PIDs & Preservation

Incorporating persistent identifiers in a preservation strategy

Helena Cousijn
DataCite
Germany
helena.cousijn@datacite.org
https://orcid.org/0000-0001-6660-6214

Ginny Hendricks
Crossref
UK
ghendricks@crossref.org
https://orcid.org/0000-0002-0353-2702

Alice Meadows
ORCID
USA
a.meadows@orcid.org
https://orcid.org/0000-0003-2161-3781

A persistent identifier (PID) provides a long-lasting reference to an entity. PIDs should be open and unique and should resolve to a digital representation of the entity. Used in this way, PIDs can serve as an important component in digital preservation strategies for academic resources.

In this panel discussion, representatives from leading PID organizations will explore roles that PIDs can play in digital preservation strategies. We invite the participants to engage in a conversation on how PID providers can work with the preservation community, and what preservation strategies they should be deploying with PID metadata.

Keywords – Persistent identifiers, metadata, infrastructure, preservation

Conference Topics – Building Capacity, Capability, and Community; The Cutting Edge: Technical Infrastructure and Implementation

I. Introduction

Crossref, DataCite, and ORCID are well-established persistent identifier providers: DOIs for research outputs (Crossref and DataCite) and ORCID identifiers (iDs) for researchers. Ensuring that these identifiers are truly persistent—even beyond the lifetime of the research output or research—is critical for continued access to the entities and/or their metadata. As such, the three organizations have committed to persistence in their organizational practices. All are open, not-for-profit, community-governed, and community-led and all are committed to a collaborative, community-driven approach to ensuring the long-term preservation of research is part of a trusted research information infrastructure.

PID organizations have a goal complementary to that of preservation organizations. Where preservation organizations focus on ensuring that the object remains digitally available over time, these organizations ensure the identifier and identifying metadata for connected objects remain discoverable over time. We work with the community to connect identifiers for the underlying person-place-object entities and make these relationships and associated metadata openly available [1]. By ensuring persistence of digital representations of entities and associated identifiers, preservation and PID organizations contribute to making research outputs FAIR [2] and enabling reuse of research outputs over time.

II. Different approaches to PID persistence

Whilst all three organizations have much in common, each one is approaching the overall question of preservation differently.

Crossref has direct agreements with many archival organizations such as Portico, CLOCKSS, and the Internet Archive, with whom its entire metadata corpus is routinely and systematically backed up, which ensures that the identifiers remain persistent.

However, the existence of a DOI does not alone ensure preservation. The more than 100 million metadata records require commitment from the asserter (e.g., publisher, funder) to maintain the records for the long-term. For this reason, we call Crossref DOIs persist-able rather than persistent. At the most basic level,

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019

when members stop publishing or go out of business (and when they tell Crossref) there is a hosted defunct DOI page that asks for information about other locations. In 2018, Crossref introduced the obligation to archive content in their membership terms [3]. In addition, Crossref also has individual agreements with some national libraries and other archiving organizations to redirect DOIs when needed, using our 'multiple resolution' service when multiple archives are involved. .

DataCite similarly focuses on the persistence of the identifier and associated entity metadata rather than the research output itself. DataCite members register DOIs with DataCite and take responsibility for maintenance and preservation of the entities for which DOIs are registered. In cases where individual organizations are no longer able to maintain the DOI record, larger organizations (e.g., national libraries) often take on this task.

While preservation and access are often treated as separate functions, DataCite treats them as complementary: preservation aimed at providing access over time, while access depends upon preservation at a point in time [4]. This is particularly critical when working with non-traditional scholarly outputs such as datasets. Providing access to datasets, related data, versions, software, etc. allow the provenance of the materials to be accessible and usable over time. To support this, DataCite systematically and regularly checks the health of its DOIs by checking for successful resolution.

In the worst-case scenario, when entities disappear from the digital domain and the connection with their DOI breaks and return a "404 error", the DOI Foundation offers a "DOI not found" form for people to report lost DOIs. The DOI Foundation also alerts the appropriate DOI Registration Agency via a daily email which is manually actioned.

Most digital preservation work has focused on digital representations of research outputs. However, the organizations where research is performed or funded, and the people that carry out research are just as important to the integrity of the research process. Digital representation of organizations (such as the forthcoming ROR IDs) [5] and persons are at a much earlier stage of community understanding and technological capability. ORCID provides a persistent identifier and landing page for researchers, open APIs, and annual public metadata files, all components of a persistence strategy. However, many of the connections to a person's identifier are to entities that have fleeting if any digital representation (e.g., student or faculty webpage, peer review activity). ORCID has therefore started conversations with preservation specialists, and are in the early stages of grappling with defining a preservation strategy.

III. Challenges remain

With the proliferation of PIDs and a growing number of low-barrier providers, PIDs are becoming commoditized. At the same time, more mandates are coming into play - whether to "get a DOI" or "get an ORCID". Without a community understanding of the needs and requirements for long-term commitment to stewarding digital content - whether about people, places, or things - PIDs cannot be useful for preservation.

IV. Panel discussion - an education-based approach

While there are clear opportunities and intentions for PID infrastructures to support the preservation of research outputs, it takes time for these services to be fully supported, and adoption is variable. Ongoing efforts need to emerge from cross-community understanding and a collective commitment to digital representations and preservation. This panel discussion will be led by Craig Van Dyck, Executive Director of CLOCKSS. To start this discussion, we will first ask the audience some important questions: 1) where should the responsibility for the preservation of the content underlying PIDs lie? 2) How should PID providers preserve their metadata? and 3) how can the different communities work together to meet preservation challenges? Based on the responses from the audience, the facilitator will lead a panel discussion on how to develop a cohesive approach to preservation, with persistent identifiers and metadata as core connecting components.

V. REFERENCES

[1] https://www.project-freya.eu/en

[2] Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management

and stewardship. https://doi.org/10.1038/sdata.2016.18
[3] https://www.crossref.org/membership/terms/
[4] Abrams, Cruse and Kunze (2009). Preservation is not a Place. https://doi.org/10.2218/ijdc.v4i1.72
[5] https://ror.org