Decentralized Persistent Identifiers: a basic model for immutable handlers

Miguel-Angel Sicilia, Elena García-Barriocanal, Salvador Sánchez-Alonso, Juan-José Cuadrado

CS Department
University of Alcalá (Madrid)

June 14, 2018

Outline

- PID systems and decentralization.
- A model of decentralized PIDs using verifiable claims.
- Example deployment using IPFS.
- Conclusions and outlook.

PID systems functions

- **Persistent Identifier (PID)** systems aim at providing long-lasting **references** and **resolution** to digital objects.
- They are actually also **resolving to metadata**.
- Examples are the Handle System (DOIs, DSpace) and Archival Resource Key (ARK).
- They are (in some sense) centralized, either in the resolution itself and/or in the assignment of domains for minting ids.

What do we mean by decentralized?

- Two types of decentralization[1]:
  - Architectural: how cooperating and equal are the computers supporting the service?
  - Political: who is controlling those computers?
- For example, the Handle System is politically centralized (patented technology with a public license, service agreements) and architecturally hierarchical (a global handle service that federates local services).
- Decentralization is important for fault-tolerance, attack resistance, collusion resistance and (arguably) sustainability.

[1] Adapted from: https://medium.com/@VitalikButerin/the-meaning-of-decentralization-a0c92b76a274

Is ARK decentralized?

- Free service, open to any archival organization.
- The Name Assigning Authority Number (NAAN) Registry is centralized (California Digital Library as BDFL?).
- Each NAAN is responsible for its ids, and there is no cooperation built-in or incentive mechanism to maintain them (beyond re-direction).

Is ARK decentralized?

`http://gallica.bnf.fr/ark:/12148/bpt6k5834013m`
`http://gallica.bnf.fr/ark:/12148/bpt6k5834013m?`
`http://other.example.org/ark:/12148/bpt6k5834013m`
`http://other.example.org/ark:/12148/bpt6k5834013m?`

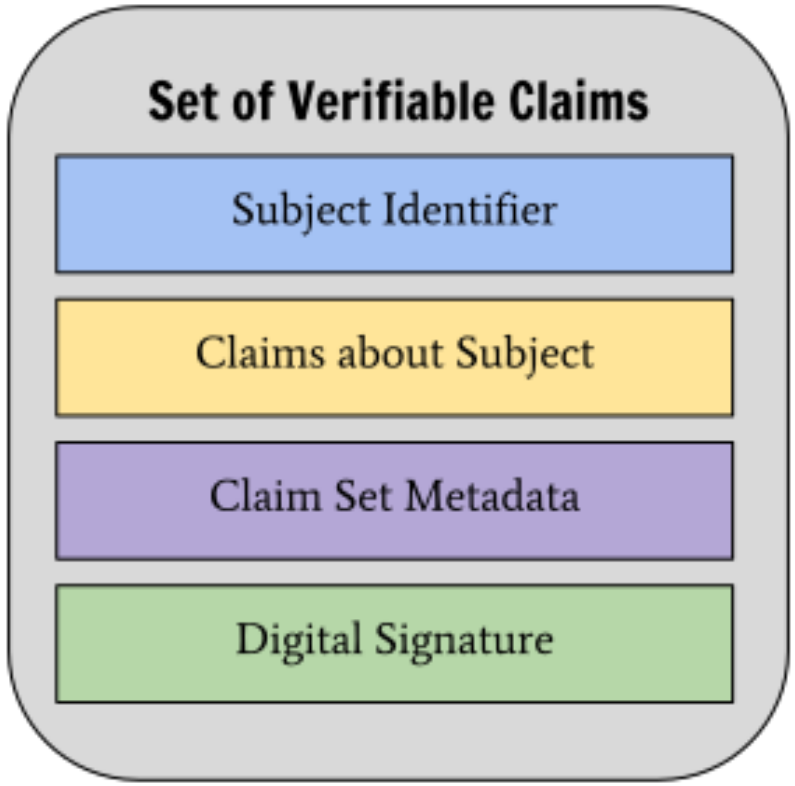
- Shall the metadata requests return the same info?
- How can we trust they return the same object?
- Can we be sure the returned resources have not been tampered?
- All this still relies on the Name Mapping Authority Hostport (NMAH), and NMAH do not cooperate among them.

Proposal in a nutshell

- Remove any need of a centralized organization in assigning and maintaining identifiers and resources ("anyone can mint a PID").
- Provide provenance proof built-in, i.e. support digitally verifiable claims (including metadata).
- Make PIDs and objects truly immutable and (eventually) permanent.
- Provide a platform for building incentive systems for cooperating in maintaining the resources and their links.

PIDs as verifiable claims

- A **verifiable claim** is a claim that is effectively tamper-proof and whose authorship can be cryptographically verified[2].
- A verifiable claim that associates an identifier to a digital object can be considered a PID iif **the object referenced is immutable**.



[2] https://www.w3.org/TR/verifiable-claims-data-model/

Byte content model and verifiable claims

- The byte content of digital resources (and the claims about them) can be identified by **content hashes**.
- Verifiable claims are thus claims associating some metadata (maybe including legacy existing PIDs) to a content hash.
- These claims and the objects they reference are deposited in a decentralized file system as IPFS, considered (pragmatically) permanent[3].

[3] BitTorrent-style, this requires some incentive mechanism or stewardship.

Decentralized file systems and IPFS

- Decentralized file systems share file storage in the net.
- The Interplanetary File System (IPFS) is one of them, inspired on the BitTorrent P2P network.
- Files and folders are referenced and stored using content hashing, and storage is distributed using Merkle DAGs.

```
ipfs add -r .
added QmSTuTEThyESv...HPD8wSHVy7 test-ipfs/donut.jpeg
added QmSR9MJ5resQLj...jZeZDtW test-ipfs/purse.jpeg
added QmUNLLsPfHbsf67hvA3Nn test-ipfs/folder1
added QmaKZ3dnc9ejBdGg...Ljk6gXUnk9sdM9 test-ipfs
```

Decentralized file systems and IPFS:pinning

- IPFS is intended to provide a global P2P file system where equal peers collaborate in maintaining the files, with no central or specially privileged nodes.
- As files are retrieved in nodes they create local replicas.
- Retrieval can be optimized using the swarm of nodes or considering locality.
- If a node "pins" a particular file, it does not garbage collect it, so it becomes a "custodian".

The identifier system

- Claims are assumed to refer to valid contents via content hashes (this may be enforced including smart contracts in the solution).
- The **claim itself has a content hash and this can be considered a PID**. It resolves to a resource+metadata and is unique.
- Anybody can sign and deposit claims, i.e. mint PIDs, no trust required.
- Users (journals, institutions, individuals) may also reference and "pin" these identifiers.
- Digital signatures make those claim's issuer verifiable.
- Existing identifier schemas and arbitrary metadata can be mapped inside the claim content.

Example: minimal claim

- Only the mandatory elements of a verifiable claim plus a specific type and the IPFS link.
- This asserts that the signer has minted an PID and nothing else.
- With `ipfs add` we have the PID.
- Syntax to be revised to comply with W3C specs.

```
{
  "@context": ["https://w3id.org/security/v1",
               "fs://ipfs/Qre...", ...],
  "id": "https://bnf.fr/credentials/3732",
  "type": ["Credential", "PID-ipfs"],
  "claim": {
    "id": "/ipfs/QmUmg7BZC1YP1...",
    "link": {"/": "/ipfs/QmUmg7BZC1YP1..."},
   },
  "signature": {...}
}
```

Example

```
{
  "@context": ["https://w3id.org/security/v1",
               "fs://ipfs/Qre...", ...],
  "id": "https://bnf.fr/credentials/3732",
  "type": ["Credential", "PID-ipfs"],
  "issuer": "https://bnf.fr", "issued": "2018-01-01",
  "claim": {
    "id": "/ipfs/QmUmg7BZC1YP1...",
    "link": {"/": "/ipfs/QmUmg7BZC1YP1..."},
    "ark": "ark:/12148/bpt6k5834013m",
    "Content-Type": "application/pdf",
    "charset": "utf-8",
    "metadata": {"/": "/ipfs/QmUmg56yiuo9.."}
  },
  "signature": {...}
}
```

Conventions used

- In the `@context` a reference to a IPFS file (in URI form) contains the schema for `type` (e.g. `PID`) and the properties inside `claim` (e.g. `ark`).
- The `id` in the claim is the same as in `link`. The former is part of W3C model and the latter is an IPLD link for IPFS.
- Additional metadata about the resource is included in the claim, in this case required to process the information.

Incentives

- Popular items in IPFS often retrieved continue to be live as in BitTorrent in any case.
- Voluntary basis: "pinning" resources and verifiable claims by institutions with archival missions.
- Clients that reuse idle space and computing as in SETI@Home and others[4].
- Filecoin: a blockchain solution for renting storage space with payments for storage and retrieval.

[4] https://en.wikipedia.org/wiki/List_of_distributed_computing_projects

Conclusions and outlook

- Current PID systems rely on conventional Web servers that do not cooperate in the maintenance of identifiers and resources.
- Some PID systems have a degree of political centralization.
- We have showed how the W3C verifiable claims models can be combined with distributed file systems for a decentralized approach to identification and de-referencing.
- Eventually, that solution can be complemented with incentive mechanisms for a sustainable solution.
- Beyond this: smart contracts to enforce additional functionality and conventions in a trustless context.