UBIQUITOUS IDENTIFIERS FOR GREY LITERATURE

Kimberly Douglas
California Institute of Technology
Pasadena, CA
kdouglas@caltech.edu

ABSTRACT: Digital identifiers are a necessary and critical feature of digital libraries and repositories. Various organizations have developed and attempted to promulgate different types of digital identifiers for ubiquitous and persistent use. The primary examples are the DOI, URN and the PURL.

Grey literature is produced at all levels of government, academia and many businesses and industries where publishing is not the primary business activity. Nevertheless, many of these materials are easily available over the Web and are regularly cited in the more formal literature.

Should organizations be encouraged to use any of these identifiers to improve identification and access to this literature?

Identifiers in the digital world make things happen – in that way they are different from the identifiers we know in the print world. Clifford Lynch uses the word "actionable" (Lynch 1997). In other words the identifiers will cause something to happen – they will translate into some thing. In the print environment there is always the human in between to interpret the identifier, e.g. ISBN, into its corresponding thing using look-up services or catalogs. Not so in the digital world – the identifier will be the key to getting the document or information object. It will have to live a long and productive life.

Identifiers are not new within the publisher and library environment. The ISBN grew out of publisher needs and was embraced by the library world by adding it to bibliographic descriptive information for verifying and tracking resources. Digital identifiers perform a similar function with the difference that without an identifier resolution system networks will not be able to reliably find and display a specific information resource. This very purpose is what makes the digital identifier critical in the Internet.

Yet digital identifiers for unique and persistent content identification via the Internet remain illusive despite three serious standardization attempts. Not just formal publishers but grey literature publishers, government, academia and many businesses and industries for whom publishing is not the primary business activity, need to seriously address the identifier issue in order to ensure the best use of their resources over time.

Properly conceived and maintained identifiers provide stability and persistence. Authors require and expect this kind of reliability. One difference between grey literature and the formally published literature is the reliability with which one can cite and retrieve a document. Attention to identifiers can make grey literature more consistently accessible in the online environment.

Three digital identifier initiatives have different origins in time, agency and purpose. From the perspective of a current grey literature author/publisher an identifier scheme and resolver system needs to require as little overhead effort as possible. Reliable access and persistence over time are the priority purposes and for these organizations current copyright law suffices for intellectual property control.

A successful identifier system has three components:
Identifier system

1. A managed identifier assigning process: There is an assignment chief though this may not amount to much if derived totally from the object. Policies for who or what gets an identifier have to be established. This is not a new activity for a grey literature publisher that has handled technical report series. However, in the web environment there may be more types of material to track and certainly versions to consider. The library, the publicity department, or documentation group, are possible entities for this job.

2. Resolution: There needs to be a technical solution for recognizing the identifier and pointing the browser to the actual place and document. Resolver systems are to URLs as Domain Name Servers are to domain numbers. It needs to be set-up and maintained by trained technical staff. This can be done locally for specific implementations as are now the example with the NCSTRL and American Memory projects. Or, with global commitment, a system similar to the DNS can be established. Currently an organization must seriously evaluate how much technical or systems support the organization can provide.

3. Reverse look-up service: The assigned identifiers need to be included in discovery databases for the type of object involved. There will and should be many types of databases that pick up this information. OPACs can be used to find ISBNs; there is also an opportunity for value-added 3rd party services. Occasionally the organization's library has provided access through the library's catalog or the documentation center has provided a database for internal use. Many government agencies have contributed records to OCLC, RLIN and other national and international bibliographic centers.

As early as 1990, the IETF (http://www.ietf.org/html.charters/urn-charter.html) began efforts to standardize syntax and management principles for the Uniform Resource Name

(URN) as part of the Uniform Resource Identifier (URI) to reliably identify, link, find and manage electronic resources on the global network..

Though the URN efforts are as yet unfinished and discussions continue to this day major accomplishments comprise:

1. Established a beginning point for all identifier standard discussions and decisions.
2. Provided a context for prototype resolver development, CNRI's Handle System that supports resolution of one name to many resources – unique at this time though not demonstrated in the http environment.

Three obstacles have to be overcome before the URN can break out of its current stalemate.

1. Web browsers must support URN syntax.
2. A management scheme similar to Domain Name Servers must exist. A proposal in the form of an RFC for the Naming Authority Pointer (NAPTR) was posted in June 1999 and is modeled after the IANA, the Internet Assigned Numbers Authority. However, an implemented management scheme is not currently in place.
3. There needs to be an advocate for the URN with a critical need that it can solve. This advocate has to have clout.

Considerations for grey literature publishers:

The URN provides no implementable technology although arguments for URNs are clear. Those organizations that have made use of the URN do so by embedding the Handle Resolution software within the http protocols of a Web server to create a proxy server for the Handle System. Such development requires considerable computer system support.

In 1995, OCLC, which had been participating in the URN work, decided to focus their attention on solutions to persistence in naming and access. Quickly, the PURL server was developed using http redirect functionality to resolve a persistent name in the form of a URL to a location – a real Uniform Resource Locator. OCLC chose not to address management issues of how to name versions or different formats. OCLC also decided to provide the PURL resolver service to anyone who cared to use it. That policy remains in effect today.

OCLC will persistently maintain the PURL resolver (http://purl.org/) even when other resolving standards and technologies are implemented. PURLs will be supported as long as the location is kept up to date by the maintainer and/or the location continues to exist. OCLC will attempt to bring the PURL into line with any URN standard that is adopted. The sense I get from conversations with Keith Shafer is that since the PURL management is so open – some registrations and registrants may not comply with URN standards. One has to believe that there will be "space" in the URN standard to register PURL servers.

Of course, syntax will be different so new use of PURLs as URNs would use the new syntax – the old would be unchanged as PURLs only. Nevertheless, OCLC will do what is possible and reasonable without compromising existing implementations. How that will be accomplished technically is not yet known nor can it be currently known. OCLC did a demonstration project that showed PURLs could be resolved by the Handle system.

Accomplishments:

1. PURL is a current solution to the persistent naming and access problem.
2. PURL server at OCLC can be used by anyone at this time.
3. Use of the PURL server requires no local systems support. OCLC provides the system support.

Obstacles:

1. PURL server is located at OCLC, Columbus Ohio and may not be viable for non North-American agencies.
2. Browsers cannot distinguish between URLs for location and PURLs for persistent naming. They look the same. Users are thus also not conditioned to tell the difference and would not necessarily understand that the location to which a PURL resolves is not the right persistent access point. Conventions for communicating a PURL versus a URL on the document or in the browser do not exist.

Considerations for Grey literature:

1. The authoring or publishing agency must administratively commit to maintaining the PURL by updating the URL it resolves to when it changes. The PURL is free for the using right now. Use of the PURL server at OCLC requires no technical or systems support. An organization must identify a department to take on the job of maintainer over time and needs to devise a method for communicating PURLs and encouraging the use of the assigned PURLs to potential users and its authors.

In 1997, the third primary identifier was born, the Digital Object Identifier, the DOI (http://www.doi.org/). The commercial publishing sector under the auspices of the American Association of Publishers gave the impetus for this development. In October 1997, the International Digital Object Identifier Foundation (IDF) was formed to serve as a central vehicle for organizing the environment under which the commercial publishing sector could productively engage in electronic publishing. Major publishers, Springer, Elsevier, Wiley, Academic, American Chemical Society and others underwrote the efforts as member organizations. Over the last two years, the President, Norman Paskin, has single-handedly forced an international, cross business sector discussion and decision-making construct that has generated more energy and global interest in identifier issues than seen before.

The main focus of the IDF is to manage intellectual property in the online environment. The IDF asked NISO to form a committee to establish a DOI Syntax standard. The IDF

has researched and contacted international organizations regarding standards for non-text forms of intellectual property, e.g. music. Currently the IDF is very much involved in the INDECS efforts to categorize rights trading information so that access to electronic documents can be properly tracked and managed.

Accomplishments:

1. Re-opened the URN discussion
2. Provided broad-base discussion and consensus forum for digital identifiers.

Obstacles:

1. URN syntax based and thus hindered by lack of web browsers that can resolve URN.
2. Require a specific assigned prefix which costs $1000.
3. Registration agency not yet clearly established.

Considerations for Grey literature

1. Heavy emphasis on rights management metadata. Such emphasis is needed within a licensing environment but may have less relevance in cases where copyright laws and other rights doctrines, fair use, are the basis for access and use.
2. Management issues may be more prescriptive and intensive than a grey literature publisher may want or need.

In examining two well-known sources for grey literature publications, XXX preprint archive at LANL and the NCSTRL distributed collection of Computer Science Technical Reports, it was discovered that neither builds or uses straightforward identifiers that provide for persistent linking now. The LANL system automatically generates a specific item identifier based on subject section, date submitted, and number-order received on the same day. However, it is not possible to identify logically an ID that will persist over time. The specific location of the preprints is within a subdirectory named 'abs.' If that subdirectory name changes over time or is removed, links using that address in other tools, lists or publications may not reliably resolve.

Example 1:

http://xxx.lanl.gov
Astrophysics, abstract
astro-ph/9909217
From: Elizabeth J. Barton <betsy.barton@hia.nrc.ca>
Date: Mon, 13 Sep 1999 23:39:49 GMT (519kb)
Tidally-Triggered Star Formation in Close Pairs of Galaxies
Authors: Elizabeth J. Barton (1,2), Margaret J. Keller(1), Scott J. Kenyon
Location: http://xxx.lanl.gov/abs/astro-ph/9909217

Based on the host and the paper number, this made-up URL is a logical identifier.

http://xxx.lanl.gov/astro-ph/9909217

It resulted in a 404 message with the text:

"The requested URL was not found on this server. If you're looking for a paper try something like http://xxx.lanl.gov/abs/astro-ph/9909217"

Wouldn't it be better to promote: http://purl.org/lanl/astro-ph/9909217 to resolve to the correct location over time?

NCSTRL was another grey literature server examined. The records give a URN but it is not clear how it could be reliably used for persistent linking. TheURL is too long and cumbersome for permanent use. The possible logical URL http://ncstrl.org/ncstrl.princeton/TR-567-97 does not work nor does http://www.ncstrl.org/ncstrl.princeton/TR-567-97.

Example 2:

http://www.ncstrl.org/
A Java Filter
Dirk Balfanz and Edward W. Felten
December 1997
URN = ncstrl.princeton/TR-567-97
URL = http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.princeton/TR-567-97

Wouldn't http://purl.org/NCSTRL/ncstrl.princeton/TR-567-97 be a much more useable and potentially permanent identifier?

Grey literature publishers need to think of how to make identifier management fit their business plan. Publishing is a means to an end. Most grey literature publishers and their authors publish their material so that it will be disseminated. The desirable outcome of the effort is wide distribution at little or no cost. If there are charges, it has been primarily to underwrite printing and mailing costs not to gain revenue. It is in the grey literature publishers' interests that third parties can make reliable links to their resources. Since a grey literature publisher is not necessarily including the cost of publishing in its business plan, it may be voluntary, the cost of maintaining resolution systems support in addition to source material servers may be more than the organization can underwrite. The priority technical effort needs to be the maintenance of the digital object for which they may be the sole source.

If a grey literature publisher is advertising or promoting URLs, the issue of persistent names naturally and logically has to be addressed. Without a doubt, systems staff will move the objects, thus changing the access point.

Failing the URN resolving browser and central agency, it appears that the PURL – the persistent URL – is an appropriate immediate step that grey literature publishers can take toward providing persistent access to their literature. Since http protocols are the means by which browsers operate, users are developing ad hoc persistent URLs. The PURL is a method now available to grey literature intellectual property managers nearly independent of system resources.

PURLs allow objects to move and the identifier to be widely and uncontrollably distributed for maximum exposure and still be reliable. The PURL allows an entire site to be redirected with a single PURL partial redirect. So, using PURLs does not require that every object needs an individual PURL. There only has to be organizational commitment to maintaining the destination URLs in the PURL service. Use of the PURL service allows the grey literature publisher to focus resources more appropriately on identifier management policies and descriptive data needed for resource discovery services.

References:

LITA Electronic Journals Electronic Publishing Interest Group. 1999. Making Sense of Digital Identifiers for Internet and Other Online Applications – Bibliography. [Online]. Available: http://www.lita.org/igs/Epej/preconf699/preconf699bib.html [November 4, 1999]

Lynch, Clifford. 1997. "Identifiers and their roles in networked information applications." *ARL: A Bimonthly Newsletter of Research Library Issues and Actions* 194 (October 1997). Washington, DC: Association of Research Libraries. [Online]. Available: http://www.arl.org/newsltr/194/identifier.html [November 4, 1999]