DOI Primer

Chris Hartgerink chris@libscie.org & Mikael Laakso mikael.laakso@hanken.fi

The Digital Object Identifier (DOI) system is like a digital address book: One DOI name resolves to information on the location of the content. It is conceptually comparable to a regular phonebook: A name resolves to a phone number with which you can contact a person; the name (typically) does not change but the number is dynamic and may change quite often. A DOI name resolves to metadata about the content, which for many intents and purposes of end-users is a Uniform Resource Locator (URL). For example, doi:10.1016/j.wneu.2019.07.027 resolves to https://linkinghub.elsevier.com/retrieve/pii/S1878875019319333 (as of 2019-07-10). The difference with the phonebook analogy is that compared to names of natural persons, which are allowed to change in certain circumstances, a DOI name is not permitted to change under any circumstance, which makes it a Persistent Identifier (PID).

A DOI name can be assigned to any physical, digital, and abstract entity of any degree of granularity. For example, an article may be assigned a DOI name, but a figure in that article may be assigned one too. Moreover, DOI names are not just limited to objects in scholarly publications; DOI names have been issued for episodes of TV shows, movies, government reports, books, datasets, and much more. A DOI name digitally identifies an object, but a DOI name does not make the object itself persistent. As a result, any kind of object may be digitally identified, but its location and content may change over time. There is no requirement for the object that has a DOI to be covered by long-term preservation services, hence, it may become unavailable if care is not taken. Moreover, despite DOI names themselves being unique, not all objects have to be; DOI names may be redundant, with multiple identifiers being registered for the same object.

The DOI system aims to facilitate persistence, data updates, interoperability, extensibility, and content management (1.7; Davidson and Douglas 1998). Persistence and metadata updates are important to reduce the effect of link rot (Klein et al. 2014), where a URL leads to an unknown webpage (e.g., a 404 error). Interoperability with data sources or data formats helps the information flow (e.g., in reference management). Extensibility of the DOI system permits for new functionality to be added over time (e.g., CrossRef’s inclusion of references). Content (i.e., intellectual property) management relates to the administration of access, which can be controlled by changing the location the DOI name resolves to (the appropriate copy problem; Mader 2001).

## What is a DOI?

A DOI name is a specific instance of a persistent Handle name, which is a Uniform Resource Identifier (URI). More technically, a DOI is a persistent identifier in line with IETF RFC [3650](http://www.ietf.org/rfc/rfc3650.txt), [3651](http://www.ietf.org/rfc/rfc3651.txt), [3652](http://www.ietf.org/rfc/rfc3652.txt) (HANDLE system), ANSI/NISO Z39.84-2010, and [ISO 26234](https://www.iso.org/standard/43506.html). There is no limit length on a DOI name and it is case insensitive. Given that a DOI name is a specific instantiation of a Handle name (3.5.1; International DOI Foundation 2018), doi:10.1371/journal.pbio.2003693 is the same as hdl:10.1371/journal.pbio.2003693. As a result, you can use a Handle resolver with a DOI; <http://proxy.handle.net/>.

Practically speaking, a DOI name is composed of a prefix and a suffix (i.e., <prefix>/<suffix>). The prefix a directory indicator (10) and the digit registrant code, separated by a dot (e.g., 10.1016 for Elsevier BV, 2.2.2; International DOI Foundation 2018). One organisation might have multiple registrant codes, as the result of acquisitions that were originally their own registrants. Vice versa, unique prefixes do not necessarily denote one organisation completely. In contrast to the prefix, the suffix is a character string that the registrant is free to assign, as long as it is unique for the registrant (2.2.3; International DOI Foundation 2018). By combining the prefix and suffix, all identifiers are unique within the DOI system.

A DOI names one object, which can be resolved to inform us about its current location. Resolving a DOI name does not provide any information on the object’s location history, as that falls outside the scope of the DOI system. The DOI system also does not track how the location has changed over time, despite being of value. Most of the time, DOI resolution occurs through following the https://doi.org/<doi> link in a browser. More systematic and complete resolution is possible through an open API provided by the International DOI Foundation (IDF); [https://doi.org/api/handles/<doi>](https://doi.org/api/handles/) (where <doi> is replaced with the DOI name, see also 3.8.3; International DOI Foundation 2018).

Most often you will see DOI names in reference lists. This is where their persistence and linked metadata comes in handy for both the authors and the publishers. It helps provide a direct link to the metadata that needs to be included, ranging from publication year to its journal issue. Some publication manuals (e.g., American Psychological Association 2010) require reference lists to include DOIs for correct identification of references. Other uses are similar, such as when TV shows and their episodes get DOIs, they are used to provide metadata in television programming.

## Who is behind the DOI?

The International DOI Foundation (IDF) is the main institution behind the DOI system. They provide, develop, and maintain the common agreements and policies to keep the DOI system functioning. This serves as the main governing body where all the involved entities come together to make operations happen in a federated manner.

The IDF is composed of various kinds of members, but the most important ones are the Registration Agencies (RAs). Only general members of the IDF can be RAs, and in order to be a general member a yearly fee of $35,000 is due, with a more complicated and potentially more expensive membership fee when an entity becomes an RA (see 8.6; International DOI Foundation 2018). Registration agencies offer services for the registration of prefixes and specific DOI names within the DOI system (8.1; International DOI Foundation 2018), as a service with an underlying business model. RAs can compete for similar customers and in the case one RA exits or goes out of business, all content is transferred to another RA (8.5; International DOI Foundation 2018). There are various RAs that have been announced previously but no longer appear on the IDF pages as current RAs (e.g., The Stationary Office, Content Directions Inc., Copyright Agency Limited). Each RA can declare the metadata for their DOI names and can have its own policies to cater for its community , which only have to be consistent with the upstream policies from the IDF.

Registrants are the entities who register the DOIs through a Registration Agency that makes sure the information is included in the DOI system. A registrant can be anyone who wants to register DOI names, and may utilise the services from various RAs at the same time. Each RA may set its own conditions for becoming a registrant, including registration fees or membership fees (see for example CrossRef’s fee structure; <https://www.crossref.org/fees/>). Registrants do not need to be members of the IDF when they are registered with the RA. Registrants have to ensure appropriate content management of their materials.

## Short history of the DOI system

The DOI system was introduced in 1998. This was the result of a committee set up by the Association of American Publishers in 1994 (Davidson and Douglas 1998) and had the aim to protect copyright. Initial members of the IDF in 1998 included the Copyright Clearance Center, the Xerox Corporation, various of the large scholarly publishing houses, and Microsoft.

Originally, there were no Registration Agencies (RAs) but only the International DOI Foundation (IDF; Davidson and Douglas 1998). Prefixes were sold for $1,000 and could not be gotten elsewhere. This version of the DOI system necessitated annual fees for each registered DOI.

In 2000, CrossRef started as the first RA (<https://www.crossref.org/about/>). This substantially altered the DOI landscape over the years, becoming more federated as a result. This evolution is not well documented. The exact order that other RAs subsequently came into existence (as best as we could document): [mEDRA (2003)](https://www.medra.org/en/who.htm); [Publications Office of the European Union (OP; 2009)](https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1425380943945&uri=CELEX:32009D0496); [DataCite (2010)](https://www.handelsregister.de/rp_web/charge-info.do) and [Entertainment Identifier Registry (EIDR; 2010)](https://eidr.org/about-us/); [airiti (2011)](http://doi.airiti.com/en/about_airiti_page1.html). We were unable to reliably document the start of the RAs for the China National Knowledge Infrastructure (CNKI), Institute of Scientific and Technical Information of China (ISTIC), Japan Link Center (JaLC), and Korea Institute of Science and Technology Information (KISTI).

### Growth of DOI system

Figure @ref(fig:plot-doi-data) depicts the amount of DOIs per year. This figure includes both the DOI names registered per year per RA and the self-assigned publication year for that object. For example, a DOI name registered in 2008 might refer to a publication from the year 1960. We retrieved these data either from public APIs ([CrossRef](https://api.crossref.org) and [Datacite](https://api.datacite.org); exact queries available in our [Rmarkdown file](./doi-primer.Rmd)), through searches in their webportals (EIDR, mEDRA), and direct contact with the RA (EIDR, OP). We could not retrieve the relevant information for each RA, explaining omissions in our plots.



Yearly counts of DOIs per publication year (top panel) or per registration year (bottom panel). A publication registered in 2008 might refer to an object originally published in 1960, is counted as one occurence in 1960 for the top panel and one occurence for 2008 for the bottom panel.

From Figure @ref(fig:plot-doi-data), it becomes clear that CrossRef is by far the largest RA for DOI names. Nonetheless, DataCite is growing so fast it might overtake CrossRef in the (near) future. Other RAs, if data is available, make up a relatively small amount, even though they are non-trivial in an absolute sense. For example, EIDR has around 1 million registered DOI names in its peak year, 2017. Table @ref(tab:doi-data) shows some summary data about each RA.

Summary data per Registration Agency (RA) and count type (i.e., deposit year or publication year).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ra | type | earliest | latest | maximum |
| crossref | deposited | 1998 | 2019 | 46011566 |
| crossref | published | 1400 | 2203 | 5883887 |
| datacite | deposited | 2004 | 2019 | 3603581 |
| datacite | published | 1 | 2019 | 2624096 |
| eidr | deposited | 2010 | 2018 | 960862 |
| eidr | published | 1878 | 2018 | 90155 |
| jalc | deposited | 2013 | 2018 | 2777478 |
| medra | deposited | 2003 | 2019 | 144155 |
| medra | published | 1950 | 2021 | 62470 |
| op | deposited | 2004 | 2019 | 213268 |
| op | published | 2004 | 2019 | 24892 |

## Limitations of the DOI system

The DOI system relies on Registration Agencies and the registrants to assure the quality of the metadata. By extension, this means that the metadata is only as good as the trust that can be placed in the RA. Inspecting Table @ref(tab:doi-data) already suggests several challenges in the metadata: (1) deposited objects prior to an RA’s existence (e.g., CrossRef started in 2000 but deposited objects occur for 1998); (2) deposited objects for unreasonable future dates (e.g., CrossRef includes publications for the year 2203); (3) deposited objects for questionable past dates (e.g., DataCite includes objects as far back as the year 1). Further manual inspection of the DOI names per publication years shows odd spikes (e.g., the year 1500 in CrossRef contains 101 DOI names). RA specific metadata can be incomplete (e.g., CrossRef metadata allows for abstracts but coverage for one of the large publishers is as low as 0.0023%). If the quality of the metadata cannot be assured, both by its coverage and accuracy, it may undermine the legitimacy of the DOI system.

The DOI system permits persistent digital identification of objects, but it does nothing to permit persistent content and may result in content drift (Klein et al. 2014). In this scenario, the content that a DOI resolves to may change over time, unbeknownst to the user. The DOI system was never designed to permit this functionality; in 1998, Davidson and Douglas (1998) wrote that “*any guarantee of persistence in the digital environment […] is going to require a lot of ongoing detail work.*” This still applies to the DOI system due to its design. Over twenty years since its inception, recent proposals for Persistent Identifiers (PIDs) have taken persistency of content as one of the major issue points of the DOI system (Wannenwetsch and Majchrzak 2016; Sicilia et al. 2019). When web-based scholarly journals shut down, their content might go down with them — unless preservation is taken care of in advance. In a recent study of vanished open access journals at least 7 out of 176 journals that were found to have disappeared also left after them DOIs that fail to resolve, and in worst cases resolve to ad-filled domain parking pages (Laakso, Matthias, and Jahn 2020).

A limitation of the DOI system is the limited scope it has been researched and by extension, our understanding of its practical functioning. Most of the research has focused on describing the DOI system’s properties, comparing the DOI system to other identifiers, describing uptake of the DOI system across various databases, or providing technical descriptions of implementing DOIs. In depth technical analyses of where DOI names resolve to and whether those pages actually exist are scarce, except for recent work on whether DOI name resolution (which indicates they are not as persistent in practice as they aim to be in theory; Klein and Balakireva 2020). Social analyses of whether the DOI system is an equitable system in a global landscape are missing as well. This lack of research makes the DOI system a powerful, but not well understood system.

# References

American Psychological Association. 2010. *Publication Manual of the American Psychological Association*. 6th ed. Washington, DC: American Psychological Association.

Davidson, Lloyd A., and Kimberly Douglas. 1998. “Digital Object Identifiers: Promise and Problems for Scholarly Publishing.” *The Journal of Electronic Publishing* 4 (2). <https://doi.org/10.3998/3336451.0004.203>.

International DOI Foundation. 2018. “DOI Handbook.” <https://doi.org/10.1000/182>.

Klein, Martin, and Lyudmila Balakireva. 2020. “On the Persistence of Persistent Identifiers of the Scholarly Web.” <http://arxiv.org/abs/2004.03011>.

Klein, Martin, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. “Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot.” Edited by Judit Bar-Ilan. *PLoS ONE* 9 (12): e115253. <https://doi.org/10.1371/journal.pone.0115253>.

Laakso, Mikael, Lisa Matthias, and Najko Jahn. 2020. “Open Is Not Forever: A Study of Vanished Open Access Journals.” <http://arxiv.org/abs/2008.11933>.

Mader, Cynthia L. 2001. “Current Implementation of the DOI in STM Publishing.” *Science & Technology Libraries* 21 (1-2): 97–118. <https://doi.org/10.1300/j122v21n01_09>.

Sicilia, Miguel-Angel, Elena Garcı́a-Barriocanal, Salvador Sánchez-Alonso, and Juan-Jose Cuadrado. 2019. “Decentralized Persistent Identifiers: A Basic Model for Immutable Handlers.” *Procedia Computer Science* 146: 123–30. <https://doi.org/10.1016/j.procs.2019.01.087>.

Wannenwetsch, Oliver, and Tim Alexander Majchrzak. 2016. “On Constructing Persistent Identifiers with Persistent Resolution Targets.” In *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*. IEEE. <https://doi.org/10.15439/2016f87>.