I originally wrote this column last December but it took so long to get scheduled for publication, I had to update it and resubmit it for a later issue. That tells us two things: the lead time for e-journals is still longer than you'd like; and things move fast in the world of the DOI.

DOI stands for Digital Object Identifier, which isn't just an identifier but also an entire system for assigning, maintaining, resolving, and using persistent identifiers. (Since this carries obvious potential for semantic confusion, I'll try to be careful to distinguish between the "DOI system" and the "DOI identifier.")) The system was originally developed for the Association of American Publishers by R.R. Bowker and the Corporation for National Research Initiatives (CNRI), but now is managed by the International DOI Foundation, a nonprofit membership organization based in New York and Geneva. The intent is to facilitate digital commerce by maintaining persistent links to the rights holder of a digital object.

One of the pesky things about URLs is that if you move a document, the URL changes. If you have the URL embedded in hundreds of references in Web pages or catalog records, you have to change all of those references or else incur the dreaded "404 File not found" message on your browser. However, if you use some other identifier in all of those references, and that identifier takes you to a directory that maps from the identifier to the URL of the document, the problem is greatly reduced. Every time you move the document you only need to locate and change the single directory entry, not every reference to it.

All schemes for providing persistence, from PURLs to URNs, are based on this idea of mapping from arbitrary identifiers to true locations. The mapping is called "indirection," and the act of getting other information (like a URL) in exchange for an identifier is called "resolution." (See the glossary below for acronym identification.)

I Want to Hold Your Hand(le)

As currently implemented, the DOI relies on CNRI's Handle System software to maintain the directory and provide resolution. Essentially, if you hand the Handle System an identifier, it will hand you back something else. In the DOI system what you get back is up to the publisher who registered the identifier. It might be the URL of the object, or of an order form for the object, or of a screen of copyright information. This can be confusing because you have to distinguish between what the DOI identifier refers to (the object itself) and what is returned in response to a query, which could in theory be all sorts of other
The identifier itself is an alphanumeric string with a prefix and a suffix, separated by a slash. The prefix has two elements separated by a dot. The first element identifies the "Directory Manager" or naming authority; the second identifies the publisher or agent responsible for the suffix. The suffix is essentially arbitrary so long as it is unique, although a standard for formatting suffixes is being developed. In the example given on the DOI home page, the full identifier is "10.1002/[ISBN]0-471-58064-3." The Directory Manager is "10," the agent assigning the suffix is "1002," and the suffix itself is ":[ISBN]0-471-58064-3." Although the original intent was to implement a distributed directory system, current plans call for a single Directory Manager. Negotiations are in progress for ISBN International to take on this role, which includes running the directory and overseeing the distribution of prefixes.

The DOI initiative, of course, is intended to do more than simply provide persistence. If that was all they wanted, publishers could have implemented a PURL server with a lot less trouble. The International DOI Foundation hopes to build a comprehensive system for managing permissions and has working groups actively addressing several aspects of this, including policy, applications, descriptive metadata, and metadata for rights management.

So don't drop everything and start assigning DOI identifiers to all the documents on your Web server: this is very much an application for publishers and rights holders. You have to register with the International DOI Foundation before you can request a prefix; you must agree to terms and conditions such as only assigning DOIs to objects for which you have electronic rights and which reside on servers under your control. It is also not free. The prefix itself costs $1,000 US dollars, with an additional annual fee based on the number of identifiers registered to that prefix in the DOI system.

The Importance of Being URNest

So, is the DOI identifier a URN? Good question! The Internet Engineering Task Force's URN Working Group has defined an architecture for name resolution and a set of minimum syntactical requirements for an identifier. Beyond that, it is up to various communities using the Internet to define identifiers within their own namespaces. Clifford Lynch and others have shown that--if appropriately prefixed--ISBNs, ISSNIs, SICIs, and other standard bibliographic identifiers can fit within the URN framework. (See "Finding More Information" below.) Presumably, the DOI identifier can, too. Most commentators consider the DOI system an implementation of the URN, but some members of the URN Working Group are uncertain whether the underlying Handle System is fully conformant. It would be nice if somebody who understood both Handles and URNs would do this evaluation and let us all know.

Is the DOI a standard identifier? In the sense of being an official standard of a national or international standards organization, like the ISSN, ISBN or SICI, it isn't. However, a NISO standards committee is being established to define a syntax.
for the DOI, so this may ultimately become an American National Standard. If it does, though, it will only apply to DOI identifiers within the DOI system, not to all identifiers for digital objects.

Why not just use existing standard identifiers to begin with? Another good question. The DOI syntax will probably encourage use of SICIs and other standard numbers within the DOI string when applicable. But current standard numbers won't work for all material. For one thing, they may not be applicable at the necessary level of granularity. You can give a DOI to a work (e.g., an article), or a portion of a work (e.g., a photograph in an article), or an aggregation (e.g., an issue)--any object for which you might want to control permissions separately. For another, there may be a need to control rights to objects that are not yet published or otherwise don't qualify for other standard numbers.

So, what's all the hoopla about? If you've already heard of the DOI, you probably know it generates a lot of interesting discussions. There are a few things, I think, of some concern--not a weeping-and-wailing-and-gnashing-of-teeth concern, but maybe worth a furrowed brow.

First, as a community of Internet users we would hope that the first widespread and well-funded implementation of distributed name resolution is compatible with URN architecture and principles. We need to ascertain if the DOI system is compatible or can be brought into compatibility. Second, unlike SICI and BICI identifiers, DOI identifiers cannot be derived from any bibliographic information about the piece. You will never know a DOI identifier unless the publisher tells you. This shouldn't be a problem in most respects as publishers will want you to get to rights information. However, identifiers are valuable to third party abstracting and indexing services also, and it is not inconceivable that independent database publishers could have to pay for, or even be denied access to, the identifiers. That could make things harder (not to mention more expensive) for all of us who are striving for more open information systems.

Third, every model of the DOI system I've seen assumes a very limited universe where there is only one copy of any given object and the object, its metadata, and the services related to it are all controlled by the publisher. This doesn't much resemble the world I live in, where I might get an article from a publisher and you might get it from UMI or OCLC, and cousin Fred gets it from his local library server. I'm assuming the model fails to address this complexity just to get off the ground, not because the designers want to move to a more closed environment.

Suppose by Any Other Name

Finally, there's a matter of misplaced expectations. I believe this is largely due to the name "DOI" itself. If this were called the "Publishers' Rights Management Identifier" we could focus on the meaning of this application. Unfortunately, the term Digital Object Identifier is so generic, people can't help but assume that it should meet their needs for, well, a digital object identifier.
If you're digitizing source material from your archives or hosting a Web site for research papers, this DOI is probably not for you. So my suggestion is, let's start calling this the "Publishers' Digital Object Identifier" when we talk and write and hold programs about it.

Publishers have a right to design systems and identifiers to meet their needs and, far from criticizing them, we should be applauding their initiative. We should, however, show an equal level of commitment to the development of a broadly applicable and open identifier system that meets the needs of the emerging national digital library. Why don't we get together in the library community and come up with our own system and standards for persistent naming and resolution--a true International Standard Digital Identifier?

+ Page 24 +

Finding More Information

There is a huge amount of literature about identifiers in general, and the DOI in particular. Here are a select few:


Acronyms

My editors always want me to spell out acronyms in parentheses after the reference. I tend to feel that technical acronyms are like street names in Boston—if you don't know what they are, the name won't help. So as a compromise, here's a glossary for this column.

ANSI: American National Standards Institute, the national
clearinghouse for voluntary standards development in the United
States.
BICI: Book Item and Contribution Identifier, a standard in
development by NISO.
IETF: Internet Engineering Task Force, the protocol engineering
and development arm of the Internet.
ISBN: Intentional Standard Book Number; they won't sell anything
at WaldenBooks without one.
ISSN: International Standard Serial Number.

+ Page 25 +

NISO: National Information Standards Organization, the ANSI
standards organization that deals with libraries, publishers, and
information services.
PURL: Persistent Uniform Resource Locator, a URL redirected by a
PURL server, a software package developed by OCLC.
SICI: Serials Item and Contribution Identifier, NISO Z39.56, a
standard identifier for issues and components of issues like
articles.
URL: Uniform Resource Locator, the information your browser needs
to get to a resource.
URN: Uniform Resource Name, a system for naming and name
resolution being defined by the IETF.

About the Author

Priscilla Caplan, Assistant Director for Library Systems,
University of Chicago Library, 1100 E. 57th Street Chicago, IL
60637. Internet: p-caplan@uchicago.edu.

About the Journal

The World Wide Web home page for The Public-Access Computer
Systems Review provides detailed information about the journal
and access to all article files:

  <URL:http://info.lib.uh.edu/pacsrev.html>

Copyright

This article is Copyright (C) 1998 by Priscilla Caplan. All
Rights Reserved.

The Public-Access Computer Systems Review is Copyright (C) 1998
by the University Libraries, University of Houston. All Rights
Reserved.

Copying is permitted for noncommercial, educational use by
academic computer centers, individual scholars, and libraries.
This message must appear on all copied material. All commercial
use requires permission.