

## REVIEW

# 20 Years of Persistent Identifiers – Which Systems are Here to Stay?

Jens Klump<sup>1</sup> and Robert Huber<sup>2</sup><sup>1</sup> CSIRO, Mineral Resources, Perth, AU<sup>2</sup> MARUM, University of Bremen, Bremen, DECorresponding author: Jens Klump ([jens.klump@csiro.au](mailto:jens.klump@csiro.au))

Web-based persistent identifiers have been around for more than 20 years, a period long enough for us to start observing patterns of success and failure. Persistent identifiers were invented to address challenges arising from the distributed and disorganised nature of the internet, which often resulted in URLs to internet endpoints becoming invalid. Over the years several different persistent identifier systems have been applied to the identification of research data, not all with the same level of success in terms of uptake and sustainability. We investigate the uptake of persistent identifier systems and discuss the factors that might determine the stability and longevity of these systems. Persistent identifiers have become essential elements of global research data infrastructures. Understanding the factors that influence the stability and longevity of persistent identifier systems will help us guide the future development of this important element of research data infrastructures and will make it easier to adapt to future technological and organisational changes.

**Keywords:** persistent identifiers; semantic web; research data repositories

## Introduction

Web-based persistent identifiers have been around for more than 20 years, a period long enough for us to start observing patterns of success and failure. Persistent identifiers were invented to address challenges arising from the distributed and disorganised nature of the internet, which not only allowed new technologies to emerge, it also made it difficult to maintain a persistent record of science (Dellavalle et al. 2003; Lawrence et al. 2001). This phenomenon, also dubbed “link rot”, affects all digital resources on the web, including research data (Vines et al. 2014).

It has been argued that “link rot” can be avoided by careful management of web servers to keep URLs stable over a long time, a principle called “Cool URIs” (Berners-Lee 1998). For semantic applications the use of Cool URIs has been proposed and it has been questioned whether DOI are necessary in a world of Cool URI (Bazzanella, Bortoli, and Bouquet 2013).

“Pretty much the only good reason for a document to disappear from the Web is that the company which owned the domain name went out of business or can no longer afford to keep the server running.” (Berners-Lee 1998).

Unforeseen to Berners-Lee, a few years after his statement the “dot.com bubble” burst and many companies went out of business, leaving many web domains orphaned. Other companies were acquired and merged into existing entities, and again sometimes losing their original web domain.

One way of addressing the root problem of the persistence of locators on the web was by the introduction of persistent identifiers which separated the identity of an object from its location on the web (see e.g. Arms 1995; Lawrence et al. 2001; Lynch 1997). Adding a system to ensure global uniqueness makes persistent identifiers a tool that allows us unambiguous identification of resources on the net. The expectations were

that persistent identifiers would lead to greater accessibility, transparency and reproducibility of research results. The discussion of PID vs. Cool URI in Bazzanella, Bortoli, and Bouquet (2013) shows that persistent access to web resources is not merely a technical question, but rather a “social contract” that needs to be entered by the stakeholders aiming to maintain persistent references to objects on the web.

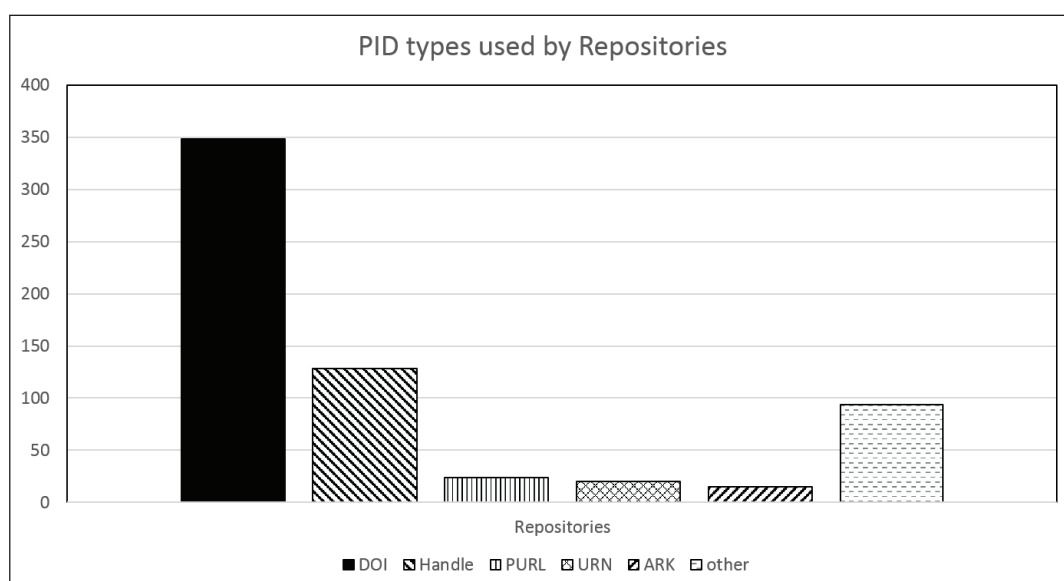
In this paper we want to review 20 years of persistent identifier practice and the uptake of different persistent identifier systems. In a series of case studies we want to characterise well known persistent identifier systems, assess their successes and failures, and extract what can be learned from these examples.

## Uptake of Persistent Identifiers

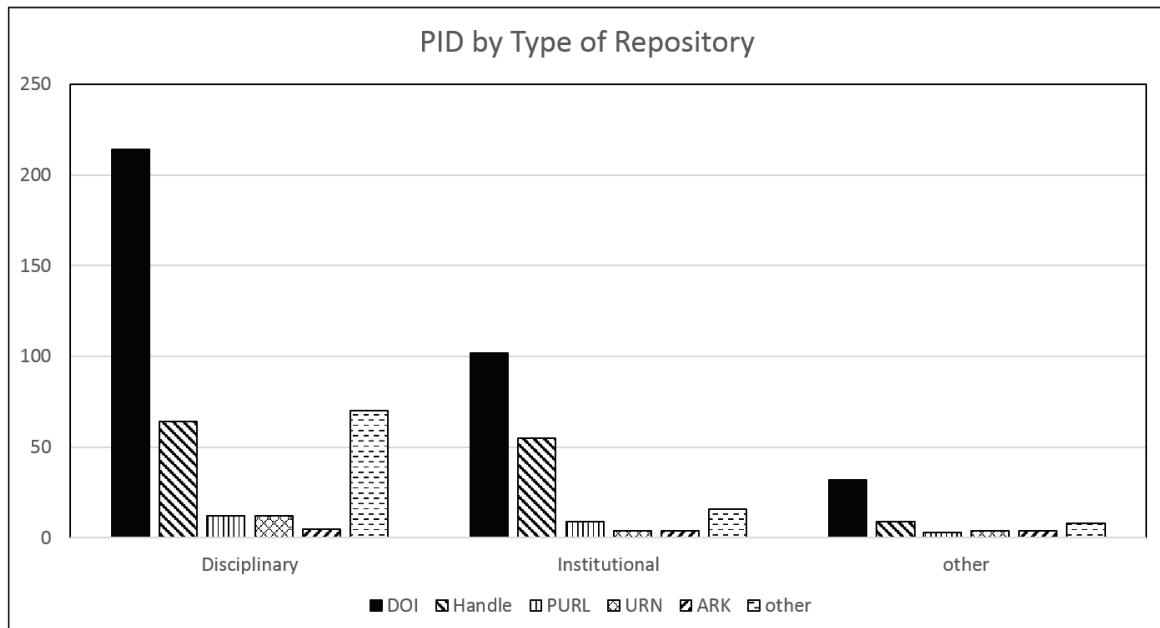
One way to assess the success of particular identifier systems is to survey their adoption by research data repositories. This might seem like a straightforward approach, but it turns out to be difficult to define a measure for the uptake of persistent identifier systems because the sizes of research data repositories and the granularities vary by orders of magnitude. Our analysis of the uptake of persistent identifier systems by research data repositories is based on data from the Registry of Research Data Repositories (re3data.org). re3data.org is a global registry of research data repositories that covers all academic disciplines. The registry arose from two separate projects, re3data.org (Pampel et al. 2013) and DataBib (Witt and Giarlo 2012) and is now managed by DataCite. The total sample we obtained from re3data.org in December 2015 listed 1381 repositories. Out of this total a subset of 475 repositories used some type of persistent identifier. Note that some repositories use more than one type of persistent identifier. **Figure 1** summarises the use of persistent identifier types used by repositories listed in the re3data.org database.

The focus of re3data.org is on research data repositories and despite its size the re3data.org registry does not claim global coverage. Still, its catalogue can be considered to be a representative overview. Not covered by re3data.org are collections of research specimens like herbaria or cultural artefacts, which might also use persistent identifiers for the identification of items in their collections. Furthermore, identifiers are also used outside of research, for example in the identification of companies on stock exchanges. In this paper we will only discuss the use of persistent identifiers in the context of research data and the record of science.

In their standardised descriptions of research data repositories re3data.org distinguish between only a few identifier systems (Rücknagel et al. 2015). Digital Object Identifiers (DOI) are clearly the most widely adopted persistent identifier in research data repository systems. **Figure 2** differentiates persistent identifiers used by three kinds of repositories, namely disciplinary repositories, institutional repositories and repositories that fall in neither of the two former categories. Remarkable is the relatively frequent use of “other” persistent identifier systems, not differentiated in the re3data.org description, by disciplinary repositories. This points at an important role of discipline specific identifiers.



**Figure 1:** Number of repositories using a particular type of persistent identifier. A total of 457 out of 1381 repositories (status of 14 Dec 2015) use some sort of persistent identifier. Some repositories use more than one type of persistent identifier.



**Figure 2:** Types of repositories using a particular type of persistent identifier. Note that some repositories use more than one type of persistent identifier and are defined as both disciplinary and institutional. “Other” PID seem to be important in disciplinary repositories, pointing to particular disciplinary groups of practice. Handle PID (non-DOI) are used by many institutional repositories, possibly due to use of Handle in repository software and early adoption.

Again, DOI are by far the most used persistent identifiers. “Other” types of identifiers seem to play an important role in disciplinary repositories and 57 of these repositories serve the life sciences. This indicates a special role that “other” identifier systems play in particular disciplines. Institutional repositories frequently use identifiers based on the Handle system, which may be due to the fact that some institutional repository software, like DSpace (Smith et al. 2003), use Handle-based identifiers to identify objects in their holdings.

## Years of Crisis

While persistent identifiers are adopted and implemented by a growing number of data archives, not every PID system experienced a success story during the last years. A few PID systems even experienced severe problems to the degree that lead to a temporary shutdown of some of their core services, which in turn led to orphaned, unresolvable or unmanaged PIDs.

As mentioned above, the years 2015 and 2016 turned out to be years of crisis for some persistent identifier systems, in particular for Persistent URL (PURL) and Life Science Identifiers (LSID). While PURL seems to have gained a new lease on life through transferring to a new organisational and technical base, the future of LSID as a resolvable persistent identifier seems uncertain.

PURL was introduced by the Online Computer Library Center, Inc. (OCLC) as a bridging technology to prepare for introduction of Universal Resource Names (URN). PURL implements the URI concept and thus it does not separate between identifier and resolving mechanism. PURL has no single global resolving mechanism and PURL resolvers do not communicate amongst each other to share resolving information like DNS or Handle servers would do. For most of its history PURL had little social infrastructure and formal governance. In 2014 OCLC withdrew its institutional support and the future of PURL became unclear while PURL experienced severe technical problems for some time and the system was put into a ‘read-only’ maintenance mode (Baker 2015). In September 2016 OCLC and the Internet Archive announced that the URL redirection service, on which PURL is based, will in future be operated by the Internet Archive (OCLC 2016). This move brought PURL back from the brink of extinction. In December 2015 a total of 16 research data repositories in re3data.org were listed as using PURL, and only few of them were using PURL exclusively. Using Google Scholar as a search engine we estimate that about 16,400 PURL identifiers are being used in the entire scholarly record indexed by Google Scholar. Of these, less than 5,000 seem to identify digital objects like data, most seem to identify semantic concepts.

LSIDs had been introduced by the Object Management Group (OMG) in 2004 as a way to naming and identifying data resources stored in multiple, distributed data stores. From 2009 onwards the biodiversity informatics communities' standardisation authority (Taxonomic Database Working Group, TDWG) strongly supported LSIDs as the preferred GUID technology. LSIDs were thought to be used by all globally leading providers for biodiversity data to identify organism names. However, LSIDs provide neither a global resolving mechanism nor a centralised provider registration. The implementation of this standard is relatively complex, resolution is DNS based and requires a multistep procedure, the associated metadata format is RDF. As a consequence the technology was controversially disputed (e.g. Hyam 2015; Page 2016) and opinions in discussion forums seemed to favour a simpler identifier system such as HTTP URIs.

As a result of these difficulties the system remained fragmented and fragile. In 2016, maintenance on TDWG's LSID resolution service was terminated and TDWG's support of LSIDs came into questioning by members of the group (TDWG 2016). After two months without a central resolving system, a resolver has been made available at <http://www.lsid.info>. However, the discussion is ongoing and significant parts of the biodiversity informatics community recommend switching from LSID to cool URI (Guralnick et al. 2015). Using Google Scholar as a search engine we estimated that about 14,000 LSIDs have been used in the scientific literature.

### Which are here to stay?

At the same time as criteria for trusted repositories were developed (Dobratz et al. 2009; Sesink, van Horik, and Harmsen 2008), similar efforts looked at criteria for trustworthy persistent identifier systems. Most notable are the criteria for trusted persistent identifier systems developed by Bütikofer (2009) in the context of the German nestor research programme on long-term preservation, and the review of persistent identifier systems as tools for science by Duerr et al. (2011). While the criteria of Bütikofer emphasize technical and organisational criteria, the review of Duerr et al. focuses more on usability of identifier systems as part of the academic record. Even though the authors come to different conclusions about which systems are likely to persist, both recognise the importance of organisational sustainability. If organisational stability is the Achilles Heel of persistent identifier systems, are there ways we can achieve better sustainability of PID systems?

A first step towards better sustainability of PID systems would be more transparency. This should include all aspects of a PID system, technical documentation, policies, governance and in particular the data and metadata which are necessary to resolve a PID.

Today, most of discussions related to the status of PID systems are hidden in online discussion fora and email lists and is only rarely made public. It is entirely unsatisfactory to publicly and officially promote a PID system while exit strategies are being discussed in the background or services are silently ceased. This needs to change, and clearly a more participatory attitude and proactive communication strategy would be beneficial for all PID systems stakeholders.

A set of criteria, analogous to the criteria for the description of research data repositories published by re3data.org (Rücknagel et al. 2015), would help with the evaluation of PID systems. In conjunction with the re3data.org criteria, they would also help to identify weaknesses in the shared responsibility of the data provider and the operator of the PID resolver service for a reliable resolution of identifiers to web endpoints.

The large number of discipline specific resolver systems tells us that there may be very specific needs in the governance of a PID system that are not met by the generic services. Here it is necessary to have a close look at the value proposition of a particular PID system and the services it provides.

In addition to organisational criteria, the value proposition of a particular PID system also asks us to evaluate its technical basis and alternative technical solutions. As we have seen from the LSID example, the seeming simplicity of Cool URIs is still attractive. The HTTP protocol is simple and universally available but the risks of "link rot" have not gone away. Other interesting technical alternatives are based on peer-to-peer networking technologies such as Blockchain (Bolikowski, Nowiński, and Sylwestrzak 2015) or MagnetLinks (Golodoniuc, Car, and Klump, this volume). Peer-to-peer technologies would allow a "Devolution of Power" and community-based backup strategies for PID resolution.

Coming back to the question of the value proposition of PID systems, do we need PID resolvers? Yes, because we do not yet live in a semantic web world where linked data graphs would lead us to resources as proposed by Sachs and Finin (2010). As an interim solution data providers should take advantage of available web search engines and make their data holdings discoverable. A possible approach would be to use mainstream web technologies as a supplement and potential fall-back solution to PID systems. Candidate technologies are microformats or JSON-LD which are suited to expose both, metadata as well as potentially

multiple identifiers associated to a digital object. Complementary sitemaps or catalogue services catalogued in a publicly available registry such as the GEOSS CSR could enable the implementation of common, generic resolution services.

There is, however, the other value proposition of PID, the persistent identification of elements of the record of science. Properly identifying these elements in a way that can be consumed by human and machine clients alike, and maintaining the persistence of objects and identifier resolution, is not a purely technical problem but is maintained through a social contract. The stability of this social contract, together with a sustainable and adaptable technological base, will determine the sustainability and resilience of a PID system. It is tempting to assume that a social contract becomes increasingly binding as user community relying on a PID particular system grows. With the examples discussed in this paper we show that this is most likely an illusion. The DOI system, which is arguably the most successful PID system today, has a strong commercial backing while minor systems such as URN and ARK have the backing of national libraries. It might be a bitter pill to swallow for some members in the research data community wary of all things commercial, but business models are essential aspects of PID systems – sustainable PID systems do not come for free.

## Acknowledgements

The authors would like to thank the Registry of Research Data Repositories (re3data.org) for providing an excerpt of the re3data.org database. We wish to acknowledge the European Commission for their funding of the projects ENVRIplus (Reference number: 654182) and THOR (Reference number 654039), as well as funding by the German Research Foundation (DFG) of the project GFBio. We also thank the reviewers for their constructive comments that helped to improve this manuscript.

## Competing Interests

The authors have no competing interests to declare.

## About the Authors

**Jens Klump** is a geochemist by training and OCE Science Leader Earth Science Informatics in CSIRO Mineral Resources. Jens' field of research is data intensive science. Research topics in this field are numerical methods in minerals exploration, virtual research environments, remotely operated instruments, high performance and cloud computing, and the development of system solutions for geoscience projects. In his previous position at the German Research Centre for Geosciences in Potsdam he was involved in the development of the publication and citation of research data through Digital Object Identifiers. This project sparked further work on research data infrastructures, including the publication and curation of scientific software and reproducible research.

**Robert Huber** is a geologist and information specialist holding a PhD in Marine Geology. He has worked for several years as information system architect for the aerospace industry and the renewable energy industry. Since 2002 he is employed at the Centre for Marine Environmental Sciences (MARUM) at the University Bremen and responsible for projects in scientific data management and IT development especially in the fields of ontology development, marine observatory networks and biodiversity in the PANGAEA working group.

## References


- Arms, W Y** 1995 Key Concepts in the Architecture of the Digital Library. *D-Lib Magazine*, July. <https://www.cnri.dlib/july95-arms>.
- Baker, T** 2015 The Future of PURLs. *JISC DC Architecture*. Retrieved from: <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1511&L=DC-ARCHITECTURE&F=&S=&P=3711>.
- Bazzanella, B, Bortoli, S and Bouquet, P** 2013 Can Persistent Identifiers Be Cool? *International Journal of Digital Curation*, 8(1): 14–28. DOI: <https://doi.org/10.2218/ijdc.v8i1.246>
- Berners-Lee, T** 1998 Cool URIs Don't Change. Cambridge, MA: World Wide Web Consortium (W3C). Retrieved from: <http://www.w3.org/Provider/Style/URI>.
- Bolikowski, Ł, Nowiński, A and Sylwestrzak, W** 2015 A System for Distributed Minting and Management of Persistent Identifiers. *International Journal of Digital Curation*, 10(1): 280–86. DOI: <https://doi.org/10.2218/ijdc.v10i1.368>
- Bütikofer, N** 2009 Catalogue of Criteria for Assessing the Trustworthiness of PI Systems. 13. Nestor-Materialien. Göttingen, Germany: Niedersächsische Staats und Universitätsbibliothek Göttingen. Retrieved from: <http://nbn-resolving.de/urn:nbn:de:0008-20080710227>.

- 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): 787–88. DOI: <https://doi.org/10.1126/science.1088234>
- Dobratz, S, Hänger, A, Huth, K, Kaiser, M, Keitel, C, Klump, J, Rödiger, P, et al.** 2009 Catalogue of Criteria for Trusted Digital Repositories. 8. Nestor Materials. Frankfurt (Main), Germany: Deutsche Nationalbibliothek. Retrieved from: <http://nbn-resolving.de/urn:nbn:de:0008-2010030806>.
- 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): 139–60. DOI: <https://doi.org/10.1007/s12145-011-0083-6>
- Golodoniuc, P, Car, N J and Klump, J** subm. Distributed Persistent Identifiers System Design. *Data Science Journal*.
- Guralnick, R P, Cellinese, N, Deck, J, Pyle, R L, Kunze, J, Penev, L, Walls, R, et al.** 2015 Community Next Steps for Making Globally Unique Identifiers Work for Biocollections Data. *ZooKeys*, 494(April): 133–54. DOI: <https://doi.org/10.3897/zookeys.494.9352>
- Hyam, R** 2015 Taxa, Taxon Names and Globally Unique Identifiers in Perspective'. In: Watson, M F, Lyal, C and Pendry, C (Eds.) *Descriptive Taxonomy: The Foundation of Biodiversity Research*. Cambridge, United Kingdom: Cambridge University Press, pp. 260–71. DOI: <https://doi.org/10.1017/CBO9781139028004.026>
- Lawrence, S, Coetzee, F, Glover, E, Pennock, D, Flake, G, Nielsen, F, Krovetz, R, Kruger, A and Giles, L** 2001 Persistence of Web References in Scientific Research. *IEEE Computer*, 34(2): 26–31. DOI: <https://doi.org/10.1109/2.901164>
- Lynch, C** 1997 (October) Identifiers and Their Role In Networked Information Applications. *ARL: A Bimonthly Newsletter of Research Library Issues and Actions*. Retrieved from: <http://www.arl.org/newsltr/194/identifier.html>.
- Page, R** 2016 Surfacing the Deep Data of Taxonomy. *ZooKeys*, 550(January): 247–60. DOI: <https://doi.org/10.3897/zookeys.550.9293>
- Pampel, H, Vierkant, P, Scholze, F, Bertelmann, R, Kindling, M, Klump, J, Goebelbecker, H-J, Gundlach, J, Schirnbacher, P and Dierolf, U** 2013 Making Research Data Repositories Visible: The re3data.org Registry. *PLoS ONE*, 8(11): e78080. DOI: <https://doi.org/10.1371/journal.pone.0078080>
- Rücknagel, J, Vierkant, P, Ulrich, R, Kloska, G, Schnepf, E, Fichtmüller, D, Reuter, E, et al.** 2015 *Meta-data Schema for the Description of Research Data Repositories*. Version 3.0. Potsdam, Germany: German Research Centre for Geosciences. DOI: <https://doi.org/10.2312/re3.008>
- Sachs, J and Finin, T** 2010 What Does It Mean for a URI to Resolve? In: *Proceedings of the AAAI Spring Symposium on Linked Data Meets Artificial Intelligence*, 3. Stanford, CA: AAAI Press. Retrieved from: <http://www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1178>.
- Sesink, L, van Horik, R and Harmsen, H** 2008 *Data Seal of Approval*. Den Haag, The Netherlands: Data Archiving and Networked Services (DANS). Retrieved from: <http://www.datasealofapproval.org/>.
- Smith, M, Barton, M, Branschovsky, M, McClellan, G, Walker, J H, Bass, M, Stuve, D and Tansley, R** 2003 DSpace: An Open Source Dynamic Digital Repository'. *D-Lib Magazine*, 9(1). DOI: <https://doi.org/10.1045/january2003-smith>
- TDWG** 2016 (September 28) Decide about Fate of Lsid.tdwg.org. GitHub. *TDWG/Infrastructure*. Retrieved from: <https://github.com/tdwg/infrastructure/issues/60>.
- Vines, T H, Albert, A Y K, Andrew, R L, Débarre, F, Bock, D G, Franklin, M T, Gilbert, K J, Moore, J-S, Renaut, S and Rennison, D J** 2014 The Availability of Research Data Declines Rapidly with Article Age. *Current Biology*, 24(1): 94097. DOI: <https://doi.org/10.1016/j.cub.2013.11.014>
- Witt, M and Giarlo, M** 2012 Databib. *Libraries Faculty and Staff Presentations*, January. Retrieved from: [http://docs.lib.purdue.edu/lib\\_fspress/1](http://docs.lib.purdue.edu/lib_fspress/1).

**How to cite this article:** Klump, J and Huber, R 2017 20 Years of Persistent Identifiers – Which Systems are Here to Stay? *Data Science Journal*, 16: 9, pp. 1–7, DOI: <https://doi.org/10.5334/dsj-2017-009>

**Submitted:** 18 November 2016    **Accepted:** 14 February 2017    **Published:** 22 March 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** 