
EDITORIAL CONTENT

Editorial: 20 Years of Persistent Identifiers – Applications and Future Directions

Jens Klump¹, Fiona Murphy², Tobias Weigel³ and Mark Parsons⁴

¹ Mineral Resources, CSIRO, Kensington, Western Australia, AU

² Department of Meteorology, University of Reading, Reading, Berkshire, UK

³ Data Management, German Climate Computing Center, Hamburg, DE

⁴ Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, New York, US

Corresponding author: Jens Klump (jens.klump@csiro.au)

Persistent identifiers (PID) have existed for more than 20 years and have become well established as a means for identifying literature and data on the web. They were invented to address the problem of disappearing internet links, also known as “link rot”, which was seen as undermining the emerging digital record of science. A number of PID systems have since been developed, and their utility for the management of the scientific record has been reviewed. Since the initial launch of the Handle System, we have seen many more uses for persistent identification, besides literature and data. A session at the European Geosciences Union General Assembly 2016 was dedicated to ‘20 years of persistent identifiers – where do we go next?’ A number of contributions from this session have since been developed into full papers that form this Data Science Journal Special Collection, with additional solicited contributions. Together, these papers give us an overview of the use of persistent identifiers in research information infrastructures and possible future directions.

Keywords: persistent identifiers; data publication; linked data; web technology; interoperability

Introduction

Persistent identifiers (PID) have existed for more than 20 years and have become well established as a means for identifying literature and data on the web. They were invented to address the problem of disappearing internet links, also known as “link rot”, which was seen as undermining the emerging digital record of science (Dellavalle et al, 2003; Harter and Kim, 1996). While some argued that careful management of web servers would prevent this problem (Berners-Lee, 1998), others recognised that a distinction should be made between an object’s identity and its location on the internet (e.g. Arms, 1995; Lynch, 1997). A number of PID systems have since been developed, and their utility for the management of the scientific record has been reviewed by Duerr et al (2011) and again by Klump and Huber (2017).

Since the initial launch of the Handle System (Kahn and Wilensky, 1995) we have seen many more uses for persistent identification, besides literature and data. PIDs are now used to identify gene sequences, proteins, specimens in scientific collections, individual researchers, research grants, and more applications are emerging. Some PID systems have become established parts of the science information infrastructure, in other areas we are seeing work in progress, and new use cases being proposed.

Persistent identifiers can also be used as anchor points in the semantic web and thus aid discovery of research resources. When semantic reasoning is applied not only to concepts in linked data but is applied to concrete identifiable objects, persistent identifiers are crucial for unambiguous identification of these objects (Güntsch et al, 2017). Furthermore, persistent identifier metadata should hold information about how they relate to other objects. This “cross-linking” has been a feature of some persistent identifier systems for more than a decade, but only now is cross-linking between identifiable objects being further developed (Mayernik et al, 2016).

A session at the European Geosciences Union General Assembly 2016 was dedicated to ‘20 years of persistent identifiers – where do we go next?’ A number of contributions from this session have since been developed into full papers that form this Data Science Journal Special Collection, with additional solicited contributions. Together, these papers give us an overview of the use of persistent identifiers in research information infrastructures and possible future directions.

Overview of the Special Collection

The special collection opens with a memorial for Norman Paskin, founding and long-term director of the International DOI Foundation (Erickson and Lannom, 2017). Norman was an influential advocate of PIDs and his contributions shaped PID usage and policies from the early days on. Without his dedication and drive many of the developments and discussions reported in this collection would have never been conceived.

The special collection itself is structured into three sections: Fundamental aspects of PIDs and PID systems, PID practice in general, and particular applications in the Earth Sciences.

The discussion of fundamental aspects of PID systems starts with a comprehensive, historical overview by Klump and Huber (2017), which reflects on 20 years of persistent identifier systems and illustrates how, despite the small community, a number of competing PID systems were built and how some of these systems developed over the years. They argue that often organizational criteria were the key factors for demise or success of PID systems. The article concludes that the PID community has to accept that good business models are essential to ensure sustainability and, as history has shown, these do not come for free.

As one pathway for future PID system evolution and following the principles established in the previous articles, Golodoniuc et al (2017) explore the range of known distributed system approaches, resulting in a comprehensive discussion and showcase of applying distributed hash tables and peer-to-peer protocols for a novel, decentralized PID system design.

Diving further into future aspects of PID systems, Car et al (2017) systematically dissect the topic of persistency in view of ever-changing technology. They propose a set of four principles for PID systems architects to follow to mitigate the conflict between persistency and adaptation. In addition, they introduce a Platform Independent Model for PID metadata and discuss its implementation. The article concludes with an outlook on activities at a current W3C group and puts forward an urgent warning that the principles must be obeyed to prevent a possible PID sustainability failure visible on the horizon.

The fundamental section concludes with a stimulating essay by Kunze et al (2017), who challenge the established vocabulary of the PID community and advocate new terminology, methodology and an overall call for action to be more precise when communicating about persistency and the policies depending on it.

The practical showcases section starts with a comprehensive overview of activities at the German National Library for Science and Technology (TIB Hannover) by Kraft et al (2017). With the TIB being a known long-term player in the PID community and hosting the DataCite office, the article reports on typical questions arising out of a DOI service’s daily business, illustrates how DOI services are embedded within other relevant TIB activities, and also gives insight into future activities concerning ORCID.

Dappert et al (2017) report on activities of the Technical and Human Infrastructure for Open Research (THOR) project that aim to solve remaining PID interoperability challenges. The article examines how to enable applications beyond simple identification and enhance support across the research lifecycle. Based on stakeholder scenarios, the authors highlight how interconnected services can enable good research practices and identify remaining limitations.

Wang et al (2017) describe practice for big data management at the Australian National Computing Infrastructure (NCI) in their article, which offers insight into the embedding of PID approaches into a national service. Identification covers data, documents, and vocabulary terms, using an approach based on wrapping existing URL-based services.

The third section on particular applications in the Earth sciences sets out with a general review of DOI usage for data across Earth science sub-disciplines by Goldstein et al (2017). While the study only covers US-based institutions, it reveals fine differences in DOI assignment policies, and the authors conclude with a discussion of what they identify as the remaining challenges for dataset DOIs.

Aquino et al (2017) illustrate exemplar practices at the National Center for Atmospheric Research on how DOIs can be used to identify Earth observation datasets and physical objects such as research platforms and instruments. The article discusses versioning, granularity, and the detailed steps of the implemented DOI workflow.

Wanchoo et al (2017) also focus on Earth observation, but at a different institution, NASA. They illustrate in detail how the operational DOI processing workflow was automated, significantly increasing efficiency,

how it is organized across multiple data centres, and how DOIs are managed between assignment and data publication.

Finally, Conze et al (2017) report about their experiences with assigning identifiers for geological samples (IGSNs) as part of a continental drilling expedition, covering aspects of identifier name construction, metadata allocation, and landing page presentation.

The articles in this special collection provide a broad overview of current PID practice, within specific communities and organizations, and draw a comprehensive arch from the history of PIDs and their supporting systems to the remaining open challenges, including questions on the terminological foundations such as persistency and the user values, towards future directions such as distributed PID system evolution, terminology, organizational concerns and uptake strategies.

The widening value proposition of identifiers

The research workflow spans a multitude of activities from data acquisition or generation over processing and analysis towards publication. Initially, PIDs emerged from needs at the publishing stage, the end of this workflow, since their primary value at that time was perceived as improving the accuracy of scientific credit giving and fostering the wider adoption of citation practices, and particularly, data citation.

However, PIDs have also gained recognized value at earlier scientific workflow stages. Their added value here does not come primarily from direct use by researchers, but more as an enabling layer for upper user-oriented services that improves their performance or capabilities. The most prominent role PIDs take is as an enabling layer for managing digital objects of various forms, independent or prior to formal publication (see, for instance the European Commission's High-Level Expert Group on Scientific Data "Riding the Wave" report by Wood et al, 2010) and led to the establishment of a fundamental PID support service in the frame of the EUDAT collaborative data infrastructure. Further roles in which PIDs have been recognized as enabling technology are regularly discussed within the Research Data Alliance and expressed for instance in the recommendations on PID Information Types (Weigel et al, 2016), Data Type Registries (Broeder and Lannom, 2014) and Data Citation (Rauber et al, 2016) or ongoing discussions in the context of research infrastructure building across disciplines (Genova et al, 2017) and automated data processing.

The benefits to all parties of harnessing this functionality are well understood by a certain proportion of the research community. PIDs are domain agnostic and enable disambiguation so improve levels of reproducibility. At the same time, they support opportunities for automation (so speeding up and increasing the accuracy of data input and linking processes), provide opportunities for metrics – around usage, re-use and other sorts of relationships between research objects. However, the take-up and understanding of these opportunities have not been well disseminated amongst the vast majority of researchers. The use of DOIs for journal articles is now a standard practice, but this has been achieved using the single research object that is universally accepted as a critical scientific output connected with an industry-wide initiative (Crossref) that inculcates standards and provides an infrastructure and back-end system that researchers do not need to interact with. In other words, the outlook for most other PID systems within research communities is still very challenging.

One of the most successful academic PID systems apart from the DOI is the Open Researcher and Contributor ID (ORCID). It is routinely used by academic-facing platforms as an authentication tool (such as the data repository, Zenodo, and some journal peer review systems), by publishers and journals to track article progress with authors, by institutions to build researcher performance profiles and also by research funders. Increasingly, functionality is being added to the research ecosystem to enable researchers to add non-standard research objects to their ORCID records – such as peer reviews (via Publons), grants and datasets. ORCID has received substantial funding (via governmental and non-governmental sources) and is evolving a membership model to build long-term sustainability and governance. It has not yet established itself sufficiently to be truly independent but its progress in this direction is encouraging for the PIDs in the earlier stages of development and adoption by their respective communities such as funder, project and grant IDs¹.

When it was first introduced to the scholarly community, ORCID needed to explicitly offer some value to researchers in the form of self-compiling publication lists in order to justify the time required for them to register and thereby adopt an identifier for themselves. This offer was made by ORCID because – in the

¹ The Open Funder Registry is being used as an identifier for funders, (<https://www.crossref.org/services/funder-registry/>), and the grant identifier project is being initiated by Crossref and DataCite as an extension of this. Meanwhile, the Research Activity Identifier (RAiD) is being developed in Australia to help institutions identify and track research projects.

first instance – the use of an identifier could not be mandated, in contrast to a researcher's social security number, tax file number, or other government mandated identification.

In the research context, scholars are constantly called upon to juggle a wide range and large volume of research and related administrative tasks. Whilst they endeavour to complete all the tasks requested of them, they must prioritize based on how they think the tasks will benefit their career. In order for scholars to benefit from the use of persistent identifiers, there needs to be a clear case made to them about the benefits of taking the time to do so. At the same time, those groups trying to increase uptake and use of PIDs need to concentrate on lowering the barriers to entry (in terms of time and technical involvement) and strengthening the rewards for joining. Ultimately, a researcher should have little to do with managing PIDs. PIDs are just an underlying technology they generally take for granted.

Future technical development and sustainability

The papers of this Special Collection describe concepts and practices in the use of PIDs, and they point to some present and future challenges. These challenges are technical and organisational in their nature.

Car et al (2017) and Golodoniuc et al (2017) ask how PID systems can address the challenge of a fundamental change in the way the internet operates. The current system of URLs and HTTP has been around for more than 25 years. But can we expect it to still exist in 25 years from now? Are distributed PID systems more resilient to these fundamental changes?

The interoperability of semantic web applications depends on the unambiguous identification of terms in controlled vocabularies. At present, these terms are identified by URIs in HTTP form. While not required, it is considered that these URI can double as a URL that points to a human or machine readable web resource (Sauermaun and Cyganiak, 2008). The Achilles heel of this proposal is the domain name, or base URL, that is assumed to be immutable (Berners-Lee, 1998). Klump and Huber (2017) challenge this assumption and propose that there is a role for PID systems in the longevity of semantic web applications.

A source of debate since the beginning of PID systems have been questions around where to target of the resolver, how to implement content negotiation, and at what granularity to identify objects. Kunze et al (2017), in their paper propose a common vocabulary to describe PID systems and their expected behaviours. Adopting a common terminology will certainly help in further developing the principles and practices of PID systems in the same way as the terminology developed for the Open Archival Information Systems Reference Model (CCSDS, 2012) was used in the development of technologies, best practices and policies around archival information systems.

But not all challenges are technical in nature. All PID systems started out as projects. The transition from project to sustainable infrastructure is challenging and requires data infrastructures, publishers, funders and other infrastructure stakeholders to engage in order to understand the value proposition of PIDs, assess priorities for action, funding and research, and determine the drivers and barriers. In order for a specific PID system to be widely trusted and used, there should be an underlying trustworthy sustainable model for its continued existence. Operating a sustainable PID system will always have costs associated, be they time or financial. This should be understood by all stakeholders before any specific PID is mandated.

Outlook

Throughout the articles of this special collection, it becomes evident that the usage and conceptual design of PIDs drives a continuing evolution both at the technological and the organizational level.

The changing value proposition has led to a broadening of the user base, bringing citation as the most prominent traditional use case to new artefacts (Aquino et al, 2017; Conze et al, 2017; Kraft et al, 2017). This is, however, today only one dimension of the overall PID uptake as visible from the changing value proposition outlined earlier. Adoption possibilities are also manifold: there are paths of hidden functionality in user tools or infrastructure services separated from dedicated PID services. Similarly, the delivery of a critical added value must clearly go beyond data citation, and the factors blocking such uptake must see more systematic investigation (Dappert et al, 2017). The applications may also suffer from remaining weaknesses at the conceptual level, with unresolved problems remaining as the foundational concerns as Kunze et al (2017) clearly demonstrate.

Organizational approaches that are sustainable and match all stakeholder interests are only now emerging. The establishment of the DONA Foundation is a key milestone in this direction, but its procedures still have a long road to becoming mature and have a lasting impact. Distributed approaches at the technical level (Golodoniuc et al, 2017) may contribute also to these organizational balance aspects, providing the

necessary framework in which diverse policies can co-exist, but they are not sufficient. The organizational level needs impetus such as the calls for transparency and evaluation criteria (Klump and Huber, 2017), which should make quality assessment procedures a key factor. But only a fully synthesized approach will be able to deliver the necessary stability, and systematic approaches such as those developed by Car et al (2017) may become essential in driving this evolution.

This special issue has set the scene for these current and possible future developments and shed light on the next evolution steps the PID community has to tackle.

There have been some clear – and strengthening – successes within the publishing industry, and some new use cases are emerging outside of publishing. However, some obvious challenges remain. These include the need for sustainability and long-term technical stability in order to build trust. There is also a behavioural gulf between community support (which can be nominal) as opposed to community adoption (which involves a cultural shift). Tensions still remain between the relative merits of distributed versus centralised systems. Finally, we conclude that there is a clear, imminent need for a reference model for PIDs.

Acknowledgements

The authors would like to thank all contributors to the session ‘20 years of persistent identifiers – where do we go next?’ at the European Geosciences Union General Assembly 2016 and all session participants for their input and the fruitful discussions that lead to this Data Science Journal Special Collection.

The authors, as editors of this Data Science Journal Special Collection, would also like to thank the reviewers who helped us to raise the quality of the papers in this collection to a very high standard.

Competing Interests

The authors have no competing interests to declare.

References


- Arms, W Y** 1995 Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine*. Available at: <http://www.dlib.org/dlib/July95/07arms.html>.
- Aquino, J, Allison, J, Rilling, R, Stott, D, Young, K and Daniels, M** 2017 Motivation and Strategies for Implementing Digital Object Identifiers (DOIs) at NCAR's Earth Observing Laboratory – Past Progress and Future Collaborations. *Data Science Journal*, 16(0). DOI: <https://doi.org/10.5334/dsj-2017-007>
- Berners-Lee, T** 1998 Cool URIs don't change. Cambridge, MA: World Wide Web Consortium (W3C). Available at: <http://www.w3.org/Provider/Style/URI>.
- Broeder, D and Lannom, L** 2014 Data Type Registries: A Research Data Alliance Working Group. *D-Lib Magazine*, 20(1/2). DOI: <https://doi.org/10.1045/january2014-broeder>
- Car, N J, Golodoniuc, P and Klump, J** 2017 The Challenge of Ensuring Persistency of Identifier Systems in the World of Ever-Changing Technology. *Data Science Journal*, 16(13): 1–18. DOI: <https://doi.org/10.5334/dsj-2017-013>
- CCSDS** 2012 Reference Model for an Open Archival Information System (OAIS). Magenta Book (Recommended Practice No. CCSDS 650.0-M-2). Greenbelt, MD: Consultative Committee for Space Data Systems. Available at: <http://public.ccsds.org/publications/archive/650x0m2.pdf>.
- Conze, R, Lorenz, H, Ulbricht, D, Elger, K and Gorgas, T** 2017 Utilizing the International Geo Sample Number Concept in Continental Scientific Drilling During ICDP Expedition COSC-1. *Data Science Journal*, 16(1): 1–8. DOI: <https://doi.org/10.5334/dsj-2017-002>
- Dappert, A, Farquhar, A, Kotarski, R and Hewlett, K** 2017 Connecting the Persistent Identifier Ecosystem: Building the Technical and Human Infrastructure for Open Research. *Data Science Journal*, 16. DOI: <https://doi.org/10.5334/dsj-2017-028>
- Dellavalle, R P, Hester, E J, Heilig, L F, Drake, A L, Kuntzman, J W, Graber, M and Schilling, L M** 2003 Going, Going, Gone: Lost Internet References. *Science*, 302(5646): 787–788. DOI: <https://doi.org/10.1126/science.1088234>
- Duerr, R E, Downs, R R, Tilmes, C, Barkstrom, B, Lenhardt, W C, Glassy, J, Bermudez, L E and Slaughter, P** 2011 On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, 4(3): 139–160. DOI: <https://doi.org/10.1007/s12145-011-0083-6>
- Erickson, J S and Lannom, L W** 2017 In Memoriam: Dr. Norman Paskin, Founding Director, International DOI Foundation Visionary Pioneer of Reference Linking, Metadata and Persistent Identifier Infrastructure. *Data Science Journal*, 16(40): 1–2. DOI: <https://doi.org/10.5334/dsj-2017-040>

- Genova, F, Arviset, C, Almas, B, Bartolo, L, Broeder, D, Law, E and McMahon, B** 2017 Building a Disciplinary, World-Wide Data Infrastructure. *Data Science Journal*, 16(16). DOI: <https://doi.org/10.5334/dsj-2017-016>
- Goldstein, J C, Mayernik, M S and Ramapriyan, H K** 2017 Identifiers for Earth Science Data Sets: Where We Have Been and Where We Need to Go. *Data Science Journal*, 16(0). DOI: <https://doi.org/10.5334/dsj-2017-023>
- Golodoniuc, P, Car, N J and Klump, J** 2017 Distributed Persistent Identifiers System Design. *Data Science Journal*, 16: 34. DOI: <https://doi.org/10.5334/dsj-2017-034>
- Güntsch, A, Hyam, R, Hagedorn, G, Chagnoux, S, Röpert, D, Casino, A, Droege, G, Glöckler, F, Gödderz, K, Groom, Q, Hoffmann, J, Holleman, A, Kempa, M, Koivula, H, Marhold, K, Nicolson, N, Smith, V S and Triebel, D** 2017 Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. *The Journal of Biological Databases and Curation*, 2017(1). DOI: <https://doi.org/10.1093/database/bax003>
- Harter, S P and Kim, H J** 1996 Electronic journals and scholarly communication: a citation and reference study. *Information Research*, 2(1). Available at: <http://www.informationr.net/ir/2-1/paper9a.html>.
- Kahn, R and Wilensky, R** 1995 A Framework for Distributed Digital Object Services (Technical Note No. tn09-01). Reston, VA: Corporation for National Research Initiatives. Available at: <http://hdl.handle.net/cnri.dlib/tn95-01>.
- Klump, J F and Huber, R X** 2017 20 Years of persistent identifiers – Which systems are here to stay? *Data Science Journal*, 16(9): 1–7. DOI: <https://doi.org/10.5334/dsj-2017-009>
- Kraft, A, Dreyer, B, Löwe, P and Ziedorn, F** 2017 14 Years of PID Services at the German National Library of Science and Technology (TIB): Connected Frameworks, Research Data and Lessons Learned from a National Research Library Perspective. *Data Science Journal*, 16. DOI: <https://doi.org/10.5334/dsj-2017-036>
- Kunze, J, Calvert, S, DeBarry, J D, Hanlon, M, Janée, G and Sweat, S** 2017 Persistence Statements: Describing Digital Stickiness. *Data Science Journal*, 16(0). DOI: <https://doi.org/10.5334/dsj-2017-039>
- Lynch, C** 1997 Identifiers and Their Role In Networked Information Applications. *ARL: A Bimonthly Newsletter of Research Library Issues and Actions*. Available at: <http://www.arl.org/newsltr/194/identifier.html>.
- Mayernik, M S, Phillips, J and Nienhouse, E** 2016 Linking Publications and Data: Challenges, Trends, and Opportunities. *D-Lib Magazine*, 22(5/6). DOI: <https://doi.org/10.1045/may2016-mayernik>
- Rauber, A, Asmi, A, van Uitvanck, D and Pröll, S** 2016 Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC) (Technical Report). Denver, CO: Research Data Alliance. DOI: <https://doi.org/10.15497/RDA00016> (Last accessed 21 September 2017).
- Sauermann, L and Cyganiak, R** 2008 Cool URIs for the Semantic Web (Technical Report, W3C Interest Group Note 03 December 2008). Cambridge, MA: World Wide Web Consortium (W3C). Available at: <https://www.w3.org/TR/cooluris/> (Last accessed 14 March 2016).
- Wanchoo, L, James, N and Ramapriyan, H** 2017 NASA EOSDIS Data Identifiers: Approach and System. *Data Science Journal*, 16(0). DOI: <https://doi.org/10.5334/dsj-2017-015>
- Wang, J, Car, N, Evans, B, Gohar, K, Trenham, C and Wyborn, L** 2017 Persistent Identifier Practice for Big Data Management at NCI. *Data Science Journal*, 16(0). DOI: <https://doi.org/10.5334/dsj-2017-020>
- Weigel, T, DiLauro, T and Zastrow, T** 2016 PID Information Types WG final deliverable (Technical Report). Denver, CO: Research Data Alliance. DOI: <https://doi.org/10.15497/FDAA09D5-5ED0-403D-B97A-2675E1EBE786>
- Wood, J, Andersson, T, Bachem, A, Best, C, Genova, F, Lopez, D R, Los, W, Marinucci, M, Romary, L, Van de Sompel, H, Vigen, J, Wittenburg, P and Giarretta, D** 2010 Riding the Wave: How Europe can gain from the rising tide of scientific data. Brussels, Belgium: European Commission. Available at: http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204.

How to cite this article: Klump, J, Murphy, F, Weigel, T and Parsons, M 2017 Editorial: 20 Years of Persistent Identifiers – Applications and Future Directions. *Data Science Journal*, 16: 52, pp. 1–7, DOI: <https://doi.org/10.5334/dsj-2017-052>

Submitted: 09 October 2017 **Accepted:** 10 November 2017 **Published:** 11 December 2017

Copyright: © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 