NISO Identifier Roundtable

March 13-14, 2006

National Library of Medicine, Bethesda, MD

Summary

NISO convened the Identifiers Roundtable on March 13-14, 2006 to clarify issues and promote consensus in this area. Experts from diverse sectors that rely on the exchange of digital resources met at the U.S. National Library of Medicine on March 13-14, 2006, to identify problems, clarify issues, and to develop an agenda for further work.

Key points that emerged were:

1. Information exchange between systems requires identifiers that are based on public standards, both for shared used of the identifiers and to prevent collisions between identifiers that are developed in different contexts.
2. Identifiers are part of an infrastructure that includes support services for creating identifiers, for binding them to information or objects, and for resolving an identifier to obtain the associated object or metadata about the object.
3. Long term sustainability of identifiers requires community and institutional support backed by viable business models.
4. Identifiers must be usable within the standards of the World Wide Web in order to operate appropriately in the current networked environment.
5. There appears to be broad general agreement on the nature and properties of identifiers, with perceptions to the contrary attributable to past disagreements that can now be seen as having arisen from the differing intended uses of specific identifiers.

A goal of the Roundtable was to develop a list of specific issues where NISO could make a helpful difference. The tasks that were identified (which are described in greater detail at the end of this report) are:

1. Establish a registry of identifier schemes that includes information about the associated services and policies for each scheme.
2. Explore use of the "info" URI registry as a focal point for community identifier needs.
3. Prepare a white paper on identifiers. This might include a glossary, a discussion of basic principles, services, and attributes for identifiers, and an implementer's guide.
4. Educate decision-makers and technology developers in the community about identifiers and their uses.

Report of the NISO Identifiers Roundtable

Key Discussion Points

What is an Identifier?

Identifiers for digital objects underpin information systems, and are essential for machine-to-human and machine-to-machine exchange of digital resources. Identifiers enable the packaging of digital information and the creation of services around such information, allowing its robust use in all sectors of human endeavor (including industry, commerce, academia, and government), affecting humans at scales from the individual to large organizations, and embodied in software that is highly visible (text processing, personal bibliographic databases) or nearly invisible (as is the case with much internal business processing). The reach of identifiers extends beyond traditional document management into scientific data sets, business processes, and commercial transactions that include a broad diversity of functional requirements and persistence expectations.

Identifiers exist to help manage information within a particular context or environment. This makes it difficult to develop universal or context-free answers. For any general question ("What are the ideal characteristics of an identifier?" "Should identifiers be resolvable?"), the answer is: "It depends." It depends on the context within which the identifier is designed; it depends upon its purposes; it depends upon its uses within a particular community. Understanding a community's specific needs is a necessary element for the development of a successful identifier system.

When identifiers proliferate, some items will have more than one identifier and some identifiers may appear to reference more than one item. Although this can be considered an error in some identifier systems, in other circumstances it is not. As an example, most individuals have many different identifiers associated with them: names, social security numbers, medical records numbers, and employee identification numbers. At times, multiple contexts will use the same number, as if the case with the social security number, which may be reused for a variety of identification purposes.

The uses of identifiers also proliferate. As identifiers age, the needs of their user community evolve, leading to "mission creep." This is how the ISBN, designed to identify books in the

publishing and retail supply chain, came to be associated with teddy bears and biscotti. This creates ambiguity around the identifier and its implementation. Mission creep may be preventable by rules and policies, but it may also be mitigated to some degree through identifier systems that are flexible enough to accommodate changing requirements. In any case, organizations that manage identifiers will have a role in directing the inevitable evolution of identifiers over time.

The Need for Identifier Standards

Private identifiers are useful for limited exchange where there are pre-existing agreements. However, the exchange of information across heterogeneous systems requires identifiers (and supporting systems) that are based upon public standards. Standards allow systems to "understand" identifiers that come from other systems, and prevent collisions between identifiers that are created independently. (1) NISO has played a role in promulgating identifiers commonly used within the library and publishing information communities, independently and through its participation in the work of ISO TC 46 / SC 9 (Information and documentation / Identification and description). With the emergence of the Internet and the World Wide Web, identifiers specified by the IETF (Internet Engineering Task Force) and W3C (World Wide Web Consortium), including URLs, email addresses, and format names, have come to play an important role both within and beyond these technical communities. (4) Unfortunately, a coherent, generalized architecture for digital identifiers is lacking, leading to the emergence of new approaches that sometime overlap and create confusion. The requirement that legacy (pre-Internet) identifiers must be accommodated in the digital world further complicates the management of globally unique, persistent identifiers.

Although many efforts are being conducted outside of the community of libraries and publishers, they will eventually influence the NISO community as it pursues its traditional role of generating, gathering, archiving, and disseminating information across all domains of human activity. The experience of NISO and its member bodies will help inform a broad interdisciplinary discussion of identifier systems and their requirements.

The Role of Identifier Systems

Some identifiers are tightly bundled with a system that manages the creation, resolution, and maintenance of identifiers. Such systems may supply a brand that a community comes to trust, although in discussions it was widely agreed that the syntactic form of an identifier can not itself convey trustworthiness. The degree to which bundling of identifiers, systems, and resolution services is desirable varies according to the application needs. Bundled identifier systems and resolution services provide complete, integrated solutions that may serve particular business needs. Many organizations, however, requiring autonomy in managing their own information resources, prefer either unbundled systems (e.g., URL-based) or partially bundled systems (such as URN, Handle, or DOI) in which complete responsibility for the final step in resolution rests not in the identifier system, but in the technical systems maintained by the organization.

Work on information systems tends to focus on technical problems, because they are generally better understood and easier to deal with than related socio-organizational problems. In the early days of the World Wide Web, identifier persistence was assumed by some to be a by-product of technology. It is now widely realized that persistence arises from a commitment of a maintenance body, and is not a technical property of an identifier.

The perceived value of an identifier rests in the user's expectation that it will produce a desired result, and that efficacy arises from the reliability of the systems using that identifier. It is the role

of organizations in the community to provide both the policy and technical fabric that supports and maintains identifier systems, and establishes the relationships of trust underlying them. At present, standards coordination is not done through a central authority, but rather through the collaboration of peer bodies (such as IETF, TC 46, NISO, and W3C).

To have viable community solutions, we need a shared understanding of the variety of roles of identifier systems in the information space and how they can be managed sustainably. Although used every day, identifiers are a mystery to many people, including people responsible for building complex information systems. Business managers may not understand how identifiers can be used to provide services and therefore may not include identifiers in their plans at the service level. Many people repeat the same trial-by-error learning experience as they develop new identifiers. For example, people may create semantically-laden identifiers and then learn how difficult they are to maintain over time. Even technology staff may not understand the various aspects of identifier use or the long term implications of the creation of identifiers in their services.

It can also be difficult to discover that useful identifiers already exist. Opportunities are missed when service creators cannot discover applicable pre-existing identifiers and identifier systems. It can be difficult to find clear and complete information about an identifier and its associated systems. A minimal description should include:

- The intended use(s) of the identifier
- The syntactic rules governing the form of the identifier
- What the identifier is intended to resolve to
- The technical infrastructure that is available to support use of the identifier, and the parties operating it
- Policies governing creation, maintenance, support, and persistence of the identifier
- Information about any metadata related to the identifier is available
- A history of the identifier, including changes in any of the above over time

The overall goal is to facilitate appropriate re-use of identifiers and to permit the creation of services using existing, publicly available identifiers.

Business Models

There are costs associated with systems that generate and maintain identifiers. These encompass the creation of standards, the development and maintenance of registries and services, and marketing and education activities. In the library and academic environment, the cost of identifier systems has traditionally been built into the overhead of curatorial organizations. This has hidden the cost from users who have come to see identifier services as free.

An identifier scheme may itself establish a brand, as with a product or an institution. This branding can be part of the business model of the identifier system and therefore is important to its social and business standing. It can be functional branding, designed to help build adoption and to directly inform users of that particular identifier's context and the way that it can be used.

Different business models can have different goals and values. In the academic community, a common business model is that of leveraging value through the sharing of resources. This business model favors open systems with little attempt to control re-use of resources. In the publishing community, the business model is focused on sales and licensing of resources. This model will favor systems that allow businesses to control access to resources as a way to protect

revenue. Both types of systems will use identifiers, but the systems and services supporting those identifiers will have different characteristics, one of which may be that identifier systems developed in the business community may support different kinds of functionality than those in the academic community. There also will be needs in some arenas for the use of private identifiers that are not known outside of their originating system.

Identifiers and the Web

The most publicly visible current application of digital identifiers is the World Wide Web. The Web's primary inventor, Berners-Lee, originally conceived of a family of identifiers, known as Uniform Resource Identifiers (URIs). Only two forms of URI have a notable presence in today's Web. Far and away the most ubiquitous of these is the Uniform Resource Locator (URL). The Uniform Resource Name (URN) is a related identifier, conceived in the early 1990's, and deliberated upon within the IETF. The URN is used, for example, to identify XML namespaces when a Microsoft Word™ document is saved as HTML. It has also been deployed within European libraries to retrieve documents through either special-purpose resolution systems, or by embedding a URN within a URL. As originally conceived, URN resolution would have been supported by the shared Internet infrastructure and invoked natively by web browsers, but this has not come to pass. Roundtable participants expressed frustration concerning the historic difficulty of applying to the IETF to register new URN and URI schemes. Nonetheless, new URN and URI namespaces have appeared (e.g., the Handle URN and the "info" URI) and the URI registration process has recently been revised to make registration more straightforward.

The URL has become the *de facto* identifier of choice for the Web because of its ease of creation and use, the ubiquity of related tools, its flexibility in disclosing (or concealing) brands, and its ability to point to a location within a document using relative anchors. An identifier system that uses URLs today is immediately functional on a large scale as long as resolution is the expected behavior. The success of URLs results in part from the success of ubiquitously distributed Web technology that has become so important to so many communities that we can anticipate the appearance of smooth migration paths when changes are made to the Web infrastructure or its descendants.

In the Web's nascence, URLs were concerned (as their name accurately reflected) with the *location* of a resource on the Internet. As such, they were less than perfect identifiers, as the objects to which they pointed often moved or disappeared altogether, leading to the familiar "404 Not Found" error returned to Web clients ("a broken link", in the vernacular). This led to a widespread perception that URLs were inherently incapable of serving as persistent identifiers. However, it became apparent over time that the flexibility of URL semantics, combined with the Web's forwarding (redirection) mechanism, allowed suitably committed organizations to provide potentially permanent identifiers in the form of URLs (as has been demonstrated by ARKs, DOIs, and PURLs, for example).

Definition of Terms

Discussions of identifiers often do not clearly distinguish between identifiers, identifier systems, and identifier resolution, leading to some confusion if the term "identifier" is used to describe different combinations of these three.

The concept ***identifier*** has been defined in differing ways. Defined by one Roundtable participant as "a relationship between a string and a resource," it is more commonly defined as a symbolic stand-in for a digital object, most concretely represented in current practice in the form of a linear

series of digital bits organized as discrete characters in accord with a defined protocol, a ***string***. The former definition is helpful; however, in emphasizing that an identifier is defined operationally, it is its ***association with an object*** that makes it an identifier, and that association is most powerfully manifested by the ability to use the identifier to retrieve its associated object. Such retrieval, however, depends upon the existence of a system for performing the retrieval, which in turn assumes the existence of an organization that operates the system.

An identifier can still exist as such in the absence of such a system, as long as a trustworthy entity asserts the relationship between identifier and object. These definitions make it clear at the outset that identifiers can not be fully considered without thought being given to ***relationships of trust***, and the organizations associated with them.

An IETF document (RFC 3404) implicitly describes identifier ***resolution*** to be a process by which an identifier string is employed to access its associated object and/or descriptive information about the object (***metadata***). This usually involves one or more intermediate mapping operations.

The term ***service*** is often understood differently depending upon one's professional affiliation. For computer scientists, it might refer to the technical systems for creating and using identifiers, whereas for librarians and businesspeople, it would more likely refer to the actions they take on behalf of their clients (for example, an interlibrary loan service), which are in turn often built on top of identifier services, and involve human staff who are managing said services and interacting with human clients.

Several attributes of identifiers and their associated objects are worthy of definition.

Resolvability refers to the ability of an identifier to be resolved as described above. The terms ***actionability*** and ***dereferenceabilty*** are sometimes used in the same sense.

Referent refers to the object which is identified by the identifier, whether or not resolution returns that object.

Granularity refers to the extent to which a collection of information has been subdivided for purposes of identification and resolution (for example, at one extreme, an entire body of technical literature could be identified as a collection, and at the other, components of individual reports, such as tables and figures, could be identified for retrieval).

Persistence refers to the degree to which the resolvability of an identifier is matched to the business process that supports the association of identifier and referent. As such, it is an attribute of the association between an identifier and its referent than of the identifier itself.

Semantic opacity relates to the extent to which an identifier may itself carry information about an object (a fully opaque identifier carrying none).

These last terms will be enlarged upon in the following discussion of key identifier attributes.

Key Attributes of Identifiers

Granularity: When a book was offered in print form, a single identifier sufficed. With books available in digital form, individual chapters, illustrations or other components may serve as free-standing information units. Different levels of granularity within the same resource may be required to serve a variety of needs. When an identifier system does not provide such flexibility, parallel identifier systems may arise for the resource. The granularity "problem" is not one that has a single solution.

Semantic opacity: When parts of an identifier string may be inferred to be assertions about aspects of the object identified, such assertions can become misleading, infringing, or offensive due to semantic drift. A fully-opaque identifier does not enable such inferences. Such an identifier can function over long periods of time during which organizations change names, new trademarks and acronyms arise, and subject hierarchies and language evolves.

Some identifiers (*e.g.* ISBNs) are probably opaque enough to be long-lived, as their semantics are widely recognizable only to information professionals. Semantically-laden (non-opaque) identifiers, while potentially perishable, provide usability advantages by allowing one to select or verify them (*e.g.*, from within a list) by drawing inferences about the related object. Such strings function as metadata containers, and their structure and semantics range from *ad hoc* and unpublished to fully standardized (*e.g.*, OpenURL, SICI).

Some providers support or tolerate exploitation of recognizable identifier semantics, with one popular application being a reverse inference that permits a user to start with the content and correctly guess its identifier. One participant, however, described a system in which such user behavior was viewed as unwelcome "identifier hacking" leading to unintended access to collection resources that were not supported, a clear practical illustration of the importance of context in identifier functionality. Semantically-laden identifiers are difficult to maintain unless their semantic qualities are likely to remain unchanged (for example, the year at the beginning of an LCCN provides useful information that will not change with time).

Persistence: The degree required varies by application and is established in the context of a business case manifested as an organizational commitment to continued access. For example, identifiers used in tracking shipments require limited persistence. For information resources, it is often important to maintain the relationship between identifier string and object even if the resource is no longer available. In such instances, a resolution service may, appropriately, no longer resolve to the actual resource, but the string is still a *bona-fide* identifier if the service can provide information about the related resource (itself a kind of resolution).

It is often impossible to provide perfect solutions that address all of these identifier attributes, as the environment is somewhat chaotic and beyond the control of any organization or application. Lessons learned from existing systems can help information providers develop "good enough" solutions that largely meet user needs.

Potential Tasks for NISO

Numerous needs were expressed during the day and a half of the Roundtable. Some of them, such as the lack of universal resolution services for URNs, were deemed outside of the scope and capabilities of NISO and its community. Other issues were seen as appropriate NISO activities. The group recommended the following as general areas for NISO to pursue, with specific goals for each.

Education

Those who create identifier systems, and those who use identifiers, need to learn from their colleagues across time and space. NISO can provide publications and courses to support services and promote interoperability between systems. NISO could prepare simple, clear documents along the lines of its "Understanding Metadata" publication, addressing those who use identifiers to provide services, system developers implementing identifier-based systems, and would-be creators of identifiers. A number of specific documents were suggested:

- A technical report (possibly written in collaboration with others) describing how URNs are intended to be used, and how they relate to “info” URLs, addressing issues of location, resolution, and identification.
- A URN “profile”: a technical report outlining the URN specification as it pertains to identifiers of most interest to the NISO community.
- An “Implementor’s Guide” describing the practical aspects of selecting and using an identifier and its related technical services.

Service & Policy Registry

The identifier community needs a way to discover available identifier services and policies. Existing services could be used within new information systems, increasing efficiency and reducing the needless proliferation of competing approaches.

A registry will require a business model, as well as development and maintenance resources. A first step would be to create scenarios demonstrating the value of the registry, which could serve as a focus for community discussions of the functional requirements for such a registry. The scenarios would demonstrate discovery services related to otherwise non-actionable identifiers, and would be

- neutral as to business models;
- demonstrate cross-sector services; and,
- define the context dependence of the service.

The scenarios would demonstrate how services can be recombined to provide new functionality. The following scenarios were discussed at the organizers' meeting at the end of the Roundtable:

1. A menu of services related to identifiers with the string "doi:". Proxy URLs are a single redirect and DOIs and Handles are often buried in proxied URLs. In response to the string "doi:", the registry would return a menu of services in structured XML; a web browser plug-in would display the response. The identifier could be within a PDF file.
2. Resolution of ISBNs to services, in particular use of the OCLC x-isbn list for an ISBN.
3. Contextual resolution of music product identifiers in support of online services such as Napster.

The group also discussed the virtue of developing a vocabulary for standardized statements about services (policies); one application would be the formation of a controlled vocabulary of persistence promises along the lines of the permanence rating scheme devised by the National Library of Medicine. Entry within the registry might serve as a stepping-stone to elaboration as a formal standard for services and policies that demonstrated strong interest among users.

"info" URI Registry

NISO is the support agency for the "info**"** URI. Developed during the time that NISO Committee AX was working on the standard for the OpenURL, this new URI allows the creation of identifiers that the NISO community needs, and includes, importantly, the ability to turn legacy identifiers into URLs. NISO is in the process of formalizing the policies that will be used in "info**"** URI creation and maintenance, and in setting up the management procedures for the "info**"** URI registry. This identifier and its registry could serve as a focal point for NISO's identifier activity, creating a trusted brand and a starting point for community members doing work that requires identifiers.