Persistent identifiers: the building blocks of the research information infrastructure

Persistent identifiers (PIDs) – for people (researchers), places (their organizations) and things (their research outputs and other contributions) – are foundational elements in the overall research information infrastructure. They enable these entities to be uniquely identified and connected, to create reliable links between them. In this paper we describe what PIDs are and how they work. We demonstrate how, if widely adopted, the connections they enable will result in improved access to information, opportunities for collaboration, reduced administrative overhead and, ultimately, increased trust in scholarship and research. To ensure they are fit for purpose, we propose that PID metadata should meet FAIR principles, and that the source of information they contain should be clear and transparent. We also recognize existing usage of persistent identifiers, and invite community support for wider adoption of PIDs in future.

Keywords
Research infrastructure; persistent identifiers; ORCID; Crossref; DataCite; metadata

ALICE MEADOWS
Director of Communications
ORCID

LAUREL L HAAK
Executive Director
ORCID

JOSH BROWN
Director of Partnerships
ORCID

The ultimate objective of anyone working in scholarly communications is to support research and researchers. We do this to ensure that research results can be shared, validated, (re)used and built upon to support creativity, thoughtfulness, innovation and decision-making in ways that we hope will ultimately benefit society.

Supporting research includes supporting the research information infrastructure: the tools and services that researchers use which enable them to spend more time doing research and less time managing it – as well as the virtual building blocks on which those tools and services depend, such as metadata, standards and, the topic of this article, persistent identifiers (PIDs).

The Merriam-Webster dictionary defines infrastructure as 'the underlying foundation or basic framework (as of a system or organization)'. Similarly, the research infrastructure is the underlying framework for the research process. It provides, or supports the development of, the products and services – physical or virtual – necessary for undertaking research, from ideation to publication and beyond.

Whether infrastructures are physical (laboratories, research equipment, etc.) or virtual (such as digital platforms and services), and whether they are intended for researchers or for the whole of society, it is critical that they work – that they do what they are intended to do in

the service of the community that they serve. Infrastructure must be both trustworthy and trusted. Although originally developed as a set of community-defined principles and practices for research data, the FAIR principles,[1] also offer a worthwhile definition of a trustworthy research information infrastructure – one in which information is findable, accessible, interoperable, and reusable.

'Infrastructure must be both trustworthy and trusted'

PIDs are an increasingly important component of the research information infrastructure. They enable clear, reliable and unambiguous connections between people, places and things. In this article we will describe what PIDs are and how they work, as well as demonstrating how, if widely adopted, the connections they enable will result in improved access to information, opportunities for collaboration, reduced administrative overhead and, ultimately, increased trust in scholarship and research.

What are persistent identifiers?

'clear, reliable and unambiguous connections between people, places and things'

According to Wikipedia, a persistent identifier is 'a long-lasting reference to a document, file, web page, or other object … Typically, such an identifier is not only persistent but actionable: you can plug it into a web browser and be taken to the identified source'.

In research infrastructure, PIDs can be broadly grouped into three main types:

- identifiers for researchers, such as ORCID iDs, ResearcherIDs and Scopus IDs
- identifiers for organizations, including GRID (Global Research Identifier Database), Ringgold IDs, ISNIs (International Standard Name Identifiers), LEIs (legal entity identifiers) and the identifiers that will be provided by the recently announced Research Organization Registry[2]
- identifiers for research objects and outputs, for example, DOIs (digital object identifiers), Archival Resource Key identifiers (ARKs), handles and IGSNs (International Geo Sample Number).

PIDs may be open, i.e. fully interoperable in any system (like those provided by Crossref, DataCite, ORCID and others) or proprietary, i.e. for use within a single organization (such as Clarivate's ResearcherID or Elsevier's Scopus ID). In their 2015 paper, Bilder et al. make the case for an open research infrastructure, arguing that, 'Everything we have gained by opening content and data will be under threat if we allow the enclosure of scholarly infrastructures. We propose a set of principles by which Open Infrastructures to support the research community could be run and sustained',[3] and some commercial organizations now make parts or all of their products and services open. For example, Digital Science's GRID database is now available under a CC0 license.[4]

In addition, PIDs may be local to an individual organization (e.g. identifiers in an internal human resources system), national (e.g. the DAI – Digital Author Identifier, used in the Netherlands), or global (all the examples in the paragraph above).

While there is no requirement for a PID to be open and/or universal,[5] certain qualities of PIDs make them more useful for making the trusted connections that enable researchers to find and use the information they need for their work. The first desirable quality is resolvability, i.e. PIDs that are URLs, or can be transformed into URLs, which link directly either to a digital document or to a human-readable landing web page. Ideally, they also provide machine-readable metadata. For example, a PID for requests for comments (RFC) – a type of publication used by the technology community[6] – could be used to generate links to a web page that contains the RFC, by converting the identifier 'rfc6750' into '**https://tools.ietf.org/html/rfc6750**'. So, while it is certainly desirable for PIDs to be resolvable, this does not address the issues around what Martin Klein and Herbert van de Sompel have termed reference rot: 'a combination of two problems common for URI (Uniform Resource Identifier) references: link rot and content drift',[7] which remain a major community challenge.

A second, highly desirable quality is for PIDs to have FAIR metadata. As well as significantly reducing the risk of reference rot, this enables the discovery of open, interoperable, well-defined (FAIR) metadata containing provenance information in a predictable manner – and the PIDs themselves are also open. DOIs are a good example. They are governed by the non-profit International DOI Foundation. The information in the metadata can be used to help establish the provenance of the item, and the metadata itself is available under a CC0 license, so it is open to everyone.

How do persistent identifiers work?

PIDs act as both unique identifiers and, critically, as connectors. By unambiguously identifying and connecting an individual researcher with their research organizations, professional activities and other contributions, we can be confident that we understand – and can assert – the relationships between each of them. And, by doing so using resolvable PIDs that incorporate FAIR metadata, we also make researchers, their affiliations and their contributions more easily discoverable. This helps researchers by enabling them to reliably find the information they need for their work; and it also allows their contributions to be recognized.

'PIDs act as both unique identifiers and, critically, as connectors'

DOIs for publications are a great example of how this works in practice. Citations have long been a central element in scholarly publishing, made all the more important by the increasing use of bibliometrics in the evaluation process. Crossref was founded in 2001, primarily because a group of publishers could see that, with journals increasingly moving online, there was an opportunity – and a need – for online citation linking.[8] But citation linking would only be possible if the article being cited could be uniquely identified and connected with the article citing it. To be citable in the digital world, every article (at the time – this has now been extended to include pretty much every type of publication from books and book chapters to data sets, videos and more) would need a PID. Crossref's PID of choice was the relatively new DOI; at the time of writing they had registered well over 100 million of them.[9]

Benefits of PID adoption

The vision statement on ORCID's website is a good example of what a PID-enabled world would look like: 'a world where all who participate in research, scholarship, and innovation are uniquely identified and connected with their affiliations and contributions across disciplines, borders, and time'. There are a number of benefits to this world, for researchers, their organizations and the wider community alike.

For researchers

For researchers, widespread PID adoption will enable significant time savings. *Nature*'s 2016 salary survey showed that, on average, researchers spend over a quarter of their time on administrative tasks, such as data storage (5%), grant application (10%) and 'other' (11%).[10] Instead of spending time on frustrating administrative tasks like online form-filling during grant application or manuscript submission, researchers could simply grant access to their ORCID record and enable data-sharing with the various research information systems with which they interact. A good example of this in practice is in Portugal, where the national funding body, Fundação para a Ciência e a Tecnologia, has integrated ORCID into the PTCRISync system following their evaluation of the time savings that could be made through reduced administrative overhead.[11]

'widespread PID adoption will enable significant time savings'

As well as the time savings to be gained, there is also a greatly reduced risk of errors, through system-to-system connections between ORCID iDs and other identifiers on ORCID records. A great example of this is Crossref's auto-update functionality,[12] which at the time of writing has enabled close to 1.5 million DOIs for works to be pushed directly into ORCID records (with the user's permission), using information provided by the

publisher in the metadata submitted to Crossref. These data are then available for systems like PTCRISync; all researchers need to do is use their ORCID iD when submitting their manuscript to a publisher.

Expanded PID adoption would also enable improved recognition for the many and varied contributions that researchers make. DOIs for publications are already at the heart of the current citation-based evaluation system, but DOIs are also increasingly being assigned to open peer reviews, for example, by Crossref.[13] Morressier, a platform for early-stage research, assigns Crossref DOIs to posters and presentations to enable researchers to get credit for their work ahead of pre-publication.[14]

'all researchers need to do is use their ORCID iD when submitting their manuscript'

For organizations and the wider community

As noted, the benefits of increased PID adoption also extend to research organizations and the wider community. For example, funding bodies that need to identify and monitor research outputs from their grants can more quickly, easily and reliably do so with the help of identifiers for their organization, their grants and their researchers, all of which can be connected to metadata in the DOI for the publication. Similarly, universities and other research institutions can rely on PIDs that connect researchers with their affiliations (former positions, memberships, services), grants and works when evaluating promotion and tenure applications. And PIDs can also enable publishers to more readily comply with funder open access requirements.

Next steps

To make this vision a reality requires community commitment. Journal publishers have already made significant steps through their almost universal adoption of DOIs, their increasing use of organization identifiers for funding information and their collection of ORCID iDs for journal authors during manuscript submission and, increasingly, for peer reviewers. However, there is still work to be done!

In particular, to further increase trust in research and scholarship, it is important that, as well as using PIDs and making connections between them, there is transparent source information: who made these connections and what is their relationship to the researcher? Making the source of this information transparent enables us to make informed decisions about the value of the information for the task in hand. For example, when considering whether to invite someone to serve as a reviewer or an editorial board member, it could be very helpful to know whether information about their employment affiliation (with organization identifier, start date and job title) has been added to their ORCID record by the researcher themselves or by their employer, or both. Similarly, knowing whether a publication was added directly by the publisher or via a third party, such as Crossref, Scopus or Europe PubMed Central, could be valuable information for any organization wanting to reuse that data.

Work on improving provenance information for PIDs in metadata is, therefore, an important focus both for our organization, ORCID,[15] and for the wider community, via initiatives such as Metadata 2020.[16] Work on expanding adoption of PIDs more generally is also increasingly being pushed at the national level as seen, for example, in the recently launched French national plan for open science, which states that 'France will help define and regulate the building blocks of the open science ecosystem, such as Crossref and DataCite for DOIs and ORCID for researcher identifiers'.[17]

'improving provenance information for PIDs in metadata'

Continued community support for – and participation in – these efforts is critical if we are to succeed in making PIDs the trusted building blocks we all need them to be in order to support the global research endeavor.

Abbreviations and Acronyms
A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the 'full list of industry A&As' link: http://www.uksg.org/publications#aa

Competing interests
The authors have declared no competing interests.

References

1. Mark Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data* 3: 160018 (2016); DOI: https://doi.org/10.1038/sdata.2016.18 (accessed 30 January 2019).
2. "The ROR of the Crowd," ROR Community website: https://www.ror.community/blog/2018-12-02-the-ror-of-the-crowd/ (accessed 30 January 2019).
3. Geoffrey Bilder, Jennifer Lin and Cameron Neylon, "Principles for Open Scholarly Infrastructure-v1," *figshare* (2015); DOI: https://doi.org/10.6084/m9.figshare.1314859 (accessed 30 January 2019).
4. "Digital Science Releases Global Research Identifier Database (GRID) Under CC0 License," Digital Science website: https://www.digital-science.com/press-releases/digital-science-releases-global-research-identifier-database-grid-cc0-license/ (accessed 30 January 2019).
5. Tom Demeranville, "Building a Robust Research Infrastructure One PID at a Time," *ORCID blog*, August 7, 2018: https://orcid.org/blog/2018/08/08/building-robust-research-infrastructure-one-pid-time (accessed 30 January 2019).
6. "Request for Comments" page on Wikipedia: https://en.wikipedia.org/wiki/Request_for_Comments (accessed 30 January 2019).
7. Martin Klein and Herbert van de Sompel, "Reference rot in web-based scholarly communications," *LSE Impact Blog*, January 21, 2015: http://blogs.lse.ac.uk/impactofsocialsciences/2015/02/05/reference-rot-in-web-based-scholarly-communication (accessed 30 January 2019).
8. "History" page on Crossref website: https://www.crossref.org/history/ (accessed 30 January 2019).
9. "Dashboard" page on Crossref website: https://www.crossref.org/dashboard/ (accessed 30 January 2019).
10. Brendan Maher and Miguel Sureda Anfres, "Young scientists under pressure: what the data show," *Nature – News*, 538, 444 (2016); DOI: https://doi.org/10.1038/538444a (accessed 30 January 2019).
11. António Luís Lopes, "Integrating a local CRIS with the PT-CRIS synchronisation ecosystem," *EuroCRIS* (2018): http://hdl.handle.net/11366/650 (accessed 30 January 2019).
12. "ORCID auto-update" page on Crossref website: https://www.crossref.org/community/orcid/ (accessed 30 January 2019).
13. "Peer reviews" page on Crossref website: https://www.crossref.org/services/content-registration/peer-reviews/ (accessed 30 January 2019).
14. Alice Meadows, "Sharing and Recognizing Early Stage Research: An Interview with Sami Benchekroun of Morressier," *The Scholarly Kitchen* (blog), December 19, 2018: https://scholarlykitchen.sspnet.org/tag/morressier/ (accessed 30 January 2019).
15. Robert Peters, "Assertion Assurance Pathways: What Are They and Why Do They Matter?," *ORCID blog*, June 13, 2018: https://orcid.org/blog/2018/06/13/assertion-assurance-pathways-what-are-they-and-why-do-they-matter (accessed 30 January 2019).
16. "About" page on Metadata 2020 website: http://www.metadata2020.org/about/ (accessed 30 January 2019).
17. National Plan for Open Science, 2018: https://libereurope.eu/wp-content/uploads/2018/07/SO_A4_2018_05-EN_print.pdf (accessed 30 January 2019).

Corresponding author
Alice Meadows
Director of Communications
ORCID, US
E-mail: a.meadows@orcid.org

ORCID ID: http://orcid.org/0000-0003-2161-3781

Co-authors
Laurel L Haak
ORCID ID: http://orcid.org/0000-0001-5109-3700

Josh Brown
ORCID ID: http://orcid.org/0000-0002-8689-4935

To cite this article:
Meadows A, Haak L L and Brown J, Persistent identifiers: the building blocks of the research information infrastructure, *Insights*, 2019, 32: 9, 1–6; DOI: https://doi.org/10.1629/uksg.457

Submitted on 18 December 2018 Accepted on 21 January 2019 Published on 13 March 2019

Published by UKSG in association with Ubiquity Press.