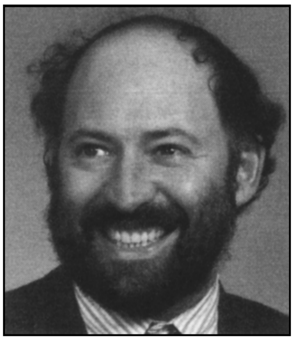
THE DOI (DIGITAL OBJECT IDENTIFIER)

Cliff Morgan

The Digital Object Identifier (DOI) was opened to the public in October 1997. It provides a regulated means of identifying 'objects' so they can be accessed and traded even with changes of ownership and location. The article describes the background to the DOI's creation, its current status, and future developments. This article is an expanded and updated version of a presentation to the Wiley Library Advisory Board, 18 November 1997.



Cliff Morgan is Publishing Technologies Director,
John Wiley & Sons Ltd,
Baffins Lane, Chichester,
West Sussex PO19 1UD
E-mail: cmorgan@wiley.co.uk

Introduction

It has been called "an electronic licence plate"[1], "an electronic sheriff"[2], "one of the most important events in publishing this century"[3], "a Dis-Organized Idea"[4], and "the 21st Century ISBN"[5]. It is unique, simple, persistent, sexy, and misunderstood. It is the Digital Object Identifier (DOI). There are many excellent articles both in print and on the Web about the DOI and its role in electronic commerce or information transfer. This article is a beginner's guide to the DOI - what it is, why it is important, how it works, how it fits in, and what is still being discussed.

What is a DOI?

A DOI is a Digital Object Identifier. Books are identified by ISBNs, cars by licence plates, and digital objects by DOIs. A 'digital object' could be the electronic form of a whole book or journal; or an individual issue, chapter or article; or an individual abstract, figure, table, chemical structure, reference, etc.; or (more contentiously) an order form, registration form, or piece of information. That is, it can be applied to more or less any discrete particle of electronic content, at any hierarchical level.

What are its characteristics?

A DOI is *unique, persistent* and *dumb.* Each digital object has its own individual number. This number never changes; and the number does not (or need not) tell you anything about the object itself.

A DOI consists of a publisher ID* (prefix) and an item ID (suffix), separated by a forward slash (/). For example, the DOI for one of Wiley's journal articles would look like this:

10.1002/0002-8231(199601)47:1<23:TDOMII>2.0.TX;2-2

The abstract for this article would be:

10.1002/0002-8231(199601)47:1<23:TDOMII>2.3.TX;2-U

* More precisely, a Registrant ID. Throughout this article, I have used 'publisher' rather than the more generic 'registrant', 'rights owner' or 'information provider'. This is because publishers are likely to form the bulk of participants in the initial roll-out.

These may look like horrendous numbers, but they really do have to be quite long to guarantee uniqueness. The publisher ID is the '10.1002' before the slash. (The '10' identifies the assigning body and '1002' the publisher.) The item ID, in this case, is simply the SICI (Serial Item and Contribution Identifier), a standard way of identifying journal articles and abstracts. You will see that the article and the abstract SICI only differ in two places: the number before 'TX' changes from '0' to '3' and the check digit at the end is thus recalculated. We are able to generate SICIs, and thus DOIs, as part of our regular production process. For book titles, we will use the book equivalent, the BICI, which is currently in Draft Standard form.

A publisher does not have to use the SICI. Any alphanumeric system can be used, as long as each object can be uniquely identifed.

It is important to realise that, although the DOI may have intelligent components (such as the SICI), as a permanent number it is dumb. For example, if the object were to be transferred to another publisher, the publisher ID would *not* change. The principle of persistence means that the above article, for example, *always* has the DOI 10.1002/..., whether Wiley still owns it or not.

Why is the DOI important?

The first step in finding and trading digital objects is to agree how to name them. The DOI mechanism also sets up a regulatory system for coordinating, managing, and ensuring compliance (policing). The ultimate aim is the automated negotiation, payment and delivery of digital objects.

The DOI is not in itself a copyright management or authentication system, but it establishes the naming framework for one.

How does it work?

The publisher applies to the DOI Agency for the publisher ID. Registration has a fixed cost, currently $1000, which may be expensive for small publishers. It has been suggested that such publishers may club together and be assigned umbrella IDs, or they may be sponsored by larger publishers or publishers' associations.

The publisher assigns item IDs to objects at the appropriate level (e.g. journal issue, article, abstract). A file of DOIs and their associated database addresses, e.g. URLs**, is sent to the DOI directory for validating and storing. Figure 1 shows a batch of DOIs with their URLs.

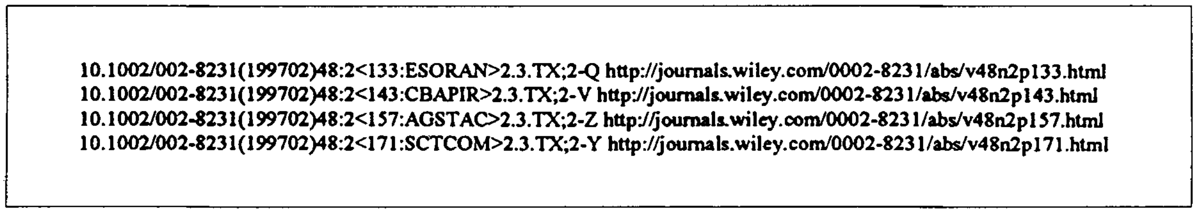
The publisher is charged per DOI. The current rate is 1 US cent.

The directory is designed to be a distributed system, so material may be sent to local directories, which could be organized on an industrial or geographical basis. Some publishers may even run their own directories under licence. All local directories will be linked to the Central Directory.

The publisher is obliged to inform the directory manager of any changes to the URLs, so that the matched DOI/URL pair is always current and valid. The directory regularly checks URLs for currency, and lets the publisher know if there are broken links.

It is this regulatory aspect that distinguishes the DOI mechanism from the more laissez-faire approach of just expecting people to keep their URLs up to date!

** The DOI system is not limited to dealing with Web addresses only - it will cope with any database location. I refer to URLs throughout the article because they are the most common current application.



```
10.1002/002-8231(199702)48:2<133:ESORAN>2.3.TX;2-Q http://journals.wiley.com/0002-8231/abs/v48n2p133.html
10.1002/002-8231(199702)48:2<143:CBAPIR>2.3.TX;2-V http://journals.wiley.com/0002-8231/abs/v48n2p143.html
10.1002/002-8231(199702)48:2<157:AGSTAC>2.3.TX;2-Z http://journals.wiley.com/0002-8231/abs/v48n2p157.html
10.1002/002-8231(199702)48:2<171:SCTCOM>2.3.TX;2-Y http://journals.wiley.com/0002-8231/abs/v48n2p171.html
```

Figure 1 Excerpt from a Wiley DOI batch file, which is sent to the DOI Directory. This particular file consists of DOIs for journal article abstracts and the associated URLs. The DOIs never change, but the URLs can be replaced by more up-to-date locations, so the DOI always takes the user to the current address.

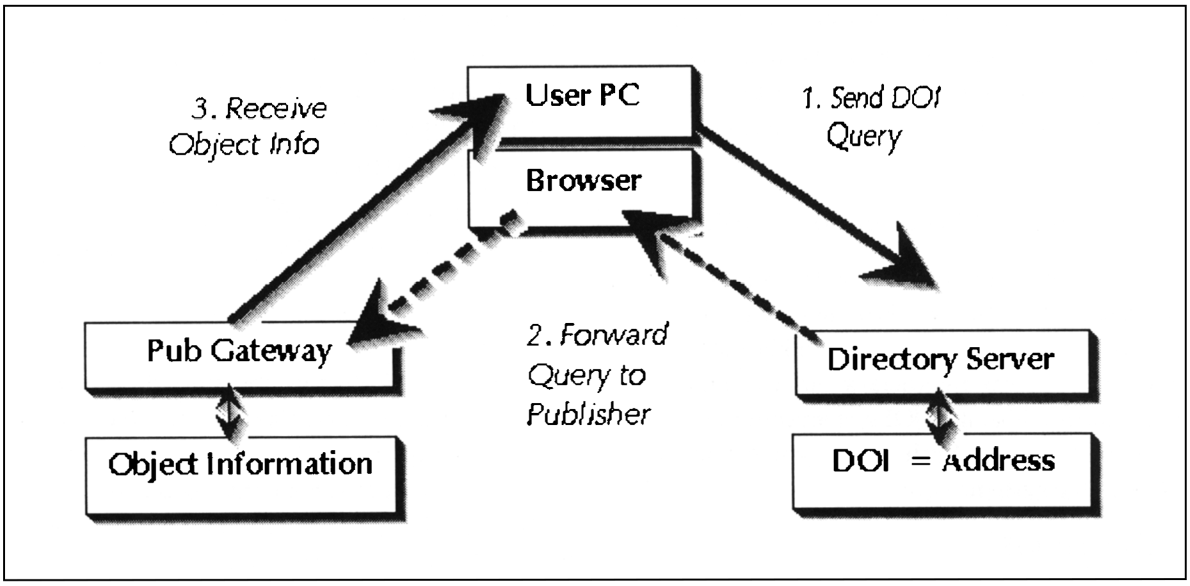


Figure 2 The DOI system. Activating the DOI button (or link) takes the user to the directory, which sends the current URL to the user's browser, for onward directing to the publisher. The publisher provides the response screen to the user. (Reproduced from the Website http://www.doi.org/system_spec.html, by permission of the International DOI Foundation)

Users would not normally see the actual DOI number. It may be represented by a button or icon, or be embedded in the text or graphic as a hyperlink. The user may select the link, or it may be selected automatically by some interactive program. A DOI query goes to the directory, which sends the current URL to the user's browser. The user is then redirected to the publisher, who provides the appropriate 'response screen'. To the user, the routing process is transparent - it appears as though the link is direct to the publisher. (See Figure 2.)

The publisher is thus in control of what the user sees. The most common uses are:

1. to provide *information* about the object, e.g. rights;
2. to *display* or *deliver* the object, e.g link from reference citation to full reference details, to abstract or full text of cited work;
3. to link to *related material*, e.g other works by the same author, supplementary material;
4. to *initiate a transaction*, e.g. subscribe to a journal, register for a conference, buy permission rights.

Work to date

The DOI was an AAP (American Association of Publishers) initiative. In March 1996, they requested bids from technology partners to design, administer, maintain and support the DOI system. In September 1996, it was announced that the Corporation for National Research Initiatives (CNRI) had been selected, using its 'handle' technology for matching DOIs to database addresses. R. R. Bowker were also initially involved in the conceptualising of the registration and directory services.

The DOI was officially launched at the Frankfurt Book Fair in October 1997. It was phenomenally well received by the industry (see the reports referred to in the introduction). It has been warmly endorsed by the IPA (International Publishers Association) and STM (International Association of Scientific, Technical and Medical Publishers).

The International DOI Foundation was set up to administer registration and directory services. This non-profit organization sets policies, selects service providers, and oversees the successful operation of the whole scheme. There are currently offices at Washington DC (crisher@publishers.org) and Geneva (doi@worldcom.ch). It is envisaged that the Foundation will consist of publishers, technology companies, Government agencies, and associations representing authors, librarians and intellectual property rights.

DOIs are already in use by a dozen or so US and European publishers (with about 250,000 created to date), and phase two of the prototype involves the testing out of various business models.

How does the DOI fit in with other standards?

The most detailed discussion on this topic can be found in Paskin[6]. Standards bodies are reviewing the DOI, but it is not yet clear whether the DOI will in fact become an official standard, and if so, when. Meanwhile, the DOI system continues to be developed by the publishers and others who are active in the system.

ISSNs and ISBNS

ISSNs and ISBNs do not identify objects below book and journal level so they are not in themselves sufficient for publishers' purposes. However, they can be used as component pieces of the DOI since they are well-established numbering systems.

SICIs and BICIs

The SICI is an established standard, and can identify down to article and abstract level, but it does not contain any publisher or rights owner identification. The BICI is still in draft form. As with ISSNs and ISBNs, these can form part of the DOI (in the same way that ISSNs and ISBNs are themselves component pieces of SICIs and BICIs).

PIIs

The PII (Publisher Item Identifier) is not a formal standard but it is used by a number of publishers. It was specifically designed to be assigned to articles in the early stages of production, before the volume, issue or pages (or even year of publication) are known. The SICI can encompass the PII, as can the DOI.

ISWCs

The International Standard Work Code is a proposed ISO standard for *all* works, e.g. musical as well as literary, but the current timetable envisages implementation in the year 2000. The DOI will be able to use this number too.

ISDIs

NISO (the US National Information Standards Organization) have proposed an International Standard Document Identifier, which would bring more regulation to the item ID by allowing reference to any specific numbering system such as SICI being used in the item ID, e.g.:

10.1002/[SICI]0002-8231(199601)47:1
<23:TDOMII>2.0.TX;2-2

This has no formal status yet, but clearly it does not conflict with the DOI so much as expand upon it.

URLs, PURLs, URNs

URLs are intrinsically unstable since they only denote the *location* of an object on the Web, which can change. Setting up aliases to reroute can be done, but this is not sufficiently regulated.

Persistent URLs (PURLs) and Uniform Resource Names (URNs) have both been proposed to overcome the URL problem, but they are still Web-oriented, and based on the naming and addressing protocols that apply here. The DOI is not limited to any one electronic environment, and it could be used as part of a URN. (CNRI handles are also URN-friendly, but not limited to the Web.)

Outstanding issues

Many of these issues are discussed more fully in Lynch[7], Paskin[8], Bide[9], Simmonds[5] and Berinstein[10], for example. These are all well worth reading to get a fuller flavour of current debates.

1. What is the *capacity* of the system? Will it *scale up* to cope with the enormous amount of material that, in principle, could be identified to a very granular level?
2. How do we deal with *legacy* material? It would be possible to assign DOIs to archival material, but how far back do we go?
3. How do we deal with *manifestations* and *versions*? A manifestation is the same content in another form (e.g. hardback and paperback, or TIFF and PDF); a version is different content (e.g. edited, annotated, or translated). Should separate DOIs be assigned for each manifestation and for each version?
4. Persistence of identification need not imply persistence of access. There will come a point when the object is no longer being made available (the electronic equivalent of being put out of print) - how will this be managed? Will the directory be purged or does it continue growing?

5. Even with regulation, the system still relies on publishers keeping the directory informed, or on them replying to prompts from the directory. If compliance is enforced through punitive measures (such as expulsion from the scheme or fines), would there then need to be an appeals procedure? In a litigious culture, this may be a real problem.
6. The DOI has been introduced as a technology without a business model to match. However, this seems to me to be inevitable when one is dealing with enabling technologies. As mentioned above, the business models are being prototyped, and it is clear that, for commercial organizations, these will dictate the eventual uses of the DOI. At the time of writing, issues concerning operational governance, funding and commercial control of the system are still being finalized.
7. The DOI will need to be integrated into rights/ permissions management, metadata standards, search-agent software, and secure transaction processing systems. It is planned that a transactions syntax will be developed by the end of this year.

Conclusion

The DOI is clearly a significant step forward in enabling electronic commerce or information transfer. Although there are still many areas requiring further discussion, the DOI story has been remarkable for its ready acceptance in a community not renowned for its collaborative nature. With such goodwill and commitment to make it work, this cross-industry initiative looks very much like an idea whose time has come.

Further reading

The Web site is excellent. The Gallery area has examples of the DOI in action, with contributions from Academic Press, the ALCS (Authors' Licensing and Collecting Society), Copyright Clearance Center, Elsevier, Houghton-Mifflin, the IPA, Scholastic, Shepard's, Springer-Verlag, and Wiley. The site can be found at http: //www.doi.org. (Please note: not .com, which is an adult site, causing much speculation as to what DOI could possible stand for in these circumstances!)

References

1. Anon., Coming soon: license plates for digital content, *The Seybold Report on Internet Publishing,* March 1997 **1** (7), 28-29
2. Carvajal, D., An electronic sheriff to battle book rustling, *New York Times,* 22 September 1997.
3. Goetze, D., quoted in: Carvajal, D., Electronic branding praised at Frankfurt Book Fair, *CyberTimes* (the New York Times on the Web), 20 October 1997.
4. Leer, A., quoted in Taylor, S., DOI: Digital Object Identifier or Dis-Organized Idea?, *Publishers Weekly,* 10 November 1997.
5. Simmonds, A., The 21st Century ISBN, *The Bookseller,* 5 December 1997, 20-22.
6. Paskin, N., Information identifiers, *Learned Publishing,* April 1997, **10** (2), 136-156. Available at http: //www.elsevier.nl/inca/homepage/ about/infoident.
7. Lynch, C., Identifiers and their role in networked information applications, *ARL: A Bimonthly Newsletter of Research Library Issues and Actions,* no. 194, October 1997. The Web version (http: // www. arl.org/newsltr/194/identifier.html) also has a very interesting response from Bill Arms of the CNRI.
8. Paskin, N., Digital information objects and the STM publisher, *STM Annual Report 1997,.* Available at http: //www.elsevier.nl/inca/ homepage/about/diginfo.
9. Bide, M., In search of the unicorn: the Digital Object Identifier from a user perspective. *British National Bibliography Research Fund Report,* November 1997. Available from BIC, 39-41 North Road, London N7 9DR, or at http: // www.bic.org.uk/bic/unicorn2.pdf.
10. Berinstein, P., DOI: a new identifier for journal content, *Searcher,* **6** (1), 72. Available at http: // infotoday.com/searcher/jan/story4.htm.