PERSPECTIVES
INTERNATIONAL MICROBIOLOGY (2006) 9:139-142
www.im.microbios.org

Miquel Termens

Department of Library Science and Documentation, University of Barcelona, and Institute for Catalan Studies, Barcelona, Catalonia, Spain

Address for correspondence:
E-mail: termens@ub.edu

DOI: The "Big Brother" in the dissemination of scientific documentation

Digital identifiers

Rapid growth in the availability and use of digital documents has prompted the development of instruments to handle them. A most important example of these instruments are digital identifiers, which provide a codification system that allows digital items, usually up to the level of a computer file, to be singled out and located. Digital identifiers make up standardized global systems applied to specific products or areas. They are part of the very many identifiers developed to handle large numbers of items and large amounts of information for transactional purposes, which often have a global span. Digital identifiers include the ubiquitous Global Trade Item Number (GTIN), a code that unequivocally identifies trade items all around the world. The GTIN can take on several configurations depending on its application. These include: EAN-13, EAN-8, EAN-14, and UCC-12. EAN-13 is the code used for retail products in order to facilitate trade at the point of sale; its widely known symbol or graphical form is the EAN/UPC-13 bar code [1].

In the area of information science, identifiers or codes to process informational data were developed well before the emergence of digital documents [2]. Examples of these include the International Standard Book Number (ISBN), the International Standard Serial Number (ISSN), and the legal deposit number and copyright registration number. The ISBN is a unique global code assigned to any printed book. Currently, the ISBN has two basic applications: (1) in the book trade, where it is widely used in publishers' catalogs and in ordering by book stores, and (2) in the bibliographic tracking carried out by libraries. The International ISBN Agency [3] coordinates management of the ISBN, which is carried out locally in each country by an ISBN national agency. The agency, in turn, is either directly dependent on the government or is commissioned to a private firm. The function of the ISSN is similar to that of the ISBN, although it applies to serial publications—especially to journals—and is administered by the ISSN International Center [4], which is based in Paris and belongs to UNESCO, and by national agencies. The legal deposit number or copyright registration number, is not a universal system; on the contrary, it is established by the particular country in accordance with its local laws. The system was initially devised to provide governments with tools to make sure that book publishers complied with intellectual property laws, but it was used to carry out state censorship of printers. Nowadays, its main use is as a method to ensure that the state gets a copy of every book printed or published inside its territory to be kept in a national library.

The proliferation of digital documents motivated the search for new solutions regarding their management. The existing identifiers, which were devised to be used on physical objects, were found to be inadequate for administering

digital intangible objects. Furthermore, the intrinsic characteristics of the electronic medium, including the new ways of using documents or the potential for computer manipulation of the codes, also made it clear that a new approach was needed. The solution that was reached encompassed three categories: (a) development of already existing codes—the most successful is the Serial Item and Contribution Identifier (SICI), which applies, among other items, to journal articles [5]; (b) creation of new codes for new digital items, such as the International Standard Recording Code (ISRC) for recorded music [6]; and (c) creation of new global identification systems —including the DOI.

The Digital Object Identifier (DOI)

The digital object identifier (DOI) [7] is an identification code with the following properties: it applies to digital objects, mostly; it does not focus on a single class of documents; its scope is global, with a unique numbering system; its management is highly centralized; its regulations and maintenance do not depend on government agencies, rather on private companies; and it has several operative applications. In brief, DOI may be defined as an identifier that published documents can include, voluntarily, whatever their nature or material, for the purpose of being handled in an automated manner. The DOI system is not meant to replace earlier systems, mainly the ISBN, ISSN, and SICI, but instead tries to cover loopholes in those systems and to devote the full extent of its capabilities to establishing an automated management environment and to handling digital documents.

At a technical level, the system is built upon four components. The first is a standardized numbering system, in which a unique identifier is allocated to each document with the granularity level preferred by the publisher. The identifier consists of two parts, a prefix and a suffix, separated by a slash. The prefix contains the number identifying the document's publisher. The suffix identifies the specific document, among those produced by a particular publisher; it is allocated by the publisher and its format is not standardized, although the use of widely accepted codes in the field is recommended, such as the ISBN for books or the SICI for journal articles. The second component is a resolution system based on the Handle System developed by the US Corporation for National Research Initiatives. The system retrieves a document's current URL from its DOI code or from metadata describing it. In the third component, a data model and a data dictionary enable the use of several metadata schemas for describing documents. The fourth component is an implementation engine, based organizationally on the International DOI Foundation and DOI Registration Agencies, and technically on a cluster of interlinked databases.

DOIs can identify, describe, and resolve the location of a digital resource. As a result, its capabilities are greater than those of most identifiers, which are limited to, as their name suggests, enabling correct identification of the object referred to by them. When a publisher allocates a DOI code, it also needs to submit to the main database metadata describing the document, which can be used by third parties to refer to it. The publisher also provides a current working URL so that the metadata are kept up-to-date, thus allowing the resolution system use the DOI code to supply the document's internet location to third parties.

The DOI system is managed by the International DOI Foundation (IDF), which was created in 1998 as a non-profit organization founded by several of the major publishing groups. The foundation is in charge of the technical details, but not of the allocation procedures, which are instead managed by the DOI Registration Agencies (DRAs). These are independent organizations, public or private, that have been authorized to confer the prefix to those publishers who have requested them and to maintain the DOI codes allocated by them. It is worth mentioning several features of this model: DRAs are not agencies within a national context, unlike those managing the ISBN and ISSN, and they will always be fewer in number, implying a highly centralized system. Governments do not fund DRAs and, consequently, their services must be paid. In addition, the DRAs' tasks are not restricted to data recording—they also manage all services related to it (provision access to Handle System, etc.). At the time of this writing, there were only seven authorized DRAs [8].

The principle DOI applications are reference-linking, persistent URL generation, intellectual property handling, and bibliographic referencing. Each one of these applications is assigned a specific conformation according to the type of document or operation it is used for. Since DRAs are not organized along territorial criteria, they have tended to instead specialize in publishing certain subjects and in providing associated well-defined services. Thus, for example, the DRA CrossRef focuses on DOI management and on services related to science journals.

CrossRef

CrossRef was the first DRA authorized and, in fact, the aims of that organization were the *raisons d'être* for the DOI system. At the beginning of 2000, the Publishers International Linking Association, Inc. (PILA), which is the CrossRef sys-

tem provider, was incorporated as a non-profit organization. Currently, the directorship is composed of members representing AAAS (Science), AIP, ACM, APA, Blackwell Publishers, Elsevier Science, IEEE, Wolters Kluwer, Nature, University of Chicago Press, Sage, Springer, Taylor & Francis, Thieme, and Wiley. Over 1400 publishers have joined CrossRef. This is a very large proportion of the science-journal publishing sector, and, in particular, includes those that are business-oriented with an international focus. In other words, CrossRef is the platform created by the main publishers of digital-science journals to allocate DOI codes, and, through them, to automate the linking of bibliographic references using hypertext between documents [9]. This is known as reference linking, a task not achievable by non-machine processes, given the large amount of manual work required, such as: locating footnote indexes and entries in the reference section at the end of an article, processing the journal references (different journals have established a plethora of citation guidelines), and then determining whether the referenced resource is online, checking the reference (volume, issue number, article), recording the URL, and, finally, creating the hyperlink within the original resource that connects to the URL.

CrossRef simplifies and automates this process: the member publishers are in charge of allocating DOI codes to their articles; filling in the form in the DOI system (something that can be automated) with the relevant metadata, such as author, article title, journal title, and the URL; and keeping that information up-to-date in case of changes. In exchange, publishers are able to retrieve data from the main database in an automated fashion. For larger and medium-sized publishers, a computer application takes care of creating the hyperlinks; it scans the text of a new article, extracts references to other articles, interprets them and initiates a search in the Handle System. In the Handle System, the search keywords are compared to metadata that point at a DOI code and with this, to a working URL. The result is then inserted into the new article. The Handle System offers an additional improvement—the working URL is always operative and does not require maintenance. URL obsolescence, and the consequently large number of broken links, is one of the major problems in present-day internet structure and it slows down the allocation of links between websites [10]. Handle System provides a solution because, for each, document, it generates a persistent address, a URN, which can be used to locate it. DOI codes can operate as URNs if they are preceded by an address resolution server, e.g., [http://dx.doi.org/10.1126/science.1088234]; hence, the reader does not need to be concerned about a link's operating state as this is attended to by the publisher of the referenced journal by maintaining an up-to-date URL in the DOI databases.

Other fields of application

It should be noted that the extensive use of CrossRef and its two related services, reference linking and persistent URLs, have been key factors in the development of electronic science journals and the readers' readiness to accept them. The applications of the DOI do not end here, although they are the most obvious outcomes. The system can also regulate access permission and users' utilization of the documents, as well as maintain centralized statistics regarding usage. In a broader sense, several services focusing on intellectual property rights, including financial aspects, are already being offered. Other uses are still in development, for example, the German National Library of Science and Technology (TIB) is testing the allocation of DOIs to primary scientific data, such as weather tables and satellite images [11].

At any rate, it seems clear that the proliferation of DOIs will be inextricably linked to the needs and interests of the publishing industry, which was the driving force in their creation. Hence, whereas major results have been obtained concerning the interconnection and production of corporate applications within this sector, less success has been achieved by the other parties involved, such as small publishers or libraries. The small or non-profit publishers have found that the requirements, technical and financial (as all DOI services require payment) of participation in the system have restricted the potential benefits of the DOI in circulating their products. Moreover, libraries have encountered the *appropriate copy problem*—the DOI system supplies the referenced document's URL, but that URL might be in a service not contracted for by that particular library. In short, DOI links comprise a *non-context-sensitive* closed system. To solve this problem, solutions such as the SFX resolution system and the OpenURL standard have been developed [12].

Finally, two other aspects must be emphasized. The DOI was created and it developed in response to the requirements of electronic-journal publishing; however, it is also applicable to other documental activities, such as e-learning systems. While the DOI currently has no impact whatsoever on e-book systems, this might be due to the difficulties that this particular market is experiencing. Furthermore, there is some distrust regarding how the identifier might be used in the future. For example, for the first time in history, the international scientific community's use of the journal *Science* is being recorded, each time an article is read, in a centralized database, which, perhaps even more controversial, is privately maintained by commercial companies with interests in the field. This framework enables a citation link analysis to be carried out that provides a better understanding of how avail-

able publications are used by scientists, as indicated by the steady processing of this information by ISI-Thomson Scientific (Philadelphia, USA). But it can also become a tool to modify the manner in which science is published.

Notes

1. The organization in charge of these standards is GS1. [http://www.ean-int.org/]
2. Vitiello G (2004) Identifiers and identification systems: an informational look at policies and roles from a library perspective. D-Lib Magazine 1. [http://www.dlib.org/dlib/january04/vitiello/01vitiello.html]
3. http://www.isbn-international.org/
4. http://www.issn.org/
5. http://www.niso.org/standards/resources/Z39-56.pdf
6. http://www.ifpi.org/isrc/
7. http://www.doi.org/
8. http://www.doi.org/registration_agencies.html
9. Brand A (2001) CrossRef turns one. D-Lib Magazine. 5. [http://www.dlib.org/dlib/may01/brand/05brand.html.] Blake ME, Knudson, FL (2002) Metadata and reference linking. Library Collections, Acquisitions, & Technical Services 26:219–230
10. A recent study of over 1000 articles published between 2000 and 2003 in the *New England Journal of Medicine*, *The Journal of the American Medical Association* (JAMA), and *Science* found that 13% of internet references were broken links (Dellavalle RP (2003). Going, going, gone: lost internet references. Science 5646:787-788). A previous study pointed out that the average life of URLs cited in computer-science articles was 4 years from the publication date (Spinellis D [2003] The decay and failures of URL references. Communications of the ACM 46:71-77. [http://www.spinellis.gr/sw/url-decay/])
11. Paskin N (2005) Digital object identifiers for scientific data. Data Science Journal 4:12-20. [http://journals.eecs.qub.ac.uk/codata/Journal/contents/4_05/4_05pdfs/DS392.pdf]
12. Van de Sompel H, Oren Beit-Arie (2001) Open linking in the scholarly information environment using the OpenURL framework. D-Lib Magazine 3 [http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html]