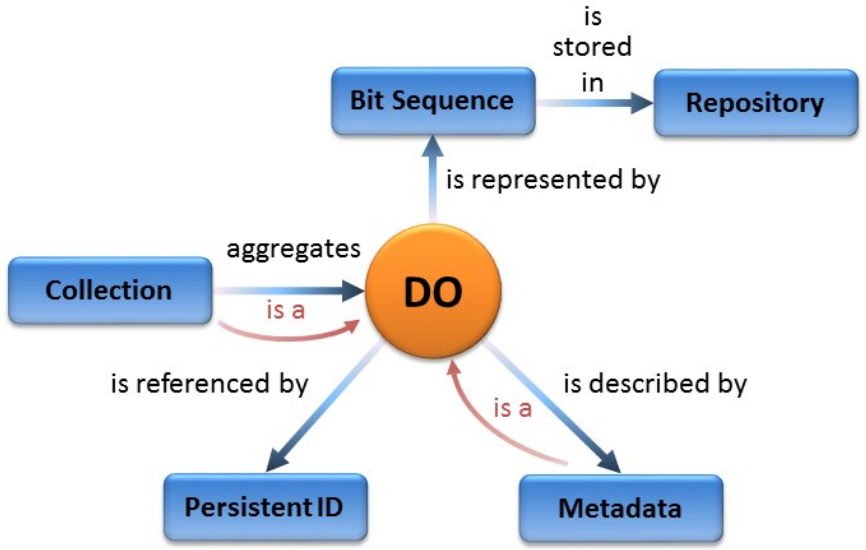
Andrey Vukolov, ELETTRA Sincrotrone Trieste

18/06/2020

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

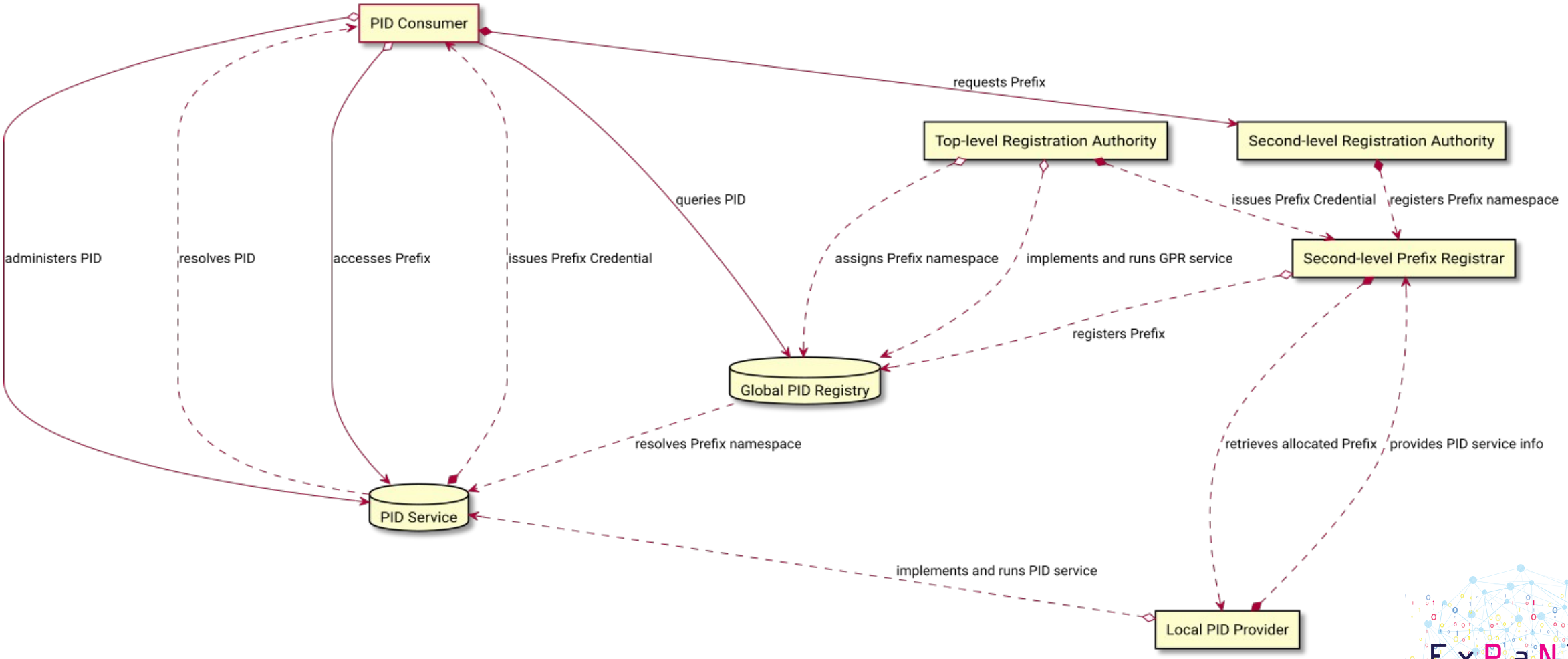
Persistent identification as the factor of FAIR

- The access model is not defined explicitly;
- In terms of resolving, in current model all kinds of the existing data representing the given Object are stored separately;
- The data represented on the diagram are existence-agnostic about each other;
- The Client obtains provenance from the trusted service that is actually self-signed.



This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Generic PID administration model



This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PID in FAIR context: current state

In context of **Findability**:

- PID (esp. DOI) provides one-and-the-only address endpoint for access to the Digital Objects;
- The **namespace**-based model with prefixes provides openly distributed list of publishing authorities;
- Findability of the both metadata and scientific data provided by a secondary authority is explicitly dependent from implementation;
- It is not possible in general to do reverse lookup or re-creation of PID from the metadata stored on RI's side.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PID in FAIR context: current state

In context of **Accessibility**:

- Each node in the PID administration workflow is a separate **point of failure** because any technical failure leads to data loss;
- The resolving tasks are **concurrent**;
- The **Global Registry** and **Prefix Provider** are the only sources of provenance for the given PID namespace;
- Irreproducibility of the prefixes creates a situation when the customer allows the root authority to **declare itself** as trusted source.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

PID in FAIR context: current state

In context of **Interoperability/Reusability**:

- The RI is separated from the PID namespace authority it uses except the case it implements the trusted service;
- Closed PID resolving algorithm reduces redundancy avoiding the RIs to act voluntarily as spare resolver for any kind of authority;
- The closed prefix assignment model makes once issued PID mutable (on RI's side) but irreproducible on the authority's side.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

How to define reproducible PID?

- The reproducible PID's **prefix** (or namespace) is **defined by known algorithm**/standard from openly accessible RI's metadata like `<prefix>{/}<body>{/}affix`;
- The prefix is **signed** digitally by RI's public key. The key is considered as immutable token and it could not be reproduced;
- There is a **distributed database** with every involved RI storing the local updated copy, containing all available prefixes;
- The database links prefixes to RI shared metadata from which **every prefix can be reproduced** on the client's side.

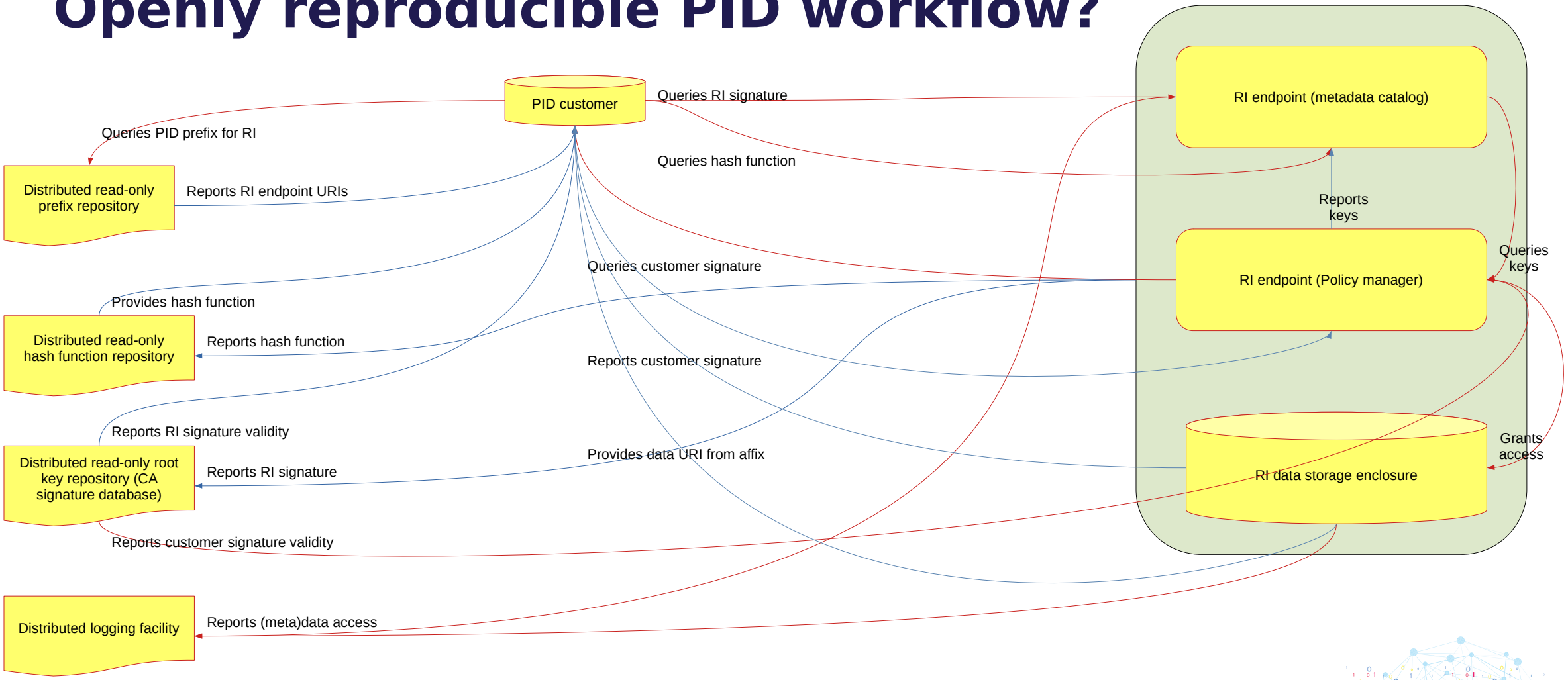
This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

How to define reproducible PID?

- The PID's body is **reproducible** from the addressed data/metadata themselves by the openly accessible algorithm;
- Authentication/authorization is yielded to RI explicitly but both operations are **signed** with customer's public key;
- Both customer and RI **report** access and authentication to the distributed database with immutable records, storing public keys as hash seeds;
- PID is encouraged to be **self-describing** and **versioned** by the affix to control integrity.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Openly reproducible PID workflow?



PID structure: `<prefix>{/}<body>{/}<affix>`

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

How reproducible PID affects FAIRness?

- *[AIR]* Prefixes reproducible from RI's public key leads the root authority only to issue the immutable keys for RIs and customers;
- *[FA]* RIs have an opportunity to voluntarily share space and resources for storage of prefix repository and reference tables;
- *[AIR]* Every RI acts like spare endpoint for the whole metadata infrastructure but also encouraged to preserve the actual data;
- [FA] The reproducible PID is encouraged to be resolvable to another PID as part of metadata.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

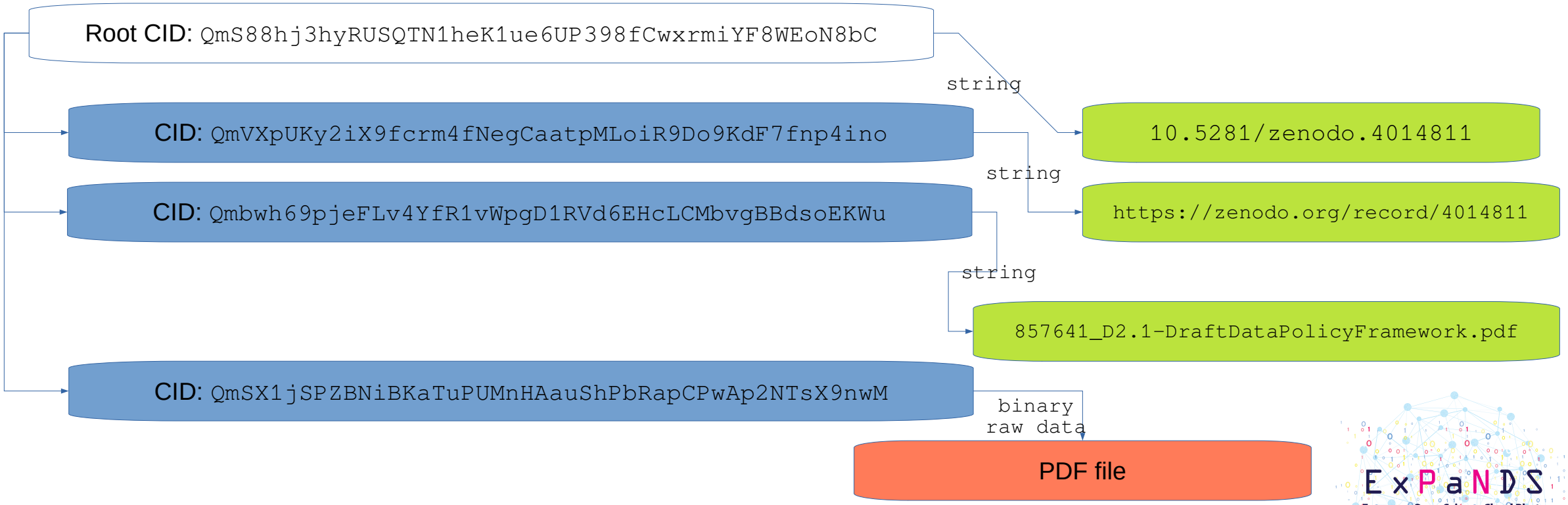
How reproducible PID affects FAIRness?

- *[FA]* In context of resolving every RI verifies keys for any other RIs;
- *[IR]* The client gathers integrity check with post-download verification reproducing PID body;
- *[IR]* The client verifies RI's signature obtaining the public key from the repository;
- *[FIR]* Encouraging self-describing PIDs the RIs can build PID graph with their own instrumentation and proposal namespaces defined through the affixes.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Example: CID as Reproducible PID

Root CID: `QmS88hj3hyRUSQTN1heK1ue6UP398fCwxrmiYF8WEoN8bC` addresses named virtual objects linked into the graph. The CID in this example is resolved by IPFS without a prefix.



This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Reproducible PID: possible concept?

As an example the following DOI-based schema could be created:

`10.XXX/<CID>/<InstrumentID>.<ProposalID>.<Version>`

Every element excluding affix is resolvable on both client and RI side.

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

Elettra Sincrotrone Trieste

Thank you for your attention!

This project receives funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 857641

ExPaNDS
European Open Science Cloud Photon and Neutron Data Services