Distributed Identification of Knowledge.

How Can Decentralized PIDs Save Human Legacy?

Andrey Vukolov
Elettra Sincrotrone Trieste, Trieste, Italy

Sergio Santamarina
Universidad Nacional de Jose C. Paz, Buenos-Aires, Argentina

Jonas Söderberg
National Bioinformatics Infrastructure Sweden, Uppsala University, Sweden

Elettra Sincrotrone Trieste

15 Años UNPAZ Universidad Nacional de José C. Paz

UPPSALA UNIVERSITET

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

I call the ongoing change "documedia revolution" because it is based on the intersection between the increase in documentality, i.e. the production of documents as a constitutive element of social reality, and the growth of the media, which today no longer work as one-to-many but as many-to-many.

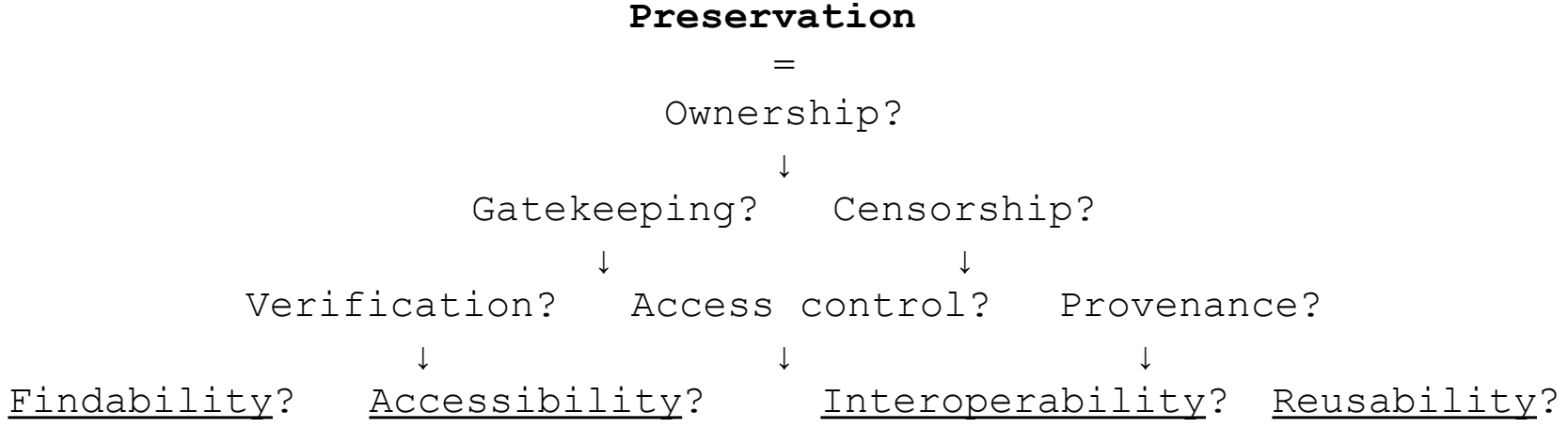
(Ferraris, 2023)

Why Preserve Knowledge?

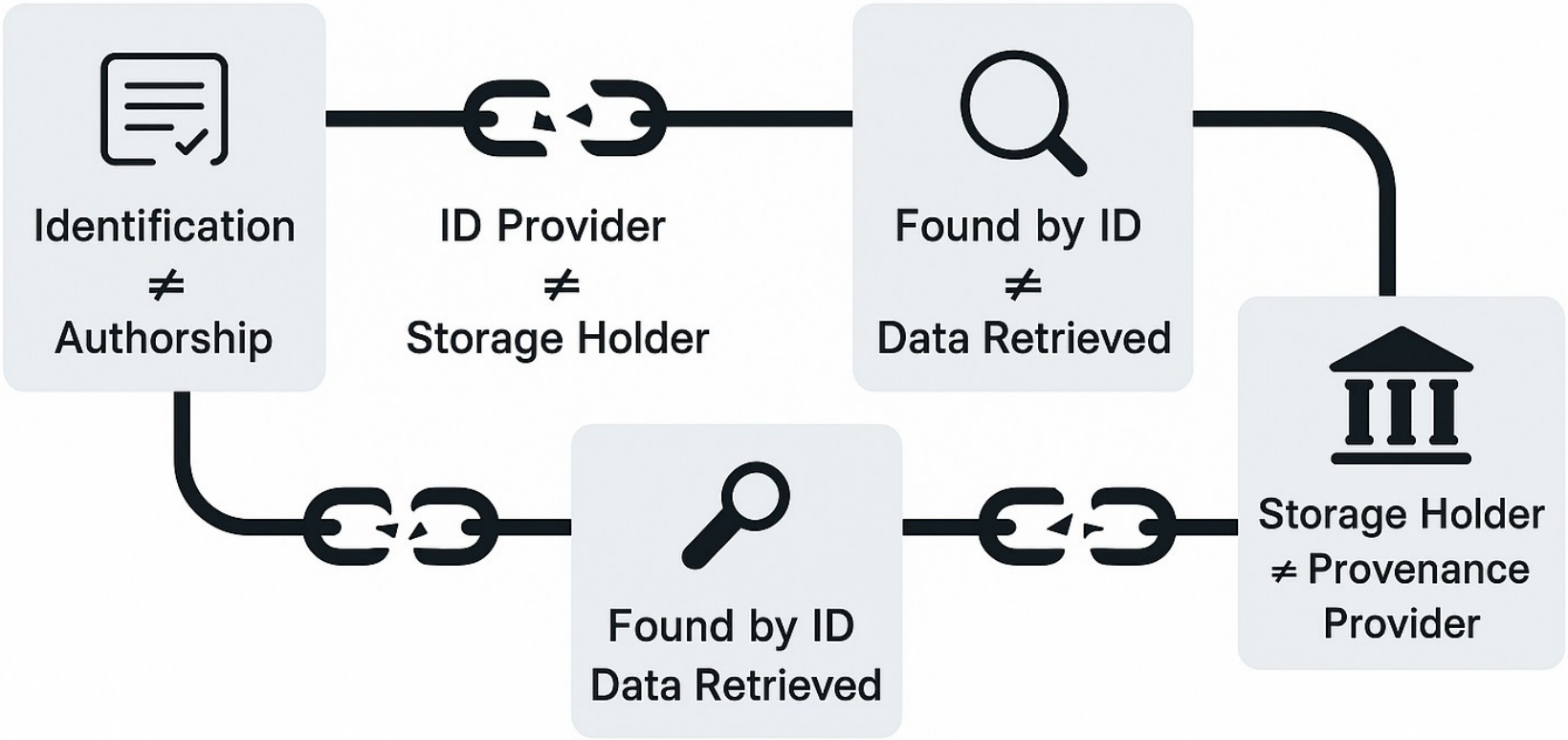
- **Knowledge is humanity's memory**
  - It enables progress, avoids repeating mistakes, and sustains culture. When data vanishes, science regresses, trust erodes, and lives can be impacted.
- **Preservation is a collective responsibility**
  - It ensures knowledge remains findable, accessible, and reusable for future generations.

Knowledge Preservation

- As an abstraction, knowledge itself is represented by data.
- Data (and thus knowledge) is produced by research
- Preservation must happen where research happens
- Not left solely to third-party organizations.



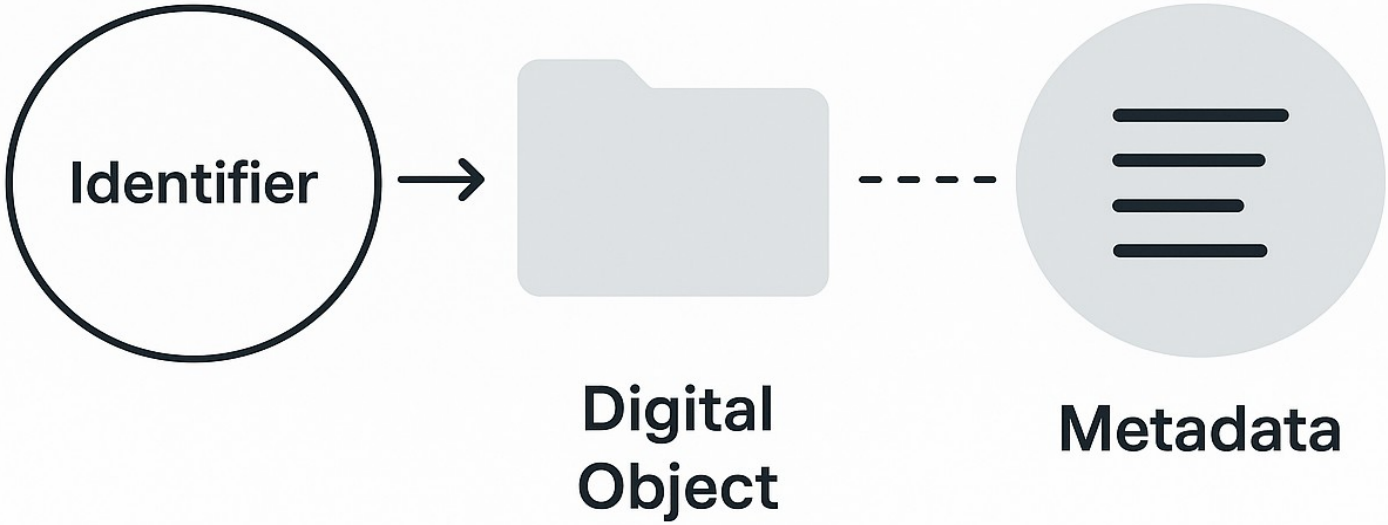
Key Preservation Gaps



Risks

- **Willing Censorship**
  Loss of access due to political or commercial decisions
- **Social Inconvenience**
  Censorship or deletion of inconvenient knowledge
- **Hidden Players**
  Lack of transparency and accountability in stewardship
- **Neglection (willing or unwilling)**
  Lack of maintenance leading to the losses of data

Identifiers in Archiving

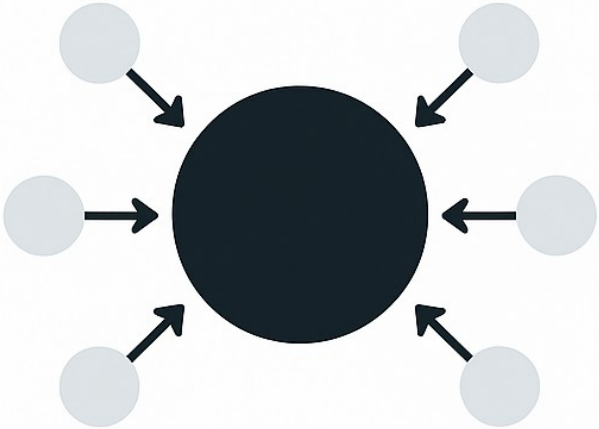


Identification and Preservation

- **Identification enables discovery**
  Unique identifiers (e.g., DOIs, ARKs) make knowledge findable and citable.
- **Preservation ensures longevity**
  Data and knowledge must remain accessible, authentic, intact over time.
- **Identification ≠ Preservation**
  Assigning an identifier does not guarantee that the data is actually stored or maintained.
- **Both are necessary and complementary**
  Sustainable knowledge infrastructures require both robust identification and reliable preservation practices.

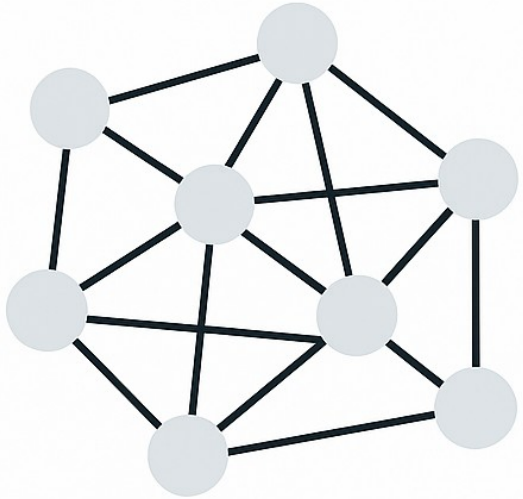
Platformisation Aspect

Centralization



Single point of failure

Decentralization



Redundant & resilient

Data Governance Vulnerabilities

- **Centralisation**.
  A few large stakeholders or state-backed entities hold disproportionate power over access, storage, and dissemination of knowledge.
- **Transparency & Provenance**.
  Without reproducibility, the stakeholder is the only one who verifies and validates the data.
- **Vulnerability to executive decisions**.
  Data preservation and access policies can change overnight, dictated by the agendas of irrational or authoritarian governments.
- **Rights management delegation problem**.
  When the data producer/owner delegates the rights management to a centralised stakeholder, he actually yields his own management rights.

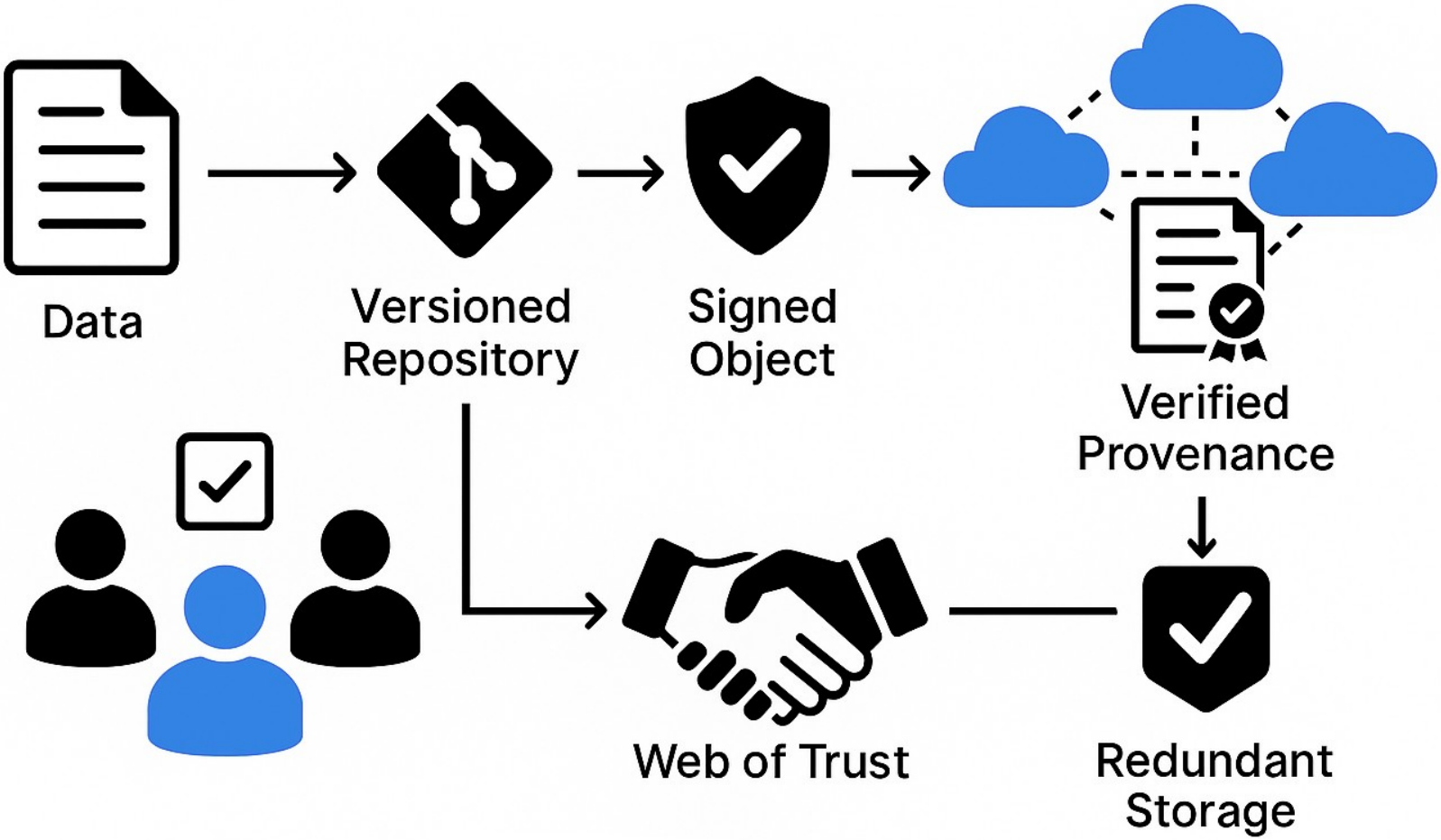
Decentralisation is Collaboration

Identification is a key to the decentralised collaborative ecosystem

- **Entity openness**.
  The data operator/provider/owner discloses himself in an open way.
- **Responsibility match**.
  In the architecturally rankless environment the computation stakeholder is not equal to identity owner.
- **Accessibility**.
  The encouraged resource sharing improves the overall redundancy of the ecosystem.
- **Interoperability** on the standardisation level is introduced by design.
- **Findability** based on the mathematically unambiguous structures.

Decentralised Identification

- By nature based on the **reproducible**, mathematical **invariants**.
- As **persistent** as the underlying network and standards are.
- **Every participant verifies all the identifiers** he interacts with at the time of the metadata retrieval.
- **Provenance and version control are more important** than the access control and interaction history.
- The decentralised workflow is **not a blockchain workflow**, but the workflow of headless retrieval + invariant reproduction (more like a BitTorrent model).
- **The signed object** is a **state of the data** but not the interaction or transaction.


Data
Versioned Repository
Signed Object
Verified Provenance
Web of Trust
Redundant Storage

Decentralised solution?

- Provide an **ecosystem** in which every participant may resolve, mint, mathematically verify and handle the PID for versioned data/metadata.
- Represent the arbitrary data/metadata as **versioned records** acting like Git repositories.
- Provide the **web of trust** over the public cryptographic entity and versioning proposal (pull request) mechanism.
- Remove the possibility to monopolise (**platformise**) the infrastructure introducing transactionless updates with short operation TTL.
- Provide the **realm mechanism** to obtain interoperability and compatibility with the existing PID systems by design.
- **Obtain the persistence** with redundant storage infrastructure (like it is done in BitTorrent).

Conclusions

- Sustainable knowledge require both persistent identification and reliable preservation practices.
- The centralised (platformised) approaches remove the sustainability in favour of social contract and public trust, leading to vulnerabilities.
- Decentralisation forms an important step in making the knowledge sustainable and compatible to FAIR model, enforcing reproducibility and redundancy by design in PIDs.
- The social trust is possible to introduce over the public reputation and programmable recognition instead of social contracts.
- Decentralisation removes the gap between provenance and storage provision responsibility, making the knowledge publicly interoperable.

Possible solution: Components

- `OpenSSL` - provides hash functions used as internal representation for the PID address space to simplify storage.
- `OpenDHT` - provides infrastructure for storing and near-realtime sharing of the realm table and realms (Git-like incremental repositories).
- `Git` - provides version control and low-level systematized storage.
- `IPFS`/`iroh` - provides internal storage with redundancy by design.

Stage: conceptual design / started TDD process for realms

Thank you for your attention!

Elettra Sincrotrone Trieste

15 Años UNPAZ
Universidad Nacional de José C. Paz

UPPSALA UNIVERSITET

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

References

1. Nosek, B. A. et al. Promoting an open research culture. Science, 348, 1422-1425(2015). DOI: 10.1126/science.aab2374
2. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). DOI 10.1038/sdata.2016.18
3. Görzig, H., Krahl, R., Matthews, B., McBirnie, A., Ounsy, M., Servan, S., Soler, N. (2023). PaN FAIR implementation framework sustainability sheet. Zenodo. DOI 10.5281/zenodo.7804713
4. Vukolov, A. (2021). Openly reproducible Persistent Identifiers (PIDs) as a factor of FAIRness in data sharing practices. EOSC Symposium 2021. Zenodo. DOI 10.5281/zenodo.4980522
5. Van Winkle, E., et al. (2025). An Overview of Decentralized Web Technologies as a Foundation for Future IPFS-Centric FDOs. Open Conference Proceedings, 5. DOI 10.52825/ocp.v5i.1054
6. Lukács, B. Education and the access to it are a public good. Following the UN Human Rights Charta and achieving the UNESCO Sustainable Development Goals. GitHub
7. Lukács, B.e, Weidener, L., Heurich, B. (2023). Opportunities and Limitations of Decentralization in Decentralized Science. DOI 10.48446/opus-14635.
8. Fehder, D. C., Murray, F., Stern, S. Intellectual property rights and the evolution of scientific journals as knowledge platforms. International Journal of Industrial Organization, Volume 36, 2014, pp. 83-94, ISSN 0167-7187, DOI 10.1016/j.ijindorg.2014.08.002.
9. Klump, J., Murphy, F., Weigel, T. and Parsons, M.A. (2017) Editorial: 20 Years of Persistent Identifiers – Applications and Future Directions. Data Science Journal, 16(0), p. 52. DOI 10.5334/dsj-2017-052

Supplementary Links

- Concepts and Definitions Draft
- OpenDHT
- OpenSSL
- Git
- iroh / IPFS
- Decentralized Persistent Identifiers: a basic model for immutable handlers
- Program document for PIDs in Photon and Neutron Society
- PID property comparison table
- Why Decentralisation is what should be done? (also on IPFS)

Distributed Identification of Knowledge.
How Can Decentralized PIDs Save Human Legacy?

Andrey Vukolov
Elettra Sincrotrone Trieste,
Trieste, Italy

Sergio Santamarina
Universidad Nacional de Jose C. Paz,
Buenos-Aires, Argentina

Jonas Söderberg
National Bioinformatics Infrastructure
Sweden, Uppsala University, Sweden

Elettra Sincrotrone Trieste

15 Años UNPAZ
Universidad Nacional de José C. Paz

UPPSALA
UNIVERSITET

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

I call the ongoing change "documedia revolution" because it is based on the intersection between the increase in documentality, i.e. the production of documents as a constitutive element of social reality, and the growth of the media, which today no longer work as one-to-many but as many-to-many.

(Ferraris, 2023)

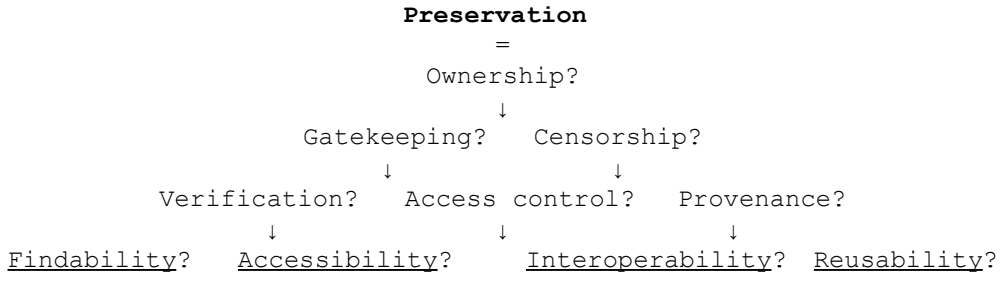
Why Preserve Knowledge?

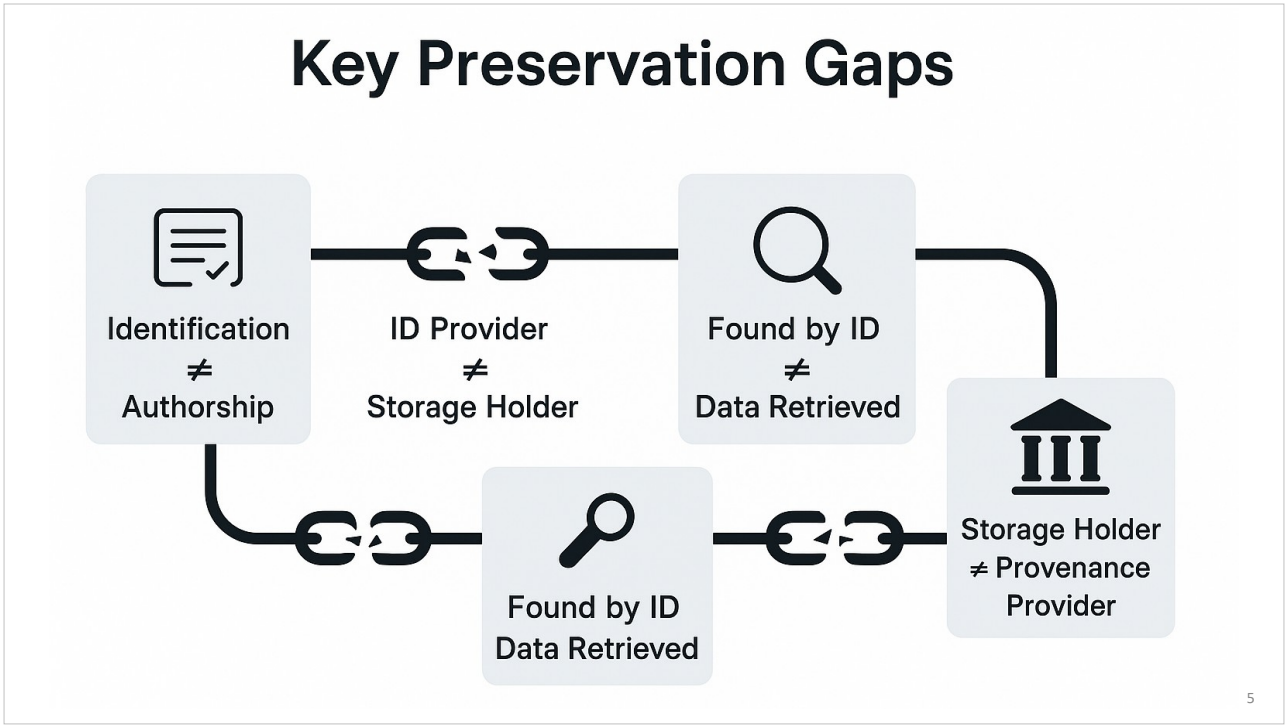
- **Knowledge is humanity’s memory**
  - It enables progress, avoids repeating mistakes, and sustains culture. When data vanishes, science regresses, trust erodes, and lives can be impacted.
- **Preservation is a collective responsibility**
  - It ensures knowledge remains findable, accessible, and reusable for future generations.

Knowledge is humanity’s memory
— It enables progress, avoids repeating mistakes, and sustains culture.
Research is fragile
— Data can be lost to time, neglect, disasters, or political agendas.
Access empowers everyone
— Preserved knowledge fuels education, innovation, and accountability.
Loss has real consequences
— When data vanishes, science regresses, trust erodes, and lives can be impacted.
Preservation is responsibility
— It ensures knowledge remains findable, accessible, and usable for future generations.

Knowledge Preservation

- As an abstraction, knowledge itself is represented by data.
- Data (and thus knowledge) is produced by research
- Preservation must happen where research happens
- Not left solely to third-party organizations.





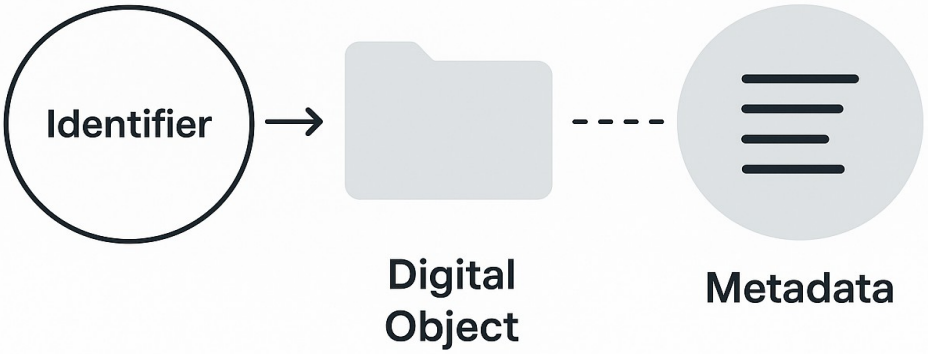
Misalignments in the knowledge preservation chain:

Identification ≠ Authorship
The entity assigning an identifier is not necessarily the creator of the work.
ID Provider ≠ Storage Holder
The service providing the identifier may not be the one storing the data.
Found by ID ≠ Data Retrieved
Having an identifier does not guarantee that the corresponding data is accessible.
Storage Holder ≠ Provenance Provider
The one holding the data might not be able to attest to its origin or authenticity.

Risks

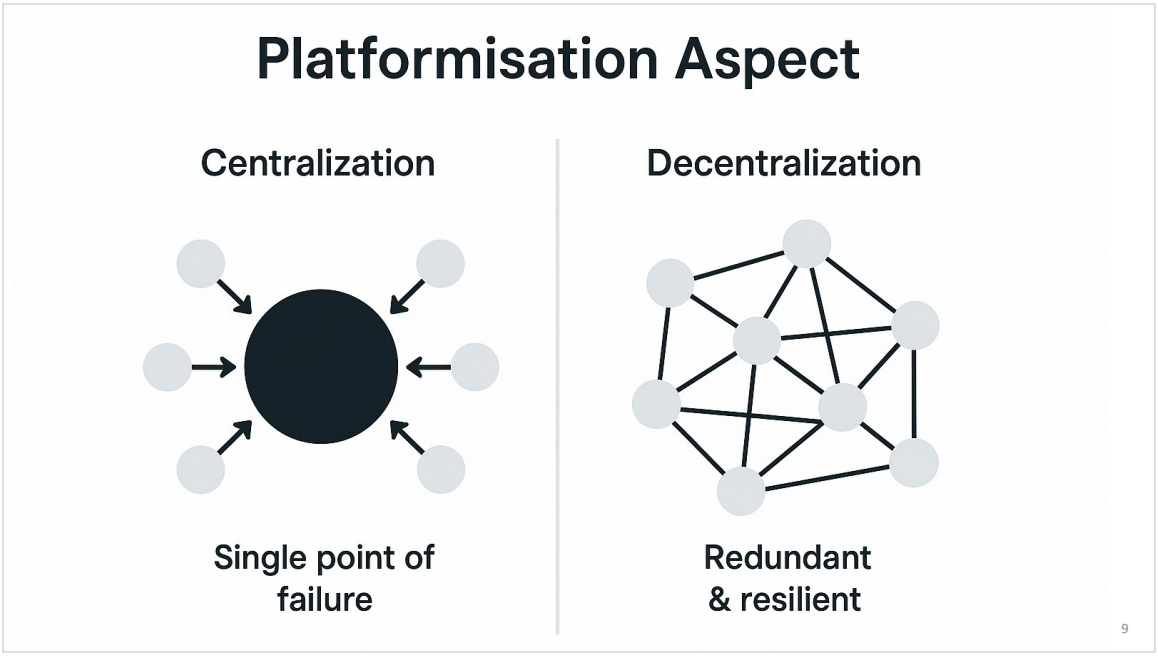
- **Willing Censorship**
  Loss of access due to political or commercial decisions
- **Social Inconvenience**
  Censorship or deletion of inconvenient knowledge
- **Hidden Players**
  Lack of transparency and accountability in stewardship
- **Neglection (willing or unwilling)**
  Lack of maintenance leading to the losses of data

Identifiers in Archiving



Identification and Preservation

- **Identification enables discovery**
  Unique identifiers (e.g., DOIs, ARKs) make knowledge findable and citable.
- **Preservation ensures longevity**
  Data and knowledge must remain accessible, authentic, intact over time.
- **Identification ≠ Preservation**
  Assigning an identifier does not guarantee that the data is actually stored or maintained.
- **Both are necessary and complementary**
  Sustainable knowledge infrastructures require both robust identification and reliable preservation practices.



Rise of Stakeholders
Critical data and research outputs are increasingly managed by large, centralised platforms.
Singular Entry Point
The platform creates the singular point of failure for the amounts of data significant for the whole tech civilization.
Gatekeepers
The stakeholder of a platform carries gatekeeping responsibility by the nature of the platform, that leads to separation of the users by design.
The data and relations between the data are products of the platform so the ownership shift becomes imminent.

Data Governance Vulnerabilities

- **Centralisation**.
  A few large stakeholders or state-backed entities hold disproportionate power over access, storage, and dissemination of knowledge.
- **Transparency & Provenance**.
  Without reproducibility, the stakeholder is the only one who verifies and validates the data.
- **Vulnerability to executive decisions**.
  Data preservation and access policies can change overnight, dictated by the agendas of irrational or authoritarian governments.
- **Rights management delegation problem**.
  When the data producer/owner delegates the rights management to a centralised stakeholder, he actually yields his own management rights.

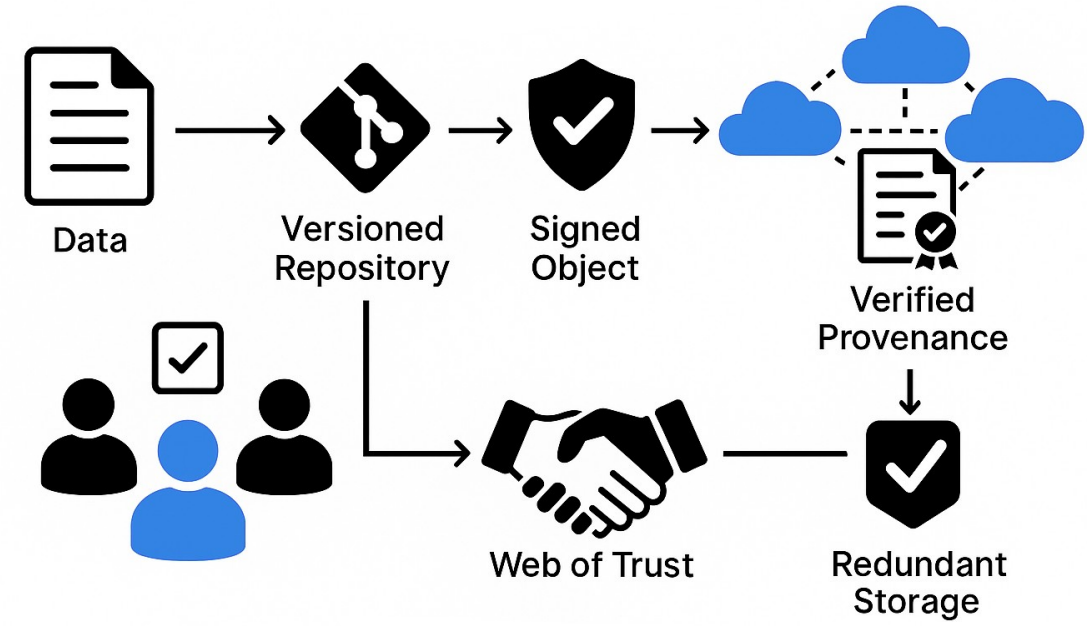
Decentralisation is Collaboration

Identification is a key to the decentralised collaborative ecosystem

- **Entity openness**.
  The data operator/provider/owner discloses himself in an open way.
- **Responsibility match**.
  In the architecturally rankless environment the computation stakeholder is not equal to identity owner.
- **Accessibility**.
  The encouraged resource sharing improves the overall redundancy of the ecosystem.
- **Interoperability** on the standardisation level is introduced by design.
- **Findability** based on the mathematically unambiguous structures.

Decentralised Identification

- By nature based on the **reproducible**, mathematical **invariants**.
- As **persistent** as the underlying network and standards are.
- **Every participant verifies all the identifiers** he interacts with at the time of the metadata retrieval.
- **Provenance and version control are more important** than the access control and interaction history.
- The decentralised workflow is **not a blockchain workflow**, but the workflow of headless retrieval + invariant reproduction (more like a BitTorrent model).
- **The signed object** is a **state of the data** but not the interaction or transaction.


Data
Versioned Repository
Signed Object
Verified Provenance
Web of Trust
Redundant Storage

Decentralised solution?

- Provide an **ecosystem** in which every participant may resolve, mint, mathematically verify and handle the PID for versioned data/metadata.
- Represent the arbitrary data/metadata as **versioned records** acting like Git repositories.
- Provide the **web of trust** over the public cryptographic entity and versioning proposal (pull request) mechanism.
- Remove the possibility to monopolise (**platformise**) the infrastructure introducing transactionless updates with short operation TTL.
- Provide the **realm mechanism** to obtain interoperability and compatibility with the existing PID systems by design.
- **Obtain the persistence** with redundant storage infrastructure (like it is done in BitTorrent).

Conclusions

- Sustainable knowledge require both persistent identification and reliable preservation practices.
- The centralised (platformised) approaches remove the sustainability in favour of social contract and public trust, leading to vulnerabilities.
- Decentralisation forms an important step in making the knowledge sustainable and compatible to FAIR model, enforcing reproducibility and redundancy by design in PIDs.
- The social trust is possible to introduce over the public reputation and programmable recognition instead of social contracts.
- Decentralisation removes the gap between provenance and storage provision responsibility, making the knowledge publicly interoperable.

Possible solution: Components

- `OpenSSL` - provides hash functions used as internal representation for the PID address space to simplify storage.
- `OpenDHT` - provides infrastructure for storing and near-realtime sharing of the realm table and realms (Git-like incremental repositories).
- `Git` - provides version control and low-level systematized storage.
- `IPFS`/`iroh` - provides internal storage with redundancy by design.

Stage: conceptual design / started TDD process for realms

Thank you for your attention!

Elettra Sincrotrone Trieste

15 Años UNPAZ
Universidad Nacional de José C. Paz

UPPSALA
UNIVERSITET

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

References

1. Nosek, B. A. et al. Promoting an open research culture. Science, 348, 1422-1425(2015). DOI: 10.1126/science.aab2374
2. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). DOI 10.1038/sdata.2016.18
3. Görzig, H., Krahl, R., Matthews, B., McBirnie, A., Ounsy, M., Servan, S., Soler, N. (2023). PaN FAIR implementation framework sustainability sheet. Zenodo. DOI 10.5281/zenodo.7804713
4. Vukolov, A. (2021). Openly reproducible Persistent Identifiers (PIDs) as a factor of FAIRness in data sharing practices. EOSC Symposium 2021. Zenodo. DOI 10.5281/zenodo.4980522
5. Van Winkle, E., et al. (2025). An Overview of Decentralized Web Technologies as a Foundation for Future IPFS-Centric FDOs. Open Conference Proceedings, 5. DOI 10.52825/ocp.v5i.1054
6. Lukács, B. Education and the access to it are a public good. Following the UN Human Rights Charta and achieving the UNESCO Sustainable Development Goals. GitHub
7. Lukács, B.e, Weidener, L., Heurich, B. (2023). Opportunities and Limitations of Decentralization in Decentralized Science. DOI 10.48446/opus-14635.
8. Fehder, D. C., Murray, F., Stern, S. Intellectual property rights and the evolution of scientific journals as knowledge platforms. International Journal of Industrial Organization, Volume 36, 2014, pp. 83-94, ISSN 0167-7187, DOI 10.1016/j.ijindorg.2014.08.002.
9. Klump, J., Murphy, F., Weigel, T. and Parsons, M.A. (2017) Editorial: 20 Years of Persistent Identifiers – Applications and Future Directions. Data Science Journal, 16(0), p. 52. DOI 10.5334/dsj-2017-052

Spare slide, to be removed later

Supplementary Links

- Concepts and Definitions Draft
- OpenDHT
- OpenSSL
- Git
- iroh / IPFS
- Decentralized Persistent Identifiers: a basic model for immutable handlers
- Program document for PIDs in Photon and Neutron Society
- PID property comparison table
- Why Decentralisation is what should be done? (also on IPFS)