Open Persistent Identifiers: The Building Blocks of Sustainable Scholarly Infrastructure

Maria Gould, California Digital Library

Maria Praetzellis, California Digital Library

Introduction

In May 2021, Microsoft circulated an announcement that it would be shutting down its Microsoft Academic Services (MAS) by the end of the year. The news of this decision reverberated through the open-scholarship community, raising questions and concerns among the many stakeholders who relied on the free service for tracking research activities in various contexts.[1]

At a time when research discovery is more necessary than ever, it is also becoming more complicated. The work of tracking and identifying publications and other research outputs is taking place in a context of increased technological complexity, competing motivations and priorities, and constrained resources. As exemplified by the Microsoft case, one of the fundamental challenges and risks in the scholarly infrastructure landscape is the unpredictable availability of the platforms and services we rely upon to perform this work. When these platforms and services go away, what do we have left?

Such challenges and risks might be overcome or at least mitigated if and when scholarly infrastructure is built with open components that can persist beyond their packaging. "The Principles of Open Scholarly Infrastructure" (POSI), which were initially outlined in 2015 and are seeing a revival in 2021, provide a set of guidelines for open infrastructure for research and scholarly communications.[2] Within this framework, open infrastructure is a strategy for sustainability. Using the POSI principles as a backdrop, we examine one essential ingredient of open infrastructure: persistent identifiers, or PIDs. We explore ways in which the use of openly available PIDs, and investments in

the services that support them, can enable the discovery of research outputs while promoting the sustainability of data and information.

Research libraries have an opportunity to adopt a "PID-centric" approach to tracking, sharing, and publishing research. PIDs have the potential to address pain points, increase efficiencies, and save time. Promoting the implementation of open PIDs and the metadata associated with them serves a broader goal of improving information connectivity.

While this article does not aim to offer an exhaustive discussion of the many complexities of funding, maintaining, and connecting the multiplicity of scholarly systems, nor does it promise a comprehensive survey of all persistent identifiers, we want to share our first-hand perspective on the dynamics of building and planning for open and sustainable scholarly infrastructure and we want to outline ideas and strategies to advocate specifically for prioritizing open PIDs and open metadata to ensure research sustainability.

Persistent Identifiers: Unlocking Discovery

Overview: Core Persistent Identifiers for Scholarly Communication

Persistent identifiers in the scholarly communication context serve as stable, long-lasting unique references to core components of the research enterprise. These components include but are not limited to publications and other research outputs, researchers and contributors, institutions and facilities, instruments and materials, funders, and grants and awards.

PIDs help to provide long-term unambiguous identification of and access to research (and information about research). This is useful in today's dynamic and diffuse research landscape: for example, a publication's URL is likely to change over time, multiple researchers have the same name, and researchers' affiliated institutions or funding organizations might be written in multiple ways across different

outlets. PIDs enable disambiguation and discovery by providing machine-readable data that can be used to track individual components of research and establish connections between these components at a given point and over time. They can associate researchers to publications, capture networks of research collaborators, link a set of related publications to each other, or identify the downstream products of a grant-funded project, among other uses.

PIDs can therefore help answer questions that are crucial for effective research discovery and management, such as:

- How can I find all of the research published at my institution?
- How can I identify the publications that resulted from a specific research project?
- How can I locate the data set associated with a publication?
- How can I track the downstream outcomes and impacts of a research project?
- How can I record collaborations with other research institutions?
- How can I ensure compliance with funder requirements for data sharing?

As scholarship proliferates across digital platforms and discovery systems, PIDs have become the essential building blocks of the scholarly communications infrastructure for finding, accessing, and tracking research outputs. In this context, the PIDs most commonly used include:

- PIDs for people
- PIDs for outputs
- PIDs for organizations
- PIDs for funders and grants

Within these categories, there may be more than one type of identifier. For instance, the Open Researcher and Contributor ID (ORCID) is a well-known global identifier for researchers. The ORCID registry is open and managed by a community-governed nonprofit organization.

However, there are other identifiers besides ORCID that can be used to identify researchers. ResearcherIDs and Scopus Author IDs are two examples; unlike ORCID, they are used in commercial databases (Web of Science and Scopus, respectively) and are not openly available. In the case of institutional identifiers, Research Organization Registry (ROR) IDs are freely and publicly available in the ROR registry, which provides an open data set and includes additional tools for working with institutional data, such as an open application-programming interface (API). Other identifiers for institutions also exist but they are not openly available, such as those in the Ringgold database and in Web of Science and Scopus.

It is important to understand the differences between open and non-open PIDs because they speak to real risks and inconsistencies in our current landscape. Therefore, our focus is on those PIDs that have broad adoption globally and that allow use and reuse of the metadata they contain.

The Importance of PIDs and Connected Metadata

A PID itself should not be seen as the end goal. Instead, the power of PIDs is not so much what they identify as the connections they enable. These connections, and the insights they offer, can only be fully realized through open metadata and open infrastructure.

As co-authors of a September 2020 report, *Implementing Effective Data Practices: Stakeholder Recommendations for Collaborative Research Support*,[3] we presented a set of recommendations for implementing and advocating for PIDs in research infrastructure as a way to "unlock discovery." A premise of the report is that PIDs are an essential element in building a more open research ecosystem. To fully realize this vision, systems and services that use or provide PIDs must follow open practices, particularly in terms of the open licensing of metadata:

> Organizations that sustain identifier registries are essential pieces of scholarly infrastructure, and beyond adoption and use of PIDs, these

> organizations need the support of the research community. The research community is also best served by **open licensing of metadata that enables interoperability across systems**. Libraries, IT professionals, and research offices that develop or purchase research support systems can help accelerate the adoption of PIDs by requiring that these systems be designed to integrate with identifier registries, and by advocating for open metadata and open code.[4]

When relying on PIDs to track and connect research, we need to be aware of the opportunities and limitations of the PIDs and the underlying research infrastructure that we use to do this. A proprietary identifier in a closed system is only useful to that system and its user base. This is a sustainability concern. PIDs developed **as** open infrastructure and for use **in** open infrastructure afford the greatest potential to implement efficient, cost-effective, long-lasting scholarly communication practices. An investment in open PID infrastructure is a strategy for making research—and insights about research—more accessible to all, and serves as a sort of insurance policy against the unpredictable events that can arise in a commercialized scholarly communication landscape. Frameworks like POSI can help distinguish which organizations or tools follow open principles and surface information about governance and sustainability. These considerations are significant factors if we are to prioritize investments in open infrastructure.

Below, we explore three critical areas for investment in sustainable, open PID infrastructure: (1) library publishing and institutional repositories, (2) data services, and (3) research data management. We focus on use cases featuring digital object identifiers (DOIs), for two reasons. First, we expect that many institutional stakeholders will be familiar at least in principle with DOIs, as they are commonly visible in publications, reference lists, and databases. Second, DOIs exemplify how PIDs can be enriched with metadata and how PIDs can work in concert with each other to make research more discoverable.

Library Publishing and Institutional Repositories

Institutions worldwide are sites of and incubators for transformations in research dissemination. Library publishing and institutional repository services contribute to the increase in digital scholarship artifacts that need to be managed and made available for discovery, access, and use.

These areas of growth also present some challenges and pose questions. Fundamentally, how can library publishers and repositories ensure the discoverability of their content? How can usage be tracked to understand how this content is being used and cited? How can the metadata in publishing and repository platforms generate insights and reports on research activities? How can content be networked to identify connections between researchers and the outputs they generate?

Incorporating open and PID-based infrastructure in these initiatives is one way to address these challenges and answer these questions. Institutions have several options in this regard. A starting point would be to make sure library publishers and institutional repositories are registering DOIs for their content and taking advantage of opportunities to enrich the DOI metadata with information to aid in tracking and discovery.

For library publishers, becoming a Crossref member or choosing a platform or service provider that integrates with Crossref means that DOIs can be registered for publications. For institutional repositories, becoming a member of DataCite or using a platform or service provider that integrates with DataCite means that DOIs can be registered for repository content.

However, registering DOIs is about more than just getting a DOI. While a DOI alone is useful insofar as it can provide a permanent reference to an object regardless of whether the object's location changes over time, a DOI becomes much more useful when it includes rich metadata.[5]

This metadata includes:

- Information about who published the work (names, unique IDs), their roles (creators, contributors), and where they are affiliated
- Information about the work itself (abstract, work type)
- Information about related works (data, older versions, series, dissertations, preprints)
- Information about who funded and/or sponsored the underlying research
- Information about copyright and licensing
- Information about referenced works

Enriching DOI metadata provided to Crossref and DataCite optimizes the work for greater discoverability and therefore reusability. Systems that harvest from Crossref and DataCite can index the additional metadata in the DOIs. Works can then be searched to find specific authors, or works associated with a particular institution or funder. Reference lists and related works can be analyzed to provide a fuller picture of the work in context.

Data Services

Researchers and research stakeholders today must navigate an array of policies and requirements around sharing data and following best practices for data publication—the FAIR principles.[6] While the landscape of data management and data sharing policies has been widely discussed, less frequently addressed is the role that persistent identifiers can play in navigating these requirements and adhering to best practices.

To illustrate how this can work, let's take the example of a single DOI for a data set. At a bare minimum, this DOI could function as a unique, long-lasting reference to the data set in case the location where it is stored changes over time. The existence of the identifier alone can make a significant difference in promoting the stability of this scholarly resource over the long term.

The DOI string itself is just the beginning, however. The discoverability and usability of the data set will be limited unless the DOI metadata contains additional information about the resource. This metadata is best enriched with other persistent identifiers that can optimize the DOI for discovery and usability. For example, including an ORCID ID in metadata for the data creator—as opposed to just including the creator's name as a text string—allows for the creator to be unambiguously identified with the work, and for scholarly reporting systems to better locate all of the research associated with this particular ORCID ID. In a similar vein, including a ROR ID—as opposed to a text string—for the creator's institutional affiliation allows for this institution to be linked to the work, and for systems to better track all of the research associated with the institution.

Enriching data-set metadata with identifiers also enables best practices with data-citation and data-usage tracking. When a data-set DOI is used in citations, services can capture this usage information.[7] Rich metadata included with the DOI provides more context about the research. When other PIDs are included as part of this metadata, that optimizes the metadata for machine-readability and for more efficient and comprehensive aggregation and reporting.

Institutions can pursue the following concrete steps to maximize the potential of identifiers in data publishing.

- If your institution hosts one or more data repositories, make sure the repository assigns DOIs to the data, and guide researchers and data managers to supply rich metadata when they register the DOIs.
- If you do not host a repository, make sure researchers are guided to submit their data to repositories that do follow best practices when it comes to DOIs.
- Researchers should be encouraged to obtain ORCID IDs so they can provide these identifiers with their data publications.
- Researchers preparing and publishing manuscripts should include data citations in their manuscripts.

Research Data Management

Data-management plans (DMPs) contain a wealth of information about research projects, including, amongst other things, project plans for access, preservation, and storage. Historically, DMPs have been two-page narrative documents that outline proposed data practices during a research project and detail where investigators will deposit research outputs upon project completion. Over the past few years, there has been a concerted push towards creating machine-actionable DMPs (maDMPs).[8] These next-generation DMPs are designed to move past the static narrative format and facilitate the creation of a living document that can guide research by integrating data-management activities with related systems and workflows in the research life cycle. Demonstrating their support for this work, the US National Science Foundation (NSF) recently recommended that researchers utilize PIDs for their data outputs and generate DMPs that allow for automated information exchange (maDMPs).[9]

PIDifying the DMP

Utilizing identifiers within DMPs allows information within a DMP to be shared across stakeholders, linking metadata, repositories, and institutions, and allowing for notifications and verification, with reporting taking place in real time. A vital goal of this system is to reduce the burden on researchers by generating automated updates to a plan and facilitating seamless integration with systems and groups that support research. Networked DMPs are a vehicle for reporting on the intentions and outcomes of a research project that enable information exchange across relevant parties and systems. They contain an inventory of crucial information about a project and its outputs (not just data). With a change history, stakeholders can query for updated details on the project over its lifetime.

The recent development of a new PID for DMPs, the DMP-ID, was a fundamentally important step toward creating Networked DMPs.[10] With the development of this new PID, which is built on DOI infrastructure, we can expose connections between the rich metadata within a DMP and related works such as project outputs, individuals, affiliations, and publications.[11]

Beyond the DMP-ID

Simply receiving a DMP-ID or creating a machine-readable DMP does not realize the true potential of the Networked DMP. Connections between DMPs and their eventual outputs are made possible through the linking of open identifiers, which form an interconnected web of research components in the form of a graph.[12] In the same way that an ORCID record will be empty if researchers do not provide their ORCID IDs when publishing works (and if publishing systems do not collect this information), a DMP-ID needs to be utilized and recorded to build on the networking capabilities of PIDs. Capturing these assertions on the DMP-ID enables the tracking of data-management activities as they occur during a grant project. Again, to facilitate these connections, we need both rich metadata records that include related identifiers and to build systems that enable seamless ways for researchers to include identifiers such as the DMP-ID.[13]

Use Case: FAIR Island Project

The FAIR Island Project, in which our organization is a lead collaborator, is an attempt to showcase how best to maximize the information-rich potential of the Networked DMP. FAIR Island addresses the current challenge of discovering and accessing research connected to field stations. Administrators generally do not have precise methods for tracking the research outcomes resulting from work conducted at their facilities. The FAIR Island Project utilizes a working field station as a controlled environment to test the implementation of optimal FAIR data policies and workflows built around the Networked DMP that address discovery and access to

research outputs.[14] The project builds interoperability between pieces of critical research infrastructure—DMPs, research practice, DOIs, and publications—to facilitate the advancement and adoption of open science. Through the Networked DMP, the project will promote the quantification of productivity of field stations, which has proven difficult despite qualitative assessments of the immense value of these centers of research.

How Libraries Can Support the Networked DMP

While the work to develop a Networked DMP is ongoing, libraries can now promote the adoption of this new best practice in data management by encouraging researchers at their universities to get DMP IDs. Currently, DMP-IDs can be generated via the DMPTool or Zenodo, or through DataCite member services.[15] As we work to build a more connected ecosystem, there will be increasing ways that researchers will be able to utilize their DMP-ID and cite this ID when publishing outputs related to their project.

Conclusion: The Path to Unlocking Discovery

This article has outlined specific opportunities that research libraries can pursue to incorporate or advocate incorporating persistent identifiers into workflows, infrastructures, and policies. Extending the premise that PIDs can "unlock discovery," we have discussed ways to choose and leverage identifiers to best achieve this goal. We emphasize that PIDs alone are not the solution, but rather that strategies and policies regarding PIDs should focus on what metadata is used with PIDs, and how PIDs are connected to each other. A core principle in this vein is one of openness—the openness of the identifiers and metadata themselves, as well as of the infrastructure in which they are embedded. Scholarly infrastructure should be open from the inside out, and PIDified from beginning to end. With open metadata and open infrastructure, we can build connections and support the long-term stability and usability of scholarship to promote open knowledge practices, save time and resources, and develop more meaningful insights about research.

In this framework, we claim that investing in PIDs and open PID infrastructure should be recognized and adopted as a core sustainability strategy that can insulate research stakeholders from an unpredictable landscape in which scholarly communication services come and go. Furthermore, openly available PIDs containing rich metadata give tool providers and builders a connected ecosystem to work from and offer the library community the flexibility and assurance that the information contained within these systems is not dependent on a single provider or platform.

Endnotes

1. Shortly after Microsoft's announcement, OurResearch—the organization behind such tools as Unpaywall and Unsub—announced that they were building a similar service: "We're Building a Replacement for Microsoft Academic Graph," *OurResearch Blog*, May 8, 2021, https://blog.ourresearch.org/were-building-a-replacement-for-microsoft-academic-graph/. For more on the shifting landscape of scholarly discovery services in the wake of the Microsoft announcement, see: Dalmeet Singh Chawla, "Microsoft Academic Graph Is Being Discontinued. What's Next?," *Nature Index*, June 15, 2021, https://www.natureindex.com/news-blog/microsoft-academic-graph-discontinued-whats-next.

2. Geoffrey Bilder, Jennifer Lin, and Cameron Neylon, "The Principles of Open Scholarly Infrastructure," 2020, https://doi.org/10.24343/C34W2H.

3. John Chodacki et al., *Implementing Effective Data Practices: Stakeholder Recommendations for Collaborative Research Support* (Washington, DC: Association of Research Libraries, September 2020), https://doi.org/10.29242/report.effectivedatapractices2020. An additional related work also explores concrete steps libraries can take to accelerate the sharing of research data: Tobin L. Smith et al.,

Guide to Accelerate Public Access to Research Data (Washington, DC: Association of American Universities and Association of Public and Land-grant Universities, 2021), https://doi.org/10.31219/osf.io/tjybn.

4. Chodacki et al., 5, emphasis added.

5. For a good example of the power of rich metadata see: Kristian Garza, "Are You There, Metadata? It's Me, the Bibliometrician," *DataCite Blog*, January 25, 2021, https://doi.org/10.5438/j4xv-y945.

6. Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (March 2016): 160018, https://doi.org/10.1038/sdata.2016.18.

7. Daniella Lowenberg et al., Open Data Metrics: Lighting the Fire (version 1), computer software, Zenodo, 2019, http://doi.org/10.5281/zenodo.3525349.

8. For an excellent overview of maDMPs see: Tomasz Miksa et al., "Ten Principles for Machine-Actionable Data Management Plans," *PLoS Computational Biology* 15, no. 3 (2019): e1006750, https://doi.org/10.1371/journal.pcbi.1006750.

9. Joanne S. Tornow et al., "Dear Colleague Letter: Effective Practices for Data," NSF 19-069, National Science Foundation, May 20, 2019, https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp.

10. Kristian Garza and Matt Buys, "A Brave New PID: DMP-IDs," *DataCite Blog*, April 7, 2021, https://doi.org/10.5438/j22a-5d79.

11. For more, see: "Machine Actionable Data Management Plan Connections," Jupyter nbviewer, accessed November 24, 2021, https://nbviewer.org/github/datacite/pidgraph-notebooks-python/blob/master/dmp/user-story-single-dmp-connections.ipynb.

12. Helena Cousijn et al., "Connected Research: The Potential of the PID Graph," *Patterns* 2, no. 1 (January 2021), https://doi.org/10.1016/j.patter.2020.100180.

13. For more information on linking DMP-IDs to other resources, see: "Link DMP IDs to Other Resources," DataCite, accessed November 24, 2021, https://support.datacite.org/docs/link-dmp-ids-to-other-resources.

14. Maria Praetzellis, "Introducing the FAIR Island Project," February 10, 2020, University of California Curation Center (UC3), https://uc3.cdlib.org/2020/02/10/introducing-fair-island/.

15. For more information on creating a DMP-ID, see: "DataCite DMP IDs," DataCite, accessed November 24, 2021, https://support.datacite.org/docs/datacite-dmp-ids.

To cite this article: Gould, Maria, and Maria Praetzellis. "Open Persistent Identifiers: The Building Blocks of Sustainable Scholarly Infrastructure." *Research Library Issues*, no. 302 (2021): 5–18. https://doi.org/10.29242/rli.302.2.