Identifiers and Their Role In Networked Information Applications

by Clifford Lynch, Executive Director, Coalition for Networked Information

Editor's Note: *When the* Association of American Publishers' *proposed Digital Object Identifier system reached a level of news that was worthy of a headline in the New York Times (First Business Page, Sept. 22, 1997), ARL turned to CNI for this reality check on the new, high profile role of identifiers.*

Identifiers are an enormously powerful tool for communication within and between communities. For example, the International Standard Book Number (ISBN) has played a central role in facilitating business communications between booksellers and publishers; it has also been important to libraries in identifying materials. The International Standard Serial Number (ISSN) plays a pivotal role in facilitating commerce among publishers, libraries, and serials jobbers; it is also vital to libraries in managing their own internal processes, such as serials check-in. Bibliographic utility identifier numbers such as the OCLC or RLIN numbers are used in duplicate detection and consolidation in the construction of online union catalog databases.

The traditional bibliographic citation can be viewed as an identifier of sorts, albeit one that is not rigorously defined; it has many variations in style, and data elements based on editorial policies. Yet the ability to cite is central both to the construction of the record of discourse for our civilization and to the development of scholarship; the citation plays an essential role in allowing authors to reference other works, and in permitting readers to locate these works.

The assignment of identifiers to works is a very powerful act; it states that, within a given intellectual framework, two instances of a work that have been assigned the same identifier are the same, while two instances of a work with different identifiers are distinct. The use of identifiers outside of their framework of assignment, though, is often problematic. For example, normal practice assigns a paperback edition of a book one ISBN and the hardcover edition another, so bookstores can distinguish between these versions, which usually vary in price and availability. But ISBNs are also used sometimes in bibliographic citations; in this situation, when the content and pagination of the hardcover and paperback editions are identical, either will serve equally well for a reader tracking down a citation, and the inclusion of an ISBN as an identifier for the cited work may actually cause problems because it is making an unnecessary distinction (for this purpose) among versions of the same work.

A great deal of scholarship involves the development of identifier systems that allow scholars to name things in a way which makes distinctions and recognizes logical equivalence--ways of identifying editions of major authors or composers, variations in coinage having numismatic significance, or the identification of chemicals, proteins, or biological species. Often the rules for assigning identifiers to objects are the subject of ongoing scholarly debate and form a key part of the intellectual framework for a field of study.

Identifiers take on a new significance in the networked environment. To the extent that a computational process can allow a user to move from the occurrence of an identifier to accessing the object being identified, identifiers become actionable. For example, in the World Wide Web links can be constructed between the entries in an article's bibliography and digital versions of the cited works, links that can be traversed with a mouse-click. The significance of making a citation actionable is so great that it has been the subject of several recent lawsuits--for example, the litigation between Microsoft and Ticketmaster about the inclusion of links to Ticketmaster's web pages in Microsoft's web service over Ticketmaster's objections, which remains pending as of this writing. Another interesting case involved a service on the Web called Totalnews, which included citations and offered access to many other services, "framed" by the Totalnews service. The case was recently settled out of court and failed to establish a precedent.

If one translates these practices under legal challenge, particularly in the Microsoft v. Ticketmaster case, into analogous practices in the print world, one can view this litigation as questioning whether one author remains free to cite the work of another without permission--which is certainly a well established practice in print, and a profoundly important right to lose in the networked environment. Of course, this is just one interpretation of the Microsoft v. Ticketmaster case--it is complicated by a number of commercial factors. Yet it helps to illustrate what is at stake in establishing identifier systems, the control of the use of identifiers, and the practices surrounding them.

In the networked information environment, we have recently seen the emergence of a number of important new identifiers, some of which are relatively mature, and others that are still under development. The remainder of this article briefly discusses a number of these identifiers.

URLs and URNs

Uniform Resource Locators (URLs) are a class of identifiers that became popular with the emergence of the World Wide Web. We first saw them on web pages, later in newspaper advertising and on the sides of buses, and then everywhere; currently they serve as the key links between physical artifacts and content on the Web, as well as providing linkage between objects within the Web.

URLs have clearly been very effective; yet they are unsatisfactory in one very major way. They are really not names, in that they don't specify logical content, but, rather, are merely instructions on how to access an object. URLs include a service name (such as "FTP" for file transfer or "HTTP" for the Web's hypertext transfer protocol) and parameters that are passed to the specified service--most typically a host name and a file name on that host, both of which may be ephemeral. From a long-term perspective, the service name is also ephemeral--for example, content may well outlive a specific service (as has already been the case with the GOPHER service). It is important to recognize that URLs were never intended to be long-lasting names for content; they were designed to be flexible, easily implemented and easily extensible ways to make reference to materials on the Net.

The Internet Engineering Task Force (IETF), which manages standards development for the Internet, realized the limitations of URLs for persistent reference to digital objects several years ago, and as a result began a program to develop a parallel system called Uniform Resource Names (URNs). The IETF URN working group recognized that the URN system must accommodate a multiplicity of naming policies for the assignment of identifiers. Roughly speaking, the syntax of a URN for a digital object is defined as consisting of a naming authority identifier (which is assigned through a central registry) and an object identifier which is assigned by that naming authority to the object in question; the specific content of the identifier may have structure and significance to users familiar with the practices of a given naming authority, but has no predefined meaning within the overall URN framework. Note that the URN syntax does not specify an access service for the object, unlike a URL.

The second key idea in the URN framework is that of resolution services or processes--which may be as complex as new network protocols and infrastructure (analogous to the Domain Name System, for example) or processes as simple as a database lookup--which translate a URN into instructions for accessing the named object. Systems which provide resolution services are called "resolvers"; sometimes the IETF work also refers to "resolution databases" which provide the mapping from names to object locations and access services. URNs are resolved to sets of URLs which provide access to instances of the named digital object. A URN may resolve to more than one URL because there are copies of the digital object that have been replicated at multiple locations such as mirror sites, or because the URN (as defined by the relevant naming authority) specifies the object at a high degree of abstraction, and multiple manifestations of the object (for example, in different formats, such as ASCII, SMGL and PDF) are available. There is no explicit requirement that the URN to URL resolution process expose the mapping from an abstract definition of content to a variety of specific manifestations; it is equally legitimate for the choice of format to be made as part of a prototol negotiation in evaluating a URL when using a sophisticated protocol such as the Z39.50 Information Retrieval Protocol which supports such negotiation. As the location and means of access for objects change, the resolver's database is updated; thus, resolving a URN tomorrow may return a different set of URLs.

Today's standard browsers do not yet understand URNs and how to invoke resolvers to convert them to URLs, but hopefully this support will be forthcoming in the not too distant future. One can reasonably view the URN framework as the means by which both existing and new identifier systems will be moved into the networked environment. The URN framework is intended to be sufficiently flexible to subsume virtually all existing bibliographic identifiers (sometimes referred to as "legacy" identifier systems); for example, the IETF working group documented how the ISSN, ISBN, and SICI might be implemented as URNs.

The IETF uses the term Uniform Resource Identifiers (URIs) as a generic name to cover both URLs and URNs, along with the still immature concept of Uniform Resource Characteristics (URCs), which can be thought of as structures which allow one or more URNs (perhaps from different naming frameworks) to be related both to sets of URLs and to metadata describing the objects identified by the URNs and URLs. The Coalition for Networked Information is active in the IETF standards work on URIs.

The OCLC Persistent URL (PURL)

As a stopgap measure to address some of the problems with the persistence of URLs, about two years ago OCLC deployed a system called the PURL (Persistent URL). Basically, PURLs are HTTP URLs where the usual hostname has been replaced with the host "PURL.ORG" and the filename is an identifier for the "real" content being referenced. The PURL.ORG host will be maintained for the long term by OCLC under that name; when someone registers an object with this PURL server they provide the current hostname and filename for the object and the PURL server creates a database entry linking this hostname and filename to the identifier that will appear in the PURL. When the PURL server is contacted because someone is evaluating a PURL, it looks up the identifier in its database, finds out where the object in question currently resides, and uses the redirect feature of the HTTP protocol to connect the requester to the host housing the object. Content providers are responsible for sending updates to the PURL server when the content file name and/or location changes.

PURLs share the idea of indirection--looking up an identifier in a database to find out where the object is currently stored--with URN resolvers as a means of achieving persistence. They are a very clever and practical design, in that they work with the existing installed base of web browsers. However, they are not truly names, since they only permit content to be accessed through a specific service, namely HTTP. PURLs will probably no longer work as new protocols appear that supersede HTTP, and as content migrates to access through such successor protocols.

The SICI Code and Related Developments

The Serial Item and Contribution Identifier (SICI) code was recently revised by a standards committee under the auspices of the National Information Standards Organization (NISO), the ANSI-accredited standards body serving libraries, publishers, and information service providers; it is described in American National Standard Z39.56-1996. The SICI relies in an essential way on the ISSN to identify the serial, and can be used to identify a specific issue of a serial, or a specific contribution within an issue (such as an article, or the table of contents).

The SICI code is starting to see wide implementation and is likely to serve a central role in a number of applications: it can be used not only to identify articles, but also to link citations from article bibliographies or abstracting and indexing databases to articles in electronic form. It is an important part of the infrastructure that supports ARL' s NAILLD program to streamline interlibrary loan and document delivery. One of the great strengths of the SICI is that it can be determined directly from an issue of a journal (or an article within the issue), assuming only that the ISSN for the journal can be somehow determined. As such, it represents an open standard for creating linkages to articles or other serial components.

Also under NISO auspices, work has just begun on a new identifier with the working name of Book Item and Contribution Identifier (BICI). The BICI can be used to identify specific volumes within a multivolume work, or components such as chapters within a book. There are still a number of unresolved issues surrounding the exact scope of this standardization effort, both in terms of the range of works that it applies to (for example, sound recordings as well as

books) and the level of granularity of the identifier (for instance, whether it can identify a specific illustration or table within a work, something the SICI is not currently designed to do).

Both ARL and CNI are heavily involved in the SICI and BICI work; Julia Blixrud of ARL chairs the BICI committee, of which Clifford Lynch from CNI is a member, as are representatives of several other ARL and CNI member organizations. ARL is an institutional member of NISO.

The Digital Object Identifier (DOI)

In the past few months, the Association of American Publishers (AAP) and their technical contractor, the Corporation for National Research Initiatives (CNRI), have issued a great deal of publicity about a new identifier called, rather grandly, the Digital Object Identifier (DOI). The DOI is based on CNRI's *The Handle System*(TM), a very general identifier system that fits roughly within the URN framework, and that provides a mechanism for implementing naming systems for arbitrary digital objects. Thus far, the DOI has been demonstrated within the context of online consumer acquisition of intellectual property and perhaps for this reason it is somewhat difficult to disentangle the proposed DOI standard, the demonstration implementation of the DOI, and applications enabled by it. Major demonstrations of the DOI system are scheduled for the Frankfurt Book Fair in October 1997.

There are a number of misconceptions surrounding various aspects of the DOI. Its development does not mean that everything on the Web will become pay-per-view; rather, the DOI provides a method for collecting revenue for access to material that is described by a DOI (either on a one-time license or pay-per-view basis), *if* the organization that owns the rights to the object wishes to do this. Some objects described by DOIs may be accessible without charge. DOIs in and of themselves are only identifiers, and do not imply that any sort of copyright enforcement mechanisms (like an "envelope" or other secure container) will be bundled with the objects that they describe; the presence or absence of such copyright enforcement technologies is an entirely separate issue. These copyright enforcement technologies can be used with objects described by all sorts of identifiers, not just DOIs.

I believe there are some legitimate concerns about the use of DOIs as a means of implementing actionable citations among works on the Web, since this is likely to mean that the author of the citing work will need to obtain the DOI of the work that he or she wishes to cite either from the owner of the cited work or from some third party, and accessing a citation would then involve interaction with the DOI resolution service, raising privacy and control issues. But the notion that the use of DOIs will make the networked environment "safe" for proprietary intellectual property in a way that it is not today is as improbable as the idea that the introduction of DOIs, as one type of commonly used URN, will somehow convert the entire Web into a pay-per-view environment.

Discussions with the DOI developers suggest that the DOI's role will be as an identifier of content that is available for acquisition; there is currently some ambiguity as to whether it actually identifies content directly or if it simply identifies a method of acquiring content (such as an order screen). It is also extremely unclear under what circumstances similar objects are assigned distinct DOIs. Current plans seem to be to carefully control what organizations are

permitted to assign DOIs, limiting the groups to "legitimate" publishers; thus, a DOI is hoped to offer some "brand name" confidence to consumers purchasing content on the Net. DOIs will be assigned to content as it is made available for acquisition, and perhaps removed from the DOI database as content is withdrawn from availability for acquisition. It is important to recognize that there does not seem to be consensus on most of these issues at present within the DOI developer community, which underscores the uncertainties about the potential roles and utility of the DOI outside of its use as a means for consumers to acquire content.

In general, one cannot determine the DOI assigned to a digital object, or even whether the object has a DOI, unless the object carries it as a label. However, this can be confusing, because some publishers use, for those digital objects which are within the scope of the SICI, the SICI code as their (publisher-assigned) identifier. The implications of this practice will require careful examination and analysis. It is also unclear what role the DOI can usefully play in identifying material outside of acquisitions--for example, for material that is already licensed and is part of a library's collection, where it would be desirable to resolve "bibliographic" links to this material, but when it is inappropriate to connect library patrons to the acquisitions apparatus defined by the DOI.

It appears that DOIs can be implemented within the IETF URN framework, though there are a few messy details having to do with character coding; to the best of my knowledge no documentation has yet been developed which spells out these details.

Recently, representatives of the DOI developer community have asked CNI to work with them to help to increase understanding of the DOI's objectives and roles, particularly as they relate to library services, and to help to suggest ways in which the DOI might be made more useful to the broader bibliographic community. NISO has also been active in trying to relate the publishing community work on DOIs to the broader needs of the full NISO constituency, and held a workshop in June 1997 to begin developing requirements for general purpose bibliographic identifiers in the networked environment.

The DOI as it currently seems to be evolving is likely to be a useful tool to permit consumers to acquire content from publishers on the Net with some confidence about who they are doing business with. My present concerns with it relate to the lack of clarity surrounding many aspects of this identifier, the very broad applicability implied by the name DOI, which doesn't seem to be consistent with its actual definition (something like Publisher Object Access Identifier, or something similar, might be more accurately descriptive), and the very real potential dangers that are raised if this identifier is pressed into broader uses, such as a means of implementing navigable citations in digital documents. In a very real sense, there are no bad identifiers, but it is very possible to put identifiers to bad or inappropriate uses.

Conclusions

Many new identifier systems are appearing; some have been developed specifically for the networked information environment, while others are long-standing identifiers that are being brought forward into the digital context. When evaluating a new identifier system, there are a number of essential questions to ask:

1. What is the scope of the identifier system--what kinds of objects can be identified with it? Who is permitted to assign identifiers, and how are these organizations identified, registered, and validated?
2. What are the rules for assigning new identifiers; when are two instances of a work the same (that is, assigned the same identifier) within the system, and under what criteria are they considered distinct (that is, assigned different identifiers)? What communities benefit from distinctions that are implied by the assignment of identifiers?
3. How does one determine the identifier for the work, and can one derive it from the work itself, or does one need to consult some possibly proprietary database maintained by a third party? To what class of objects are the identifiers applicable? Within this class of objects, is there an automatic method of constructing identifiers under the identifier system, or does someone have to make a specific decision to assign an identifier to an object? If so, who makes this decision, and why? Note that, if the identifier cannot be derived from the identified work, it is unsuitable for use as a primary identifier within any system of open citation. The act of reference should not rely upon proprietary databases or services.
4. How is the identifier resolved--that is, how does one go from the identifier to the identified work, or to other identifiers or metadata to permit the instances of the work to be located and accessed? Again, what is the role of possibly proprietary third party databases in resolving the identifier? Do the operator or operators of these resolution services have monopoly control over resolution? What are the barriers to entry for new resolution services? What are the policies of the resolution services in areas such as user privacy and statistics gathering?
5. How persistent is the identifier across time? Can one still resolve it after the work ceases to be commercially marketed? Identifiers that rely on the state of the commercial marketplace are very treacherous for constructing citations or other references that can serve the long-term social or scholarly record.

All of the new identifiers are likely to be useful to some community, for some purpose, but it will be essential to determine what roles each new identifier is suitable for, and to avoid using various types of identifiers in roles that are inappropriate. The URN framework being established by the IETF invites all communities who are coming to rely on networked information to carefully consider what they need from identifier systems, and whether those needs are best served by defining new identifier systems.

Resources on Identifiers

URLs are defined in Internet RFC 1738. Functional Requirements for URNs are defined in Internet RFC 1737, and the syntax details are defined in RFC 2141. There are also a number of experimental resolver systems that are currently being deployed on a prototype basis on the Internet (see, for example, RFC 2168). There are also a number of internet drafts that are currently moving towards RFC status (see under "draft-ietf-urn" in internet drafts) that cover areas such as resolver system requirements and the use of bibliographic identifiers as URNs. See http://ietf.org.

OCLC Persistent Uniform Resource Locator

http://www.purl.org

National Information Standards Organization
http://www.niso.org

Digital Object Identifier System
http://www.doi.org

Source: *ARL: A Bimonthly Newsletter of Research Library Issues and Actions* 194 (October 1997). Washington, DC: Association of Research Libraries.

Digital Object Identifiers (DOIs) and Clifford Lynch's five questions on identifiers
by William Y. Arms, Corporation for National Research Initiatives
October 13, 1997

Table of Contents for Issue 194 | Other Networked Information Articles | ARL Newsletter Home

ARL Home