Persistent Identification Systems

Report on a consultancy conducted by Diana Dack for the National Library of Australia. May 2001

Introduction

1. Background
2. Persistent Identifiers and the NLA
3. The Achievement of Persistence of Access to Resources
4. Summary of Recommendations

Appendices

1. NLA Guidelines for the development and application of a persistent Identification scheme for digital resources
2. Adoption of an NLA persistent redirect service
3. Handles test
4. Managing web resources for persistent access
5. Select reading list on persistent identifiers and naming schemes

This paper looks at the following issues involved in assuring the long term availability of references and citations made to digital objects in the context of open access on the Internet:-

- The role of persistent identifiers in ensuring consistent availability and accessibility of resources in an open networked environment.
- Existing schemes for persistent identifiers and the long-term considerations involved in establishing naming systems and the local and global considerations which must be addressed
- Immediate options available to the NLA.
- Recommended courses of action

PERSISTENT IDENTIFICATION SYSTEMS
Part 1 : Background

1. BACKGROUND

1.1 The distributed digital library

The WWW has produced a revolution in the availability and accessibility of electronic information for users of all types and has the potential to continue to effect vast changes in the way libraries, archives, publishers, etc. deliver information to their clients.

One of the benefits of the www for libraries is that it has enabled them to deliver services in a way undreamed of less than a decade ago. Digital collections can be delivered over open networks using standard, multipurpose software in the form of standard web browsers already on the desktops of their clientele. Prior to this development, the ability to deliver these kinds of services electronically was limited by the need to use (multiple) proprietary systems and to have some control over the software on the desktop of remote users.

The technology and infrastructure currently available makes possible for the first time a de facto distributed digital library, providing direct access to a multitude of resources through commonly available technology using open standards. Digital resources can be accessed directly by links embedded in databases, catalogues, indexes, gateways and by links embedded in other digital objects or resources.

Such a distributed digital library would consist of a vast range of digital resources housed and maintained by a variety of different institutions. Some “born digital” materials will be maintained and distributed by the publisher, some will be duplicated by archival institutions or “taken over” once the publisher no longer maintains them. Much of the material available from libraries and archives will be digitised versions of items in their own collections, other material will be produced specifically for public distribution, yet other material, initially produced as part of business operations, will be made available for public distribution. Libraries and archives will also make available archived materials.

The strength of this distributed model is that it leverages off an architecture that is open and almost

universally deployed.

Its major weakness has now been apparent for several years and relates to the almost universal use on the web of identifiers that are based on the location of the digital resources they identify and which are therefore subject to change as resources move and systems change. For the open architecture digital library or archive, this is a critical flaw.

One of the keys to the development of a distributed digital library or archive is the ability to rely on the links between the digital resources to remain static over time, whether these are links embedded as citations in other resources or links embedded in metadata in bibliographies, indexes, and catalogues or in bookmarks and gateways. Much of the value of digital resources for scholarly communication lies in enabling resources to be cited reliably with actionable links over long periods of time. Libraries and archives collect for the long-term to support scholarly research and need to support access to materials beyond the life of current technology and possibly beyond the organisational structures which exist today.

Thus libraries, archives, academic institutions and publishers have an interest in the persistence of resource identification. Persistent naming is a characteristic of all open architecture digital library experiments. Persistent identifiers and their associated metadata are needed to support a variety of transactions including the search, retrieval, selection and use of digital objects.

There are two elements of the task of maintaining the availability of resources in the distributed digital library:-

- To preserve, or archive, the item in order to ensure that it survives;
- To maintain its accessibility to ensure that it can be located and displayed by those who wish to use it and that citations and links made to it continue to provide access to it.

The issue of persistent identification is primarily concerned with the second, but is also closely related to the first.

William Arms sums up the importance of persistent identifiers in his article on the “Key concepts in the architecture of the digital library”

Names are a vital building block for the digital library. Names are needed to identify digital objects, to register intellectual property in digital objects, and to record changes of ownership. They are required for citations, for information retrieval, and are used for links between objects.

1.2 Existing formal identifier schemes

There are many formal identifier or naming schemes which have been discussed in the context of the naming of digital resources (e.g. URIs, URNs, DOIs, Handles, ISBNs, ISSNs, SICIs, BICIs, ISWC, PII), although very few of these will be fully effective as persistent identifiers facilitating access to online resources in a distributed system unless they are either registered as URI naming schemes and supported by a resolution system or they are incorporated into another naming scheme which has

some form of resolution system associated with it.

The importance of resolution cannot be overestimated. The strength of the existing URL/URI standard which supports most of the information on the www is the widely deployed resolution system in the form of the DNS (Domain Name System). Without a resolution system, requests using the identifier cannot be routed to the appropriate server and used to retrieve the resource, or a reasonable substitute for it in the form of metadata.

1.3 General URN/URI Issues

1.3.1 Uniform Resource Identifiers (URIs)

The identifiers used on the www at present are URIs. Two broad categories of URI, the URL and the URN, are described in *RFC 2396 – Uniform Resource Identifiers (URI): Generic Syntax (August 1998).*

"*A URI can be further classified as a locator, a name, or both. The term "Uniform Resource Locator" (URL) refers to the subset of URI that identify resources via a representation of their primary access mechanism (e.g., their network "location"), rather than identifying the resource by name or by some other attribute(s) of that resource. The term "Uniform Resource Name" (URN) refers to the subset of URI that are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable.*" RFC 2396, p.2.

Of these two categories, the only one that is widely deployed at present is the URL, and most resources currently on the www are described using a URL. As mentioned above, the URL generally describes a resource in terms of its current location.

It is the location-based nature of the URL that is its major weakness as a persistent identifier, since the identifier will change if the material is moved either to a different server or a different place in the file structure of the same server. If this happens, all links to this resource embedded in other documents or databases will cease to work. The volatility of material on the web generated by movement of files and re-organisation of sites, not to mention deletion of materials, has highlighted the difficulty in guaranteeing continuous access to materials in a system which uses a naming structure based on file location to provide that access.

1.3.2 Uniform Resource Names (URNs)

This need for persistence in the naming of networked resources was recognised early in the life of the Internet and has been the subject of discussion for several years within the IETF (Internet Engineering Task Force).

Work on the development of a unique identifier that would be independent of location and would remain permanent even though resources might move or disappear began at the same time as work was proceeding on the formalisation of URL schemes. The identifier was named the URN (Uniform

Resource Name). However there was little agreement on the details of such a scheme and consensus on the syntax and semantics of the identifier remained elusive for a period of years. A specification outlining the minimum requirements of the URN, *RFC 1737 Functional Requirements for Uniform Resource Names* by K. Sollins and L. Masinter was issued in December 1994.

This document outlined a general agreement on the basic requirements for URNs (Uniform Resource Names) and by definition any useful persistent identifier scheme. These are:

"Global uniqueness: The same URN will never be assigned to two different resources.

Persistence: It is intended that the lifetime of a URN be permanent. That is, the URN will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name.

Scalability: URNs can be assigned to any resource that might conceivably be available on the network, for hundreds of years.

Legacy support: The scheme must permit the support of existing legacy naming systems, insofar as they satisfy the other requirements described here. For example, ISBN numbers, ISO public identifiers, and UPC product codes seem to satisfy the functional requirements, and allow an embedding that satisfies the syntactic requirements described here.

Extensibility: Any scheme for URNs must permit future extensions to the scheme.

Independence: It is solely the responsibility of a name issuing authority to determine the conditions under which it will issue a name.

Resolution: A URN will not impede resolution (translation into a URL, q.v.). To be more specific, for URNs that have corresponding URLs, there must be some feasible mechanism to translate a URN to a URL."

The RFC goes on to outline the necessary characteristics of a URN specification which would be required to meet this list of requirements, the most significant of which are:

"To satisfy the requirements of uniqueness and scalability, name assignment is delegated to naming authorities, who may then assign names directly or delegate that authority to sub-authorities. Uniqueness is guaranteed by requiring each naming authority to guarantee uniqueness. The names of the naming authorities themselves are persistent and globally unique and top level authorities will be centrally registered."

and

"It is strongly recommended that there be mapping between the names generated by each naming authority and URLs. At any specific time there will be zero or more URLs into which a particular URN can be mapped. The naming authority itself need not provide the mapping from URN to URL."

Although these general principles were accepted, reaching agreement on the detailed syntax of a URN was difficult, and in 1995, the URI group split into two, with a separate group working on the URN specification. The differences between the protagonists were evident at two levels. The first was a division between supporters of the URN concept and those who did not believe that there was a need for a URN. The latter believed that the URI could be managed to achieve the kind of persistence required. The second was a division among supporters of the URN concept who were having difficulty in reconciling their various implementation proposals in order to reach consensus on the details of the specification.

At the end of 1995, the implementers reached a broad consensus on the syntax of a URN and agreed to work towards a technical solution that would accommodate the needs of most users. Nevertheless it was May 1997 before URN Syntax was formally agreed upon and published in *RFC 2141 - URN Syntax*.

The syntax of the URN as expressed in *RFC 2141* is as follows:

urn:<Namespace identifier (NID)>:<Namespace specific string (NSS)>

The NID, which must be registered and approved by IETF to avoid duplication, ensures the global uniqueness of the identifier. The namespace specific string can take any form specified by the naming authority provided that it is unique within that namespace and avoids the use of a small number of restricted characters as specified in the *URN syntax, RFC 2141*.

The simple structure of the identifier reflects recognition of the need to accommodate different requirements and different schemes. Because the local, or namespace specific, string can be in any form, the identifier structure allows maximum flexibility in the identifier while providing a mechanism to assure global uniqueness and facilitating interoperability between discrete systems.

1.3.3 Requirements for resolution

Despite the differences expressed over the form and utility of a URN, there is general agreement that there is a need to distinguish between naming schemes and resolution systems. That is to say that a naming scheme, as a procedure for creating unique URNs that conform to a specific syntax, is independent of the resolution service which resolves them to locate the resource. A naming scheme should not be tied to any specific resolution system and a resolution service should be capable of resolving a URN from any given name scheme.

This is consistent with the intentions behind the development of the URN. A persistent identifier, especially when used for archival data must of necessity be capable of outlasting any systems and protocols that are currently in use. However the lack of a commonly agreed upon resolution system is also a major obstacle to the wide deployment of URNs.

A variety of solutions have been proposed, including a new DNS resource record, NAPTR (Naming

Authority PoinTeR), as described in RFC 2915, that provides rules for mapping parts of URIs to domain names. This was made obsolete by a revision and renaming of the proposal issued as an Internet draft in February 2001. The new proposal is named the Dynamic Delegation Discovery System (DDDS). The DDDS approach is to use the DNS to locate "resolvers" that can provide information on individual resources, potentially including the resource itself.

The use of http proxy servers is also a possibility. However wide deployment of URNs remains elusive.

1.3.4 What is the outlook in the near to medium term?

"*There is a pressing need amongst the Internet community for a widely available, preferably decentralised, means of managing URNs and associating them with URLs. This is necessary to avoid the problems associated with temporary names (URLs), which are stored in directory services such as archie and would be helpful in providing support for rudimentary file caching on the Internet, as it provides a mechanism which is suited to the automatic selection of geographically ``close'' copies of a widely distributed resource.*"

This quotation is an extract from a message sent by Martin Hamilton to the uri@bunyip.com mailing list on 13 Jul 1993 (http://wwwwbs.cs.tu-berlin.de/html/urls/0205.html). As mentioned earlier, it was a further year before *RFC 1737 - Functional requirements for URN* was issued in December 1994, and a further 3 years before agreement was reached on the syntax and *RFC 2141 URN Syntax*. was issued in May 1997.

It is disappointing to note that in the four years since then, little progress has been made and the URN appears to be no closer to widespread deployment.

It is not easy to predict how the situation relating to URNs and URIs will develop. The length of time taken to arrive at an agreed syntax for the URN reflects the lack of consensus amongst some of the major players on the need for a URN at all. This lack of consensus is still present in the IETF and W3C community, as can be seen in the relevant discussion lists.

There is a fundamental philosophical difference between the two positions adopted. The push for a "stronger" identifier than the URL, as it is currently used, is largely coming from those in the digital library field who recognise that persistent naming is the cornerstone of the architecture of an open distributed library/archive system and from the rights management community who need an absolutely watertight persistent identifier system to support commercially oriented rights management systems. This position also is increasingly attracting the support of the community of xml developers who require persistent links for xml-schemas. To fulfil the needs of these groups (which include NLA), links must be stable over very long periods of time and be, as far as possible, location and system and protocol independent. Those who believe the URN development to be unnecessary and misguided argue that persistence is an administrative, social and management issue and that the URI as outlined in RFC 2396 will, if managed properly, meet all the needs of those advocating a URN standard.

There is a reluctance to accept that the concern of those who will be making huge investments in digital archives, whether of digitised content or born digital material, has a legitimate basis. To date there has been little concession that there are deficiencies arising from the current confusion and uncertainty over the URI/URL/URN status and supporting processes which weaken the case for the URI as a stable foundation on which to build a global digital library/archive. The URI/URL standards and registration processes need to be tightened up and clarified before it can fulfil the claims of its proponents.

For URNs to achieve widespread deployment, not only is consensus on functional requirements and syntax needed, but the ability to recognise and route URNs must be incorporated into standard Web browsers and a resolving infrastructure developed either by leveraging off the existing Domain Name System or by some other route. This is unlikely to happen in an environment where browser makers are uncertain what is accepted as a standard and developers of naming schemes are uncertain how to register their schemes.

The IETF (Internet Engineering Task Force) is responsible for the development of Internet protocols and standards. However in the area of WWW standards, W3C has the major role. It is responsible for Web resources and protocols for access and distribution, e.g. HTML specifications, PICS, RDF. The URI scheme is regarded as a Web specification and its extension or modification falls within the scope of W3C activities.

W3C has to date shown little interest in URN, or "name style" identifiers. In spite of this, the issue has continued to be debated. There is currently a W3C interest group looking at issues associated with URI schemes. The issues identified by this Group will be recommended as a workplan for a possible working group to work on and hopefully solve.

W3C has strongly resisted any move to develop a standard outside the framework of the URI. The position was stated early by James Miller in an article in D-Lib Magazine in November 1996, when he argued that all the technical infrastructure for the development of persistent identifiers was already available and it remains only for those interested to develop administrative and organisational infrastructure to realise their implementation. The continued adherence to this position is evident from current documentation on the W3C website and contributions to online discussions by W3C staff.

Whilst this position has some truth to it, it is also true that the support of W3C would make the task easier by encouraging the acceptance of name style identifiers by the major browser manufacturers. There is also as yet no standard resolution infrastructure for the resolution of URNs. The DDDS proposals mentioned earlier are still Internet drafts and will take some time to achieve the status of standards, let alone significant implementation.

In short, there is little prospect of an early solution to the resolution issue. All URN systems implemented in the near to medium future will require proxy servers to enable them to be used by standard web browsers and to route requests to the host server. This means that the identifiers will have to be encapsulated in a URL using the proxy server address.

1.4 Specific naming schemes

1.4.1 NBN (National Bibliography Number)

The NBN is a URN Namespace Identifier (NID) which was developed and registered by the National Library of Finland. It was subsequently considered and accepted in principle by the Persistent Identifier subcommittee of the CDNL (Committee of Directors of National Libraries) as a potential vehicle for a common URN based identifier for national libraries.

The NBN as originally conceived by the National Library of Finland was simply an electronic National Bibliography Number, and was essentially a running number preceded by a country code assigned to a single digital resource. The specification was subsequently further developed to accommodate a variety of needs and in the most recent registration proposal as an NBN namespace (January 2001) it was defined more broadly, although still in terms of a national bibliography.

Syntax of the NBN

The NBN consists of the Namespace Identifier (NID) “nbn” and a Namespace Specific String (NSS) The NSS commences with a prefix which identifies the origin of the number. The prefix may be a two character country code based on ISO 3166 or some other registered prefix. The prefix is followed by a delimiting character, usually a hyphen, and the string assigned by the national library. The resulting syntax is as follows:

urn:nbn:<ISO country code>-<assigned string>

e.g. urn:nbn:fi-20002561

urn:nbn:<registered prefix>-<assigned string>

e.g. urn:nbn:LCCN-9920567 (a Library of Congress Control number)

Resolution

There is no resolution system defined for the NBN. It is suggested in the Registration document that resolution would be through national bibliographic databases, or through web indexes or archives. In earlier drafts, it was stated that for global resolution, a NAPTR (Naming Authority PoinTeR) record would be created for each prefix. However neither NAPTR nor DDDS are mentioned in the current draft (issued January 2001).

As a URN, use of the NBN will be dependant on the deployment of URN resolution services, and until they are developed, will require http proxy servers to resolve.

1.4.2 PURLs

Background to PURL development

PURLs (Persistent URLs) were developed by OCLC primarily to reduce the maintenance burden of the URLs contained in records catalogued in Intercat, an Internet Cataloguing Project to investigate and establish cataloguing practice and standards for Internet resources. It therefore has its origin in library applications.

OCLC was an active participant in the IETF working groups on URNs and were fully aware of the issues being discussed and of how far the groups were from consensus on a standard solution. They therefore developed PURLs as an intermediate solution to address the lack of progress in Internet naming. OCLC planned at a later stage to automatically translate PURLs to the eventual URN form.

Resolution of PURLs

Functionally, a PURL is a URL. However, instead of addressing directly the location of an Internet resource, a PURL addresses an intermediate resolution service. This PURL Resolution Service associates the PURL with the actual URL and returns that URL to the client software. The client then completes the URL transaction to obtain the resource. The transaction is a standard Hypertext Transfer Protocol (HTTP) redirect.

PURLs creation and maintenance

PURLs are created and maintained using a standard Web browser to access a forms based service hosted by the relevant PURL server.

PURLs are not updated automatically when their associated URL changes. PURLs must be maintained by using the PURL Resolver's maintenance facility to update the location of the resource. Only authorised PURL maintainers can modify a PURL.

Availability of PURLs

PURLs use standard HTTP protocols to connect to PURL resolvers and standard HTTP redirects to return information to the requesting client. They do not require the acceptance of new protocols or any modifications to client software.

OCLC has made available the PURL source code to enable institutions to make use of the technology. Since the introduction of the PURL model and services, a number of institutions have installed their own PURL servers, including the NLA.

1.4.3 The Handle System

The Handle system is a distributed persistent naming system developed for digital library applications.

The handle system was developed by CNRI (Corporation for National Research Initiatives) and had its origin in a computer science technical reports project, NCSTRL (Networked Computer Science Technical Reports Library), funded by DARPA (Defense Advanced Research Projects Agency) in the US. Part of this project was to develop an architecture for the underlying infrastructure of an open distributed digital library. The high level architecture developed by the project is described in a seminal paper *A framework for distributed digital object services* May, 1995 by Robert Kahn and Robert Wilensky and subsequently by William Arms in an article *Key Concepts in the Architecture of the Digital Library* in D-Lib Magazine, July 1995.

In addition to its initial deployment in the NCSTRL project, the Handle system is also implemented by the Library of Congress in its National Digital Library Program, and in the Copyright Office (CORDS) where handles are assigned to deposited digital objects. The Handle system is also used by the DOI system for resolution of DOIs and in the US Defense Technical Information Center.

The system supplies a location independent identifier and a resolution system which can be used over the Internet.

Syntax of Handles

The “handle” itself is a persistent identifier consisting of two parts. The syntax is as follows:

<naming authority>/<name>

where the naming authority is an administrative unit authorised to create and maintain handles and the name of the resource is a string which must be unique to that authority but which has no prescribed syntax.

e.g. nla.pic/123456t

The naming authority may authorise sub authorities. The sub-authority identifier is separated from the name of the higher authority by a full stop, e.g.

10.1045

Where “10” is the identifier for the DOI system and thus identifies a DOI, and “1045” is the identifier for D-lib magazine.

loc.pnp

Where “loc” is the identifier for the Library of Congress and “pnp” identifies the Prints and Photographs Division

Structure of the Handle System

The handle system is more than a simple naming scheme, it is supported by a resolution system consisting of a distributed system of global, local, and caching servers.

The handle system provides all the capabilities listed in *RFC 1737, "Functional Requirements for Uniform Resource Names*" described above (See Section 1.3.2).

The handle system consists of the following components:

Naming authorities are administrative units that are authorised to create and maintain handles. The name of a naming authority consists of one or more strings, separated by a full stop, as in the examples above.

The creation of naming authorities is delegated in a hierarchy. A Global Handle Registry run by CNRI registers the top level naming authorities, both to ensure the uniqueness of the names and to route requests for handle resolution. Each naming authority can then create sub-authorities, which in turn can have sub-authorities.

Handle servers store handles and provide a service to resolve them. Resolution of handles is carried out by handle servers, at the request of a client. There is a single Global Handle Server or Registry and many associated local handle servers maintained by naming authorities. The Global Handle Registry stores and resolves handles for all naming authorities and local handle servers and provides the client with information about which local handle server(s) can service a request. The Global Handle Server ensures uniqueness amongst handles stored at the global level.

Local handle servers are used by naming authorities to manage, store and resolve handles locally. These services resolve the requests and return the current address(es) or other information about the resource sought.

Caching servers associated with local servers allow frequently accessed handles to be resolved without need to request the address from the Global Registry.

Proxy servers allow Web browsers and other clients to resolve handles.

Availability of the Handle system

The Handle software is made publicly available and can be downloaded from the CNRI Handle site.

CNRI makes available local service software, client software and simple management tools, a caching handle server, tools for the creation and administration of handles and naming authorities, and a proxy server to enable Web clients to resolve handles. CNRI also provides a test server to allow trial resolution on non-critical data.

Use of handles can be through standard web browsers using a plug-in, or through unmodified web clients using proxy servers. Communication with the Handle System uses Handle System protocols.

The resolution protocol has a corresponding API available as a library of C functions. This library, also known as the client library, has been used by CNRI in the creation of an http-to-handle proxy server, handle-aware extensions to the Netscape and Microsoft web browsers, and a handle caching server.

In the absence of ubiquitous support for the Handle protocols in standard browsers, both LC and DOI use proxy servers to enable standard browsers to access handles.

For a more detailed and technical description of the system see *CNRI. The Handle System, Version 3.0 An Overview <http://www.handle.net/docs/overview.html>*

1.4.4 Digital Object Identifier (DOI)

The DOI was developed by the AAP (American Association of Publishers) as a response to the growing concern over rights protection in open electronic networks. It was officially launched at the Frankfurt Book Fair in 1997 and it has recently been registered as an American National Standard ANSI/NISO Z39.84-2000. The DOI is now the responsibility of the IDF (International DOI Foundation), a not for profit foundation established by AAP to manage the DOI.

The intention was to develop an industry wide, standard identifier that would be assigned to a work, or manifestation of a work, at the time of creation. However, the DOI is intended to be more than a simple identifier. It was developed as a first step in the creation of an infrastructure to support rights management in the digital environment. The identifier is intended to support systems to control transactions, to provide a key for rights management systems and to facilitate communication between publishers and their clients.

In order to achieve this, the DOI has been developed as a complete system. It consists of three components:

- The syntax and assignment rules of the DOI identifier itself
- A resolution system
- A metadata structure to unambiguously identify the item represented by the DOI

DOI syntax

The DOI is composed of a Prefix and a Suffix. Within the prefix are the Directory Code <DIR> and the Registrant Code <REG>. The suffix is named the DOI Suffix String <DSS>.

The syntax of the DOI is a handle syntax:

<Directory code>.<Registry code>/<DOI Suffix String>

As the maintenance agency (or naming authority), the International DOI Foundation (IDF) assigns the Directory code and the Registry Code. The Directory Code is currently always “10” for a DOI,

although additional codes may be assigned in the future, and the Registry code is the value of the code assigned by the IDF to a particular publisher, rights holder or registrant. The DOI Suffix String can be any string that meets the requirements of the registrant to whom the registry code has been assigned. It may include local identifiers or identifiers that are constructed according to other naming schemes, e.g. SICI, ISBN, PII. Examples taken from the NISO standard document indicate the diversity of numbering used in the local string.

DOI for the Authors' Licensing and Collecting Society's Byline service:

10.054/1418EC1N2LE

DOI (incorporating a SICI) from an article in the Journal of the American Society for Information Science, published by John Wiley & Sons:

10.1002/(SICI)1097-4571(199806)49:8<693::AID-ASI4>3.0.CO;2-O

DOI for an article from JAMA, the Journal of the American Medical Association:

10.1001/PUBS.JAMA(278)3,JOC7055-ABST:

DOI for the article "ABO Blood Group System" from Encyclopedia of Immunology 2e

Online, published by Academic Press:

10.1006/rwei.1999.0001

DOI for the article: "Digital Libraries and the Problem of Purpose" published in D-Lib Magazine

10.1045/january2000-levy

The system derives much of its flexibility from the fact that it uses the DOI as a "dumb" number whose only function is to link items to a database entry. It can thus accommodate any other identifier system used by the participating publishers without the need to re-number existing items. The local string may itself have meaning (e.g. a SICI), but that is of no significance in the DOI.

Resolution of the DOI

Unlike many other identifier systems, the DOI is supported by a resolution system. It is an application of the Handle system developed by CNRI. (See above Section 1.4.3).

A record of the DOI together with the current location of the object is registered in the DOI system server. This server acts as a resolver to the objects themselves, which continue to reside on the publisher's site or on a site licensed by the publisher. In the initial implementation, most DOIs resolve only to the publisher's site, however with full deployment of the DOI system and the

development of multiple resolution capabilities, and associated metadata, this will change. Changes of location are registered with the central server.

DOIs may be resolved through a native handle server for those with a handle enabled browser or through a proxy server which resolves to the handle server for those with standard browsers.

DOI Metadata

In order to achieve the functionality required by its developers, the DOI must be unambiguously associated with the item it represents and must be capable of being resolved to a number of different locations (multiple resolution). To facilitate automated selection and transactions, it must be possible to determine which work, version, or manifestation of an object is represented without recourse to inspection of the item. This can only be achieved successfully through carefully constructed metadata.

In addition, some basic metadata associated with the resolution process is necessary to support an intelligent selection between multiple locations for the same resource.

Thus, to support the development of the DOI in a move beyond its present state as a simple one on one resolution service, registration of a minimum set of standard metadata with the item will be required as an integral part of the assigning of a DOI.

For this purpose, the DOI has adopted the INDECS (Interoperability of Data in E-Commerce Systems) metadata set. The INDECS project, designed to provide a common metadata framework to support E-commerce in intellectual property, meshes neatly with the requirements of DOI. Eight mandatory elements are prescribed in the minimum, or kernel set of metadata, three of which (Genre, Type and origination) are to be derived from standard lists.

DOI governance and charging

The development of the DOI system has been the responsibility of the International DOI Foundation (IDF). The IDF is a not for profit organisation founded by the AAP, and its members are drawn from major publishers and software companies, as well as organisations which represent the interests of publishers and rights holders. E.g. International Publishers Association, International Association of Scientific, Technical and Medical Publishers, Authors Licensing and Collecting Society (ALCS), CSIRO Publishing, CAL, Copyright Clearance Centre, JISC, OCLC, and The Open University.

The DOI system is run on a cost recovery basis. It is an open system, free at the point of use and anyone may use a DOI to link to services. The DOI can be integrated into local environments and databases without charge. The cost of assigning a DOI is borne by the publisher or rights holder when it is assigned. It is intended primarily for use with intellectual property rights, but IDF policy does not rule out assigning DOIs to items in the public domain. The development costs are currently being borne by IDF membership fees.

The governance, administrative structure and charging mechanisms are in the process of change. The IDF is moving from an organisation supported by its members to one in which the day to day operation of the system is increasingly undertaken by a federation of registration agencies. Crossref was appointed the first Registration Agency in late 2000, and two others were in active negotiation in February 2001. It is expected that a number of additional DOI Registration agencies will be approved in 2001

Under these new arrangements, the IDF will remain a non-profit making maintenance agency, taking responsibility for the continued development of policies and standards. The IDF will also maintain the contractual relationships with Registration Agencies and with CNRI for the provision of Handle resolution services.

The role of the registration agencies will be to undertake day to day services to registrants, e.g. allocation of prefixes, registering DOIs and providing the infrastructure to allow registrants to declare and maintain metadata and state data (the current location(s) of the resource). Their responsibilities will also include quality assurance by ensuring that metadata is consistent and compliant and that state data is maintained by registrants and by providing adequate security for the registrants using the system. They are also expected to co-operate with IDF in the promotion and development of the DOI system.

The Registration Agencies are permitted to offer value-added services over and above the basic requirements (e.g. exploitation of registered metadata) and these services may be charged for, provided they do not conflict with the policies of the DOI as defined by the IDF.

Registration Agencies may be run for profit, but no mandatory model for the operation of a Registration Agency is laid down. They may charge for basic services (prefix allocation and DOIs registered) or cross subsidise these basic services from value added services.

Registration Agencies will have contractual arrangements with registrants and must provide consultative arrangements for addressing issues of common concern. Registrants may be affiliated to more than one Registration Agency.

IDF will remain a Registration Agency in the foreseeable future because it wishes to ensure that all markets are served.

DOI Fee Structure to December 2001

A Registration Agency must be a member of the IDF.

The annual membership of IDF is US$30,000. (No discretionary reductions are available for those subscribing in the Registration Agency category)

A franchise fee of 2 US cents per DOI registered is payable by the Registration Agency (US$20,000 per million DOI registered). A minimum franchise fee of US$20,000 is payable in first year regardless of how many DOIs are registered during that time. Payment of the franchise fee is to be

made on a 6 monthly basis.

Future developments for DOI of interest to libraries and archives

The work done on the DOI is impressive. The intellectual and analytical skills brought to bear on the task have ensured that consideration of the issues have been wide ranging and thorough. This has been essential for this kind of operation. Because it is operating in a commercial environment, the structure of the system must be watertight and there is no room for the kind of ambiguity engendered by the use of URLs as identifiers.

DOI has been, as far as possible, designed to conform to existing standards to maximise interoperability. DOI metadata is based on the INDECS metadata set and will be expressed in RDF using XML; the DOI itself conforms to the syntax for URNs laid down in *RFC 2141 URN Syntax* and has been registered as a NISO standard.

Many of the considerations of persistence and identification of electronic resources facing DOI are those facing libraries and archives, even though the latter are not operating in the full commercial paradigm.

The IDF's plans for full deployment of the DOI system include the refinement and mandatory implementation of metadata standards and the development of multiple resolution capabilities and the ability to link the declared metadata to the DOI. These are considered essential prerequisites to a successful digital rights management infrastructure. (And, I would venture to say, a successful distributed archiving infrastructure.)

Points of concern expressed re DOIs

There has been considerable discussion and concern expressed in many quarters about the direct relationship of the DOI with rights management and copyright considerations. It is seen by many as satisfying the needs of major publishers while not necessarily meeting the needs of rights holders and those interested in "fair use".

The majority of publishers are not in the same league as the major publishers and there are fears that the system may not be affordable for them. In addition to the charges that may be levied by Registration Agencies for the purchase of prefixes and assignment of identifiers, they may see participation and maintenance of the system, or the stringent requirements for the provision of metadata as a barrier.

Amongst libraries and academic institutions, there is the concern that because of the rights management aspect, full implementation and widespread use of the DOI may have the effect of making it easier to close off materials by making them unaffordable.

1.4.5 ARK (Archival Resource Key)

The most recently proposed identifier scheme is ARK (Archival Resource Key), which has been developed by John Kunze in his work for the US NLM (National Library of Medicine). The ARK proposal was issued as an Internet draft in March 2001.

The ARK is a scheme intended to facilitate the persistent naming and retrieval of information objects and is developed specifically to meet the needs of the custodians of archival digital objects. It is designed to be both location and protocol independent, while still retaining the ability to operate with today's protocols and tools, specifically with http, the DNS and current browsers. The total scheme comprises an identifier syntax and three services.

Further information can be found at http://www.ckm.ucsf.edu/people/jak/home/ark-01.txt

Syntax of the ARK

An ARK has four components with the following syntax:

<ark>:<NMAH>/<NAAN>/<Name>

Prefix "ark" denotes the scheme.

NMAH (Name Mapping Authority Hostport) is an optional and mutable component consisting of a hostname followed optionally by a colon and port number. The NMAH is a temporary address where ARK requests may be sent. It consists of an Internet hostname that has the same format and semantics as the domain name part of a URL. The significance of the NMAH is that it is not a part of the identifier and has no effect from the point of view of object identification. In other words, ARKs that differ only in the optional NMAH component are considered to identify the same object. For example, the following ARKs would identify the same resource:

ark:nla.gov.au/1703/1234567

ark:some.other.org/1703/1234567

ark:/1703/1234567

Because it is optional and changeable, over time the NAMH string may be replaced by a currently active service provider. Such a provider may be located using the NAAN by querying a global database to identify which NMAHs will service or resolve ARKs issued by that NAA. Meanwhile, in the current environment, a browser can create a valid http URI by substituting "http" for the "ark" prefix.

NAAN (Name Assigning Authority Number) is a registered number representing the authority that assigned the identifier.

The Name is assigned to the resource by the Name assigning Authority.

The NAAN and Name together form the immutable persistent identifier for the object.

ARK services

Like the DOI, metadata is a key component of the ARK proposal. Metadata provides the association between an ARK and the object it identifies.

Three ARK services are defined:

- request for the object itself (or a sensible substitute)
- request for the metadata describing the object
- request for a description of the commitment made by the Name Mapping Authority regarding the persistence of the object.

These services are defined initially to use the HTTP protocol, given that the www is the most widely deployed Internet delivery system. When the NMAH is specified, the "ark:" prefix may currently be replaced with "http://", to produce a valid URL that can gain access to ARK services using an unmodified Web client.

1.4.6 Other naming systems (SICI, BICI, ISSN, ISBN, etc.)

Existing naming schemes that are not designed specifically for use on the Internet are not examined here in detail. They are, if used in this context, most likely to be used in conjunction with database searching or another system that will provide the resolution. e.g. some publishers are using SICIs as the Namespace Specific String in the DOI. If these systems are registered as namespaces in their own right, as has been suggested in the case of ISSN and the ISDS International Centre, it may be possible to resolve them either via the registering agency, or by incorporation within another system.

Sources of detailed information on the standards are supplied in the Reading list, Appendix 5

Return to Table of Contents | Proceed to Part 2

About This Site | Copyright | Privacy | Accessibility | Site Map | Site Search | Content A-Z | Contact Us

PERSISTENT IDENTIFICATION SYSTEMS
Part 2 : Persistent Identifiers and the NLA

There are three aspects to the issue of persistent naming in the context of distributed digital information systems that must be addressed by the NLA:

- the semantics of the identifier itself.
- the issue of resolving the identifier to a resource or to further information on how to access the resource (metadata, another file, an html file etc.…)
- the issue of encouraging others to share responsibility for maintaining persistence of citations and to join a national system.

Of these, the more difficult are those relating to the issue of resolution in a distributed system. All numbering systems posited which match the URN requirements accept as a basic premise that the naming and assignment rules are the domain of the owner of the resource, or naming authority, and they prescribe no semantics other than a basic syntax to enable the number to be identified uniquely in a global context (and to give some clue as to how the identifier might be routed for resolution). Beyond that, all that is required is that the name be unique within the area of responsibility of the naming authority.

2.1 Requirements for a national persistent identifier system in Australia

In developing a national persistent identifier system, the following questions need to be considered:

- What are the optimal organisational arrangements for a national persistent identifier system?
- What should be the syntax of the persistent identifier?
- What are the options for resolution?

2.1.1 Organisational arrangements

Australia has a federal system of government with a traditional distrust of central control. It is therefore likely that a model in which responsibility for the management of the system is devolved to the owners of the resources is likely to be more widely acceptable than a fully centralised system,

provided that it is well supported and easy to maintain. Such a system also has the advantage that it encourages organisations and individuals to take responsibility for the persistence of the resources they own rather than fostering the assumption that once registered with another agency, all responsibility for the management of the resource is discharged.

It is therefore recommended that any model adopted by the NLA for a persistent naming service to Australian institutions should be a distributed one in which NLA takes responsibility for:

- Managing the naming system and registering participating organisations as naming authorities.
- Establishing, at an appropriate time, a resolving service that would redirect requests based on the persistent identifiers to the appropriate naming authority for local resolution by the naming authority.
- Providing the option of central resolution for those who cannot undertake it locally.
- Providing advice and support for users of the service.

Organisations wishing to participate in such a system would be responsible for:

- Registering with the central body at NLA.
- Identifying those resources they believe merit long term attention to maintain persistence of access.
- Assigning persistent identifiers to those resources.
- Establishing procedures to maintain a local resolving service to direct requests for resources they have named to their current locations.

2.1.2 Syntax and semantics of a national identifier

There should be no requirement for uniformity beyond that required for a URN, namely that the namespace identifier or naming authority be globally unique and that the namespace specific string be unique within an authority's namespace. Different organisations have different naming requirements for different categories of material and maximum flexibility and autonomy within the system is more likely to encourage early adoption.

2.1.3 Resolution of a national identifier

It is not desirable that all persistent identifiers in a national system are resolved centrally. Ideally, if NLA plays a role in resolution, it should be through a resolver discovery service redirecting requests for local resolution by the naming authority that named and manages the resource.

This will involve operating a registration system that will record the allocation of namespace identifiers together with a resolution system that will allow identifiers to be resolved into the resource itself or to meaningful citations.

2.2 Options for a [national] system

There is no obvious path for NLA to follow in the area of persistent identifiers. No identifier systems today will provide the functionality required for a national distributed digital archive, and any system implemented at the present time will have at least some aspects which make it less than ideal until current browsers are able to recognise and route location and protocol independent identifiers.

The issue of resolution of the identifiers remains problematic.

As stated earlier (See Section 1.3.3), currently there is no deployed resolution system for URNs, or other "name style" persistent identifier systems such as DOI or handles without the use of a proxy server. This means that identifiers such as "urn:nbn …." or "hdl:…" or "doi:10.1006/rwei.1999.0001" cannot be resolved unless the address of a proxy server is included as part of the identifier in the link. The result is identifiers such as "http://urn.nbn.org.au/urn:nbn:au.nla/12345 …" or "http://hdl.loc.gov/ …" or "http://doi.org/ …". This defeats the purpose of having location and protocol independent identifiers, since the protocol and address of the proxy server are present with the identifier when a link is created.

If a proxy server is not used, systems such as the Handle system and DOI, which have a built in resolution scheme, require the installation of a browser plug-in to enable a standard web browser to recognise the identifier and make the appropriate routing to a Handle server. Many users are reluctant, or do not have the confidence, to download and install plug-ins.

There are a number of options currently available to the NLA to implement a persistent identifier scheme. None of these are ideal for reasons discussed earlier, nor are any of the models discussed sufficiently compelling to totally exclude consideration of the others. The possible options considered here are:

- Use standard redirects
- Continue with a simple resolving system such as PURL
- Proceed with a URN style implementation using a proxy server
- Implement a handle server
- Adopt the DOI system

All of the options provide varying degrees of location independence, (i.e. the name of the item is unrelated to the physical location of the item) and persistence if the links are properly managed. It is important to note in this context that persistence is a function of organisation and management, not technology. Whatever system is used will require commitment to manage the naming and maintain the relationship between the name and location of the resource.

2.2.2 Use standard redirects

The simplest option is to use standard redirects to map all requests to the current location of a resource if it is moved from its original location. This is used at present on the NLA and many other websites.

Advantages

- It is the simplest to implement
- Standard redirects require no software changes or new standards.
- It is low cost and low risk

Disadvantages

- There is no indication from the URL that the link is a managed link that is maintained by an organisation with a commitment to persistence of access to the resource.
- It can supply one on one resolution only and cannot support multiple resolution
- It is tied to existing protocols
- It is a single site option and does not facilitate development of distributed naming systems
- Because there are no management tools, it can be difficult to keep track of redirects.

2.2.3 Continue with a simple resolving system such as PURL

The NLA has already installed a PURL server. As a national system, this has not been particularly successful, partly because it has not been widely publicised and partly because the publicity it has been given has been largely in the context of archiving material, rather than in the broader context of persistence of citation.

Advantages

- The Library already has experience with a PURL resolver system.
- PURL indicates commitment on the part of the assigner of the identifier to the persistence of the name, and by extrapolation, the resource.
- It is a simple solution with minimal investment. It uses standard HTTP redirects and URLs
- It can be deployed in a central or decentralised model with many organisations installing their own PURL server.
- PURL server has an advantage over a standard redirects in that it comes with a collection of utilities that support the management of URLs by allowing users to update their own resource, a critical aspect of assuring the integrity and maintenance of links.
- May be migrated to another URN system with the caveat that a dual system would be required indefinitely to support older PURLs.

Disadvantages

- Is not location independent
- Is protocol dependent - it is a URL implementation using standard http redirect.
- It can only support a one to one resolution as it is based on a redirect; it would be difficult to support one to many resolution, e.g. multiple instances or multiple versions.
- Standard browsers return the URL in the location field. This would in many cases become the citation unless the PURL is clearly indicated on or in the object itself or browsers are modified to display the persistent URL that was requested rather than the actual URL representing the

current location of the resource.

Use of the PURL option would be the simplest option and would require little effort beyond more active promotion and a closer examination of its use on the NLA site together with the development of implementation guidelines for the NLA site. If PURL were adopted, and the Library intends to provide a service for others, it might be worthwhile to generalise the domain name to be used for non-NLA resources by removing the “nla” and replacing it with something like “purl.org.au”. This may be more attractive to users.

If the NLA were to follow this path, it would also be worthwhile to encourage others to set up their own PURL resolvers to assist in managing their resources.

2.2.4 Proceed with a URN style implementation.

URN

As outlined earlier, there is no resolution infrastructure that can support URN or any other “name style” identifiers. If URN were implemented in the near to medium term, the identifiers could only be resolved using an http proxy server. Nevertheless, the use of URN style identifiers would have some advantages over the use of a URL as a persistent identifier.

Advantages

- Flexibility of the naming system.
- Independence of a particular protocol and location.
- A URN used this way could still refer to multiple instances of an item.
- Supports centralised and decentralised model.
- Can accommodate existing identifier schemes easily.
- Will differentiate from URL and address based schemes and implies commitment to persistence of access on the part of the assigner.

Disadvantages

- No significant deployment and no resolution service for URN identifiers.
- Uncertain deployment prospects (See section 1.3.4)
- Not accessible using standard browsers except through a proxy server.
- No browser plug-in available.
- No administration and management tools currently available.

NBN (National Bibliography Number)

If the National Library opts for an URN style implementation, it may feel committed to the use of the NBN, having agreed to support its registration and further development at a meeting of CDNL. There is some advantage to be gained from using the NBN:

- It identifies a namespace with the qualities required, i.e. it is maintained by stable organisations committed to the persistence of the resources for which they use the identifier.
- It fits within the URN syntax [but see below].

On the other hand:

- There is as yet (as with other URNs) no resolution service available for resolving the identifiers.
- The name National Bibliography Number is a little limiting and does not fit well with the full range of resources described by NLA. NPI (National Persistent Identifier) or NPN (National Persistent Name) might be a better option and more accurately reflect the range of resources which will be described using the identifier.
- The use of the namespace identifier (NID) “NBN” adds a level of indirection for what are essentially many separate namespaces defined by a combination of the NID and the country code prefix of the namespace specific string (NSS), which is designed to be transparent and resolvable locally. Resolution of the NBN to the correct local national library or naming authority will require some decision based on the country code in the namespace specific string itself. An alternative might be to register the names as separate namespaces or sub namespaces, e.g “urn:nbn.fi:fe12345”, “urn:nbn.au:nla.pic56789” “urn:nbn.lccn:200034567”. CDNL could still act as the “owner” of the extended namespace(s).

If the NBN is used how does it fit into the general requirements?

Bearing in mind the above, it would be possible to implement the NBN to accommodate the distributed service model envisaged for Australia. NBNs in Australia could be further divided by the name of the naming authority, e.g.

urn:nbn.au:nla…

urn:nbn.au:101…

urn:nbn.au:tas200…

The NLA could take responsibility for routing all requests for such names to the appropriate organisation for resolution.

The use of a proxy server would be required. The NLA would register a domain name such as nbn.org.au – to which all Australian NBNs (nbn.au) could be sent for redirection and local resolution.

2.2.5 Implement a Handle server.

The NLA has installed and trialled a Handle system and has determined that it is a technically feasible option. The implementation of a Handle system has the potential to meet the requirements of a distributed national system.

Advantages of the handle system:

- Flexibility of the naming system
- Naming system is URN compliant.
- Independence of a particular protocol
- A handle can refer to multiple instances of an item.
- Software and management tools are available from CNRI
- It was developed specifically for digital library applications
- Can support a centralised or decentralised model.
- Implemented in major applications which share the requirements of the NLA (Library of Congress and DOI)
- Easier interoperability with DOIs if the system takes off.
- Can accommodate existing schemes easily.
- Has a built in resolution infrastructure

Disadvantages of the handle system:

- More complex than the other systems to establish
- Not yet useable through standard browsers without plug-in software for the client - but can be used with a proxy server as an interim measure (See Section 1.4.3)

CNRI have stated that they have not registered the Handle system as a URI or URN scheme because the procedures for doing so are unclear and the relative advantages of registration as either a URI or URN are also unclear. One of the work items identified by a W3C interest group set up to look at URI issues is to clarify and streamline the URI administration and registration processes to ensure that they are easily negotiated and thus encourage the registration of additional URI and URN schemes. If this effort is successful, it may open the way for schemes such as Handles to be formally recognised and implemented by standard browser developers. On the other hand, Handles may be registered as a URN namespace, so that if standard browsers recognise URNs, they can be routed to the Global Handle registry. Either way, such a development would remove a major impediment to the use of Handles.

CNRI is a research initiative, which raises the question of support for the handle system if funding is withdrawn and the body ceases to exist. Given the investment of LC and the IDF (International DOI Foundation) in the system, it is unlikely that, should this happen, the system will suddenly die. The IDF has negotiated guaranteed levels of support and service level standards with CNRI, who have, as a result, moved support for the system to a commercial agency and now guarantee a 24 x 7 operation.

2.2.6 Adopt the DOI system

It would be possible for the NLA to become a user of the DOI system, either as a Registrant or as a Registration Agency (See section 1.4.4). The administrative, organisational and financial arrangements of the DOI system have recently been revised, and the exact long-term costs and requirements are not clear.

The costs associated with becoming a Registration Agency are specified until December 31st 2001, but thereafter will be subject to change. The costs of becoming a Registrant remain uncertain at this stage. As outlined in Section 1.4.4., there is no defined business model to which the Registration Agencies must adhere, and until a number of Registration Agencies are established, it is impossible to second guess what their charging mechanisms will be. One further option for the Library is to become a registrant directly under the IDF. Currently, this involves one time payment of US$1,000 for a prefix, which allows an unlimited number of DOIs to be allocated.

However, there is little advantage in the NLA becoming a registrant simply to register its own resources. The Library is quite capable of managing its own resources and creating its own identifier system and resolving its identifiers. Therefore the following comments on advantages and disadvantages are based on the scenario of the NLA becoming a Registration Agency

Advantages

- As a handles implementation, the DOI inherits many of the advantages of the Handles system
- Flexible naming system that can accommodate existing schemes easily.
- Has a built in resolution infrastructure
- Independence of a particular protocol
- A handle can refer to multiple instances of an item. (multiple resolution)
- Supported by freely available software
- Supports centralised and decentralised model.
- The real advantages of using the DOI for the NLA lie not in allocating DOI to its own resources, but in becoming a Registration Agency. The requirement to lodge metadata with the registered DOIs and the ability of Registration Agencies to define application profiles for that metadata would enable the NLA as a Registration Agency to build value added services on the metadata, e.g. to use the data as a component of a metadata registry. It would also acquire a great deal of knowledge of what was being published.
- The development of the system will be done co-operatively and many of the features needed by the NLA in a persistent identifier system (multiple resolution, automatic identification based on associated metadata) are being developed by the IDF for the DOI implementation.
- Involvement in a system which is likely to become widely deployed for digital resources of all kinds, especially in the commercial publishing fields. Many of the archival items the NLA will have custody of will have DOIs allocated.

Disadvantages

- The high costs of start up (Fees, negotiations, business plan, IT and business systems set up)
- Not yet useable through standard browsers without plug-in software for the client - but can be used with a proxy server as an interim measure (See Section 1.4.4)
- There could be a degree of loss of autonomy involved and a level of uncertainty about the degree to which the interests of large commercial publishers will impact adversely on those of archival institutions. On the other hand, close work with these organisations may lead to increased mutual understanding and co-operation.
- Charging structures may make it unacceptable for smaller publishers/website owners

- The link with rights management may make it unacceptable to some of the constituency that the NLA wishes to engage in the task. Many not-for profit organisations are suspicious of the commercial nature of the system and see it as the domain of the big publishing houses.

It is still unclear that NLA could apply DOIs to other people's intellectual property. Discussions with Norman Paskin indicate that there are differing views within the IDF on this issue and it is not clear from the DOI manual that these have been reconciled. The DOI Handbook (Version 1.0.0 February 2001) states:

"*Currently the rules relating to the registration of DOIs are relatively loose; there is an implicit assumption that any organisation that has acquired a DOI prefix is unlikely to register a DOI for an intellectual property entity in which it does not have a legitimate interest.*" -P. 39

This leaves the issue wide open. Unless the definition of legitimate interest is circumscribed, an archival institution could reasonably argue that it has a legitimate interest in any resource which it is archiving in the public interest.

The uncertainty concerning long-term costs and the governance issues associated with the establishment of a DOI registration agency make it an unattractive option at this time. It may be worth re-examining this option when the new charging structure is made available in January 2002. The pattern of activity and take up levels will also be clearer at that time, allowing for a better informed judgement to be made.

2.2.7 ARK Persistent identifier

If it is accepted as a standard, the ARK is a promising avenue for the identification of digital archival materials. (See Section 1.4.5)

Advantages

- It is designed for custodians of digital archival materials
- It is location and protocol independent
- Can be used with standard browsers today
- It can support multiple resolution
- Associated services are based on metadata

Disadvantages

- Too new to judge the likelihood of acceptance.
- Requires a resolution discovery service at a later date and there is no guarantee that this will be developed.

As the proposal was only issued in March 2001, it is too early to determine what will become of it. Nevertheless, it is worthy of detailed consideration and support.

The concept of making the Name Mapping Authority Hostport (NMAH) optional and mutable and recognising only the NAAN and assigned name components as the ARK identifier is elegant in its simplicity. If it works and is accepted, it can be used today and will only require minor adjustments to browsers to enable them to substitute the “http” string for the “ark” string for resolution. Development of the alternative resolver discovery facility based on the NAAN will take a little longer, but is feasible with support from archival institutions. The metadata required to support the services can easily be derived from the existing metadata created by custodial institutions for their internal management and access purposes.

The Library should encourage discussion of the proposal, possibly at CDNL.

2.3 Choice of a resolution system

There is an inherent risk in embarking on a particular course of action when the standards and processes involved are not settled and the various technologies are untested. The Library wishes to work within international standards and in cooperation with other agencies with like interests and needs. However the options at present are limited and any implementation that is more than simple URL management would need to be embarked upon with caution. This is not an area where a single organisation can successfully “go it alone” and “just doing something” is not a sensible way of proceeding.

As outlined at length in this document, there is no ideal solution at present. If we want to do more than offer managed URLs in the current environment, any system we use will have to operate using an http proxy server at least in the short to medium term. This reduces the value of the system, since all identifiers, and therefore citations, will effectively be URLs. Of the systems available at present, only the Handle and DOI offer a built-in resolution system.

The Handles test undertaken by the Library has shown that the system, although not well developed in terms of user interface, is a technically feasible option, and that by extension, the Library would be capable of operating a handle service as a DOI registration agency. Thus both the Handles and DOI options are technically feasible for the Library.

Given this, overall, the use of the Handle system seems at present to offer the best option. It offers a naming scheme which is flexible and extensible and which will support distributed naming and resolution with an option to resolve centrally. It has deployed systems in LC and DOI. It has the capability for multiple resolution which will be further developed and exploited by the IDF and we may be able to take advantage of theses developments. There is the possibility that it may be registered either as a URI scheme or a URN namespace. It also has management, security, registration and resolution capabilities.

However I hesitate to recommend the adoption of the Handle system until such time as it is registered as a URI scheme and is on the way to attaining recognition by standard web browsers. In addition, if the ARK is adopted and accepted it may be a better alternative.

Recommendation 1

That the Library postpone a decision on implementation of an external resolution mechanism for persistent identifiers and that, for the time being, persistence of our own resources be achieved through a combination of managed URLs and a resolver system based on directing query strings to our application delivery mechanisms for digital materials. (See Appendix 2)

That this decision be re-examined in the first quarter of 2002 in the light of developments in the intervening period.

2.4 Development of local and national naming systems

2.4.1 Identification of NLA resources

In the absence of a fully operational system of naming authorities, the most useful option for identifying NLA digital resources is to use a system of unique identifiers that are not dependent on a domain name to provide unique identification of the resource. Although not derived from an existing scheme, the identifiers should nevertheless be compatible with such schemes. The identifiers should be globally unique outside the context of a “container” such as a URL or a Handle or a DOI. They would uniquely identify the file or digital object and could therefore be used in a number of applications or contexts, including a URL. It would not rely on a domain name or a namespace identifier to make it unique, but at the same time could be used as the namespace specific string in the context of any naming system implemented at a later date.

A unique identifier of this kind could be used in the context of a number of identification schemes; in a straight URL, as a PURL, or as the NSS in a URN or as a Handle or an ARK, or some other scheme.

For example, the identifier “nla.pic-12345-t” could be used in the following ways:-

http://www.nla.gov.au/nla.pic-12345-t

http://diglib.nla.gov.au/nla.pic-12345-t

http://purl.nla.gov.au/nla.pic-12345-t

urn:nbn:au-nla.pic/nla.pic-12345-t

urn:<SomeOtherNamespace>:nla.pic-12345-t

hdl://1703/nla.pic-12345-t(or hdl://1703.001/ nla.pic-12345-t if we go for a national system)

http://hdl.nla.gov.au/1703/nla.pic-12345-t(if a proxy server is used)

doi://<NamingAuthority>/nla.pic-12345-t

http://www.nla.gov.au/<NAAN>/nla.pic-12345-t

ark://<NAAN>/nla.pic-12345-t

In all of these examples, the inclusion of a complete identifier means that the identifier will be unique prior to any decision on a global resolution system and the identifier is readily identifiable as the same in all these “containers” even automatically. Because it is a unique identifier, it could be used as a search parameter in its own right in search engines and databases.

A proposed system of numbering for the Library’s own digital resources which is currently being used in the NLA’s digitisation program is attached at Appendix 1.

2.4.2 Towards a cooperative national identifier system.

There is also a case to be made for the development of an identifier with these characteristics for general use by publishers and archival institutions. As described above, the identifier would uniquely identify a resource and could be used in a variety of contexts as a search key in databases and search engines and in URIs for access.

The Library should develop an identifier for digital archival resources in cooperation with other custodial institutions, e.g. the state libraries and archives. Cooperative development would assure a greater chance of acceptance and more effective promotion of the concept and the identifier scheme could, if carefully constructed, eventually form the basis of a national system of persistent identifiers.

An identifier developed for national use should follow the pattern described in RFC 1737. It should consist of a naming authority component to provide uniqueness and a local string which would be determined by the identifying institution to meet its own needs. The naming authority component would be registered centrally to avoid duplication, and the assignment of the local strings would be decentralised and undertaken by the naming authorities. (Centrally in this context would mean within a particular state if the naming authority substructure were delegated on a state based model). Use of this pattern would have the double benefit of guaranteeing uniqueness and facilitating identification of the naming authority, thus acting as a guide to possible resolution at a later date. Naming and resolution are separate issues in name style identifiers, but in the real world, hints on routing are needed. If the identifiers contain some structure that aids resolution, it increases the chances that future software will be able to make use of it.

It is also advisable that in a national system, the naming authority component be numeric, or at least have minimal semantics associated with it, to accommodate the frequent change in the names and relationships of government departments and authorities. It is anticipated that these bodies will be users of such a system. It is also recommended that the syntax of the identifier specify the separation of the naming authority component from the local string in a way that will allow machine identification of the two components, e.g. <naming authority>/<locally assigned name> or <naming authority> : <locally assigned name>. This will facilitate migration to another system at a later date if

one of them is implemented. In the short term, it will be accessed using standard URIs, i.e. http://<domain name>/<persistent identifier>.

Recommendation 2

That the Library formally adopts the naming guidelines outlined in Appendix 1

That the Library develops a naming scheme for archival resources in consultation with the state libraries and archives. The scheme should be designed for distributed assignment based on a pattern of naming authority and local string to guarantee uniqueness and that the syntax of the identifier separate the naming authority component from the local string in a way that will allow machine identification of the two components. This will facilitate migration to a global system at a later date if one of them is implemented. In the short term, it will be accessed using standard URLs.

2.5 Continued monitoring of developments

In the absence of a recommendation on resolution at this time, there will be a temptation to try to simplify the persistence issue by treating the problem of resolution and access and the issue of identification as separate issues and concentrating only on identification. The development of a naming system as described may be seen as the solution. However, to ignore the issue of resolution would be to reduce the value of the development of a naming system. The aim of persistent identification is to reduce failure of access over time and, as stated elsewhere, this cannot be achieved simply by assigning unique identifiers. Thus, the issue of resolution for location and protocol independent identifiers must be addressed, or at the very least be taken into account when identification schemes are considered.

2.5.1 Developments of particular interest

It is important that in pursuing the development of a cooperative national naming system, the Library should not consider the task complete. As stated earlier, an identifier system alone will not reduce access failures over time. It will be advisable to assiduously monitor developments in the next year or so in the area of resolution and access with a view to implementing a more location and protocol independent resolution and access system if one should emerge as the front-runner.

In particular, the Library should note:

- Whether a W3C working group on URI issues is established. (The notional time frame for completion is 2 years!). There is some doubt as to whether this will happen, as there is no real commitment from W3C to support it. Most of the interest is coming from the IETF members of the Interest Group. The Library should join W3C working group if one is established and support rationalisation of URI and URN procedures to assist acceptance of standards and development of URN enabled browsers.
- The progress of DOI deployment and the establishment of Registration Authorities. In particular, it should note whether the DOI begins to take off rapidly and whether there is a

registration agency established in Australia. It will be a question of judgement as to whether the NLA should apply to become a registration agency. Technically, the Library can support a Handles service, but the cost of establishing a Registration Agency is high. It would also be worthwhile to make use of this period to canvas the The progress of DOI deployment and the establishment of Registration Auth

- The Library should also monitor the progress and possible deployment of the latest IETF proposals on the Dynamic Delegation Discovery System (DDDS) issued as an Internet draft in February 2001. The DDDS approach is to use the DNS to locate "resolvers" that can provide information on individual resources, potentially including the resource itself. If this becomes a standard, it may be possible for an interested group of custodial institutions to support resolution of their identifiers using the domain name sThe Library should also monitor the progress and possible deploy
- Monitor ARK developments. As outlined above (See Section 2.2.7), the ARK has the potential to be a significant development for libraries, archives and other custodial institutions. Although it is too early to tell whether it will be accepted and implemented, if it is, it will meet the needs of the Library for an identifier which is location and protocol independent and which can be resolved by whatever current standards are commonly in use.

Recommendation 3

That the Library actively monitor developments relating to existing identifier schemes, in particular, the issues outlined above.

2.5.2 CDNL Group

The library may wish to revisit discussions in CDNL. There are two fronts on which CDNL might make progress.

Firstly, there is no reason why CDNL should not support a number of “approved” and registered namespaces that might be used by cooperating national libraries and this would offer a more flexible approach than supporting a single namespace (NBN). CDNL should look to supporting a number of URN namespaces used by National Libraries. For example, LCCN might be registered as a separate namespace and possibly resolved to a number of institutions e.g. to a Z39.50 query on the LC database, OCLC or on Kinetica, all of which have LC tapes loaded. The Australian system recommended above (Section 2.4) might also be registered with its own prefix(es).

Secondly, CDNL needs to seriously consider work on developing shared resolving services for name style identifiers if it is to make any significant progress towards either a viable shared URN namespace, or the registration and use of other “name style” identifier schemes. This will not be an easy task, as it will require commitment of resources for the long term in an uncertain environment and possibly requiring contractual agreements. The CDNL group has in the past focussed on identification and paid little attention to the more difficult area of access. If it is to be productive, it must widen its focus to encompass the issue of persistent access and resolution.

In this context, CDNL might wish to discuss and evaluate the ARK proposal or explore the

possibility of taking part in experiments with DDDS.

Recommendation 4

That the Library revisit discussions in CDNL with a view to:

i) affirming support for a number of identifier systems used by national libraries and.

ii) exploring the possibilities of developing and trialling shared resolution systems for identifiers used by national libraries and other custodial institutions.

2.6 Other related issues

2.6.1 Issues related to metadata and metadata structures

There are a number of issues related to the kind of metadata that needs to be stored with, or linked to, a persistent identifier. What information about the files, the objects they make up and the aggregates they are part of needs to be stored in structural metadata to support direct citation at the level of granularity sought and the selection of the appropriate copy or version? How can versions and relationships be recorded and made available to the user? How can selective resolution be supported? How can a user discover if a particular work has a persistent identifier?

It is worth noting in this context that both the DOI and ARK recognise the importance of a minimum set of metadata associated with the identifier and the resource.

What to resolve to?

Many of the aggregates to which we will wish to apply a persistent identifier have no tangible manifestation, i.e. a virtual digital object made up of other digital objects. A manuscript collection is made up of hundreds of page images that will be organised into documents, series, or collection; there may be several images which make up a map. Some form of meta-object or structural metadata would generally represent such aggregates. However to assist direct citation, the persistent identifier needs to be linked to the logical object. These aggregates might be dealt with in various ways. e.g. the persistent identifier might resolve to:

- A software program which can parse the identifiers (if intelligent numbers) and create an object for display which represents the serial issue, diary, letter, etc to which the persistent identifier refers.
- As above, using structural metadata from which an object can be created for display, including links to the objects making up the object
- An object such as an html page
- A point within an EAD finding aid.
- A metadata description with links to the sub-objects.

Decisions on this are problematic for some digital surrogates. Although the structure of the aggregates and objects are static and evident at the time of creation of the files, the ability to directly cite groups of files will be dependent on metadata.

For compound objects on the www server and in Pandora, this is initially simpler, since in most cases, the digital objects are self-referencing, i.e. there is an object (html page) which represents the logical collection containing links to the various files and this object can be addressed by the persistent identifier.

Identification of the persistent identifier

The full value of persistent identifiers to ensure persistence in citations can only be realised if they are actually used by people when citing documents/objects. In order to use them, people must know that there is a persistent identifier and must be able to discover what it is. At present this is difficult because of the nature of the redirects used in most existing systems.

For example, the PURLs on the items in Pandora are not evident. They are present in the NLA catalogue and Kinetica and can be found by looking at the links on the alphabetical listing on the site, but if an item had been found by following a reference from another object or by serendipity, it is unlikely that the user would find the PURL, and would cite the URL found in the location box on the browser.

The same scenario would apply to any object accessed via a proxy server while standard browsers display the actual URL in the location window and print the actual URL when documents are printed rather than the target URL.

Possible ways of facilitating lookup need to be explored. Possibilities include inclusion of the information in the document/object itself. “This document has a persistent identifier and should be cited as …/be linked to as …” This would be possible in pages which are constructed on the fly and pages owned by the NLA, but would be difficult in digital objects hosted for others (Pandora). Other possibilities are allowing the user to do a reverse lookup on the resolver database, i.e. do a search on the URL currently displayed and be given the persistent identifier currently associated with it. Metadata associated with the persistent identifier to allow remote lookup can be facilitated by a metadata database.

Versions and relationships

Many relationships can only be recorded in a way which is actionable and meaningful using metadata. Appropriate metadata will be required to select between versions. E.g. is X the same work as Y even though it is in a different file format? And where should the relationship be recorded. The information may be needed by users, presentation software, or those responsible for preservation of the resources.

Selective resolution.

Persistent identifiers may represent multiple copies of a resource held in different locations. Often decisions are required to determine the most appropriate copy. There are a variety of conditions on which a decision may be based including:

- Access privileges, e.g. a user may have the right to access a local copy held by an institution to which she is affiliated
- Choice between free and paid copies
- Format, e.g. text or pdf or word document
- Location, e.g. a local mirror site may be preferred
- "Live" or archival copy. The user may be directed to the publisher's site for more recent material, or to the archive for older material that may no longer be available on the publisher's site

In many of these cases, decisions can only be made based on local conditions e.g. an organisation might use a proxy server to route all requests for a particular item to a local mirror site or direct all requests from its clientele for a paid service to which it subscribes via whatever mechanism is used locally to access that service. All requests for archived copies may also be directed to the publisher's site this way.

Although it is not an immediate concern, it may be worthwhile revisiting metadata standards used in the Library to evaluate how the structural, descriptive and administrative metadata will support various functions related to the uses envisaged for persistent identifiers.

2.6.2 Relationships with publishers' data

The Library must also keep a watching brief on the progress of the DOI system from a user perspective. If the IDF plans are successful, it is likely that major publishers will begin to use the system seriously. There are already signs that the publishers see citation linking in serials databases as the "killer application" which will see the DOI take off. This is evidenced by the development of services by Cross Ref, and the fact that Cross-Ref has recently become the first DOI Registration AgencyThe

The full implementation as envisaged in the IDF documents will take time to achieve because of the complexity of the applications needed to support fully fledged rights management systems, but libraries will need to be able to deal with DOIs as identifiers of materials deposited in their collections.

It is too early to say how this will be done, since it is not clear what publishers will require and how rights management systems associated with DOI will work over the networks. Possibilities include using the multiple resolution capabilities of the DOI system to record the archiving institution as a second location for deposited material and possibly at a later date the only one if the publisher decides to abandon its interest in it. Access restrictions for items held locally can be enforced with a proxyIt

The IDF have stated publicly that the DOI may be used in local systems and databases (in much the

same way as we use ISSN and ISBN). In the meantime, if items with a DOI are encountered the DOI should be used as an identifier of the item if it is an exact copy or at the very least be recorded as an identifier associated with the item.

This is equally true of other forms of identification used by publishers of online materials. The relationship between the publisher's identifier and the archival copy of the resource must be recorded, maintained and made evident to users.

Recommendation 5

That the Library continue to address the outstanding issues of metadata, publishers identifiers and their relationship with archival copies and their identifiers.

Return to Table of Contents | Proceed to Part 3

About This Site | Copyright | Privacy | Accessibility | Site Map | Site Search | Content A-Z | Contact Us

PERSISTENT IDENTIFICATION SYSTEMS
Part 3 : The Achievement of Persistence of Access to Resources

A system of persistent identifiers is only one tool that can assist in the achievement of persistent access to resources. Good information management practices are a prerequisite for the effective use of persistent identifiers.

It is important to avoid instilling in resource owners and publishers the expectation that if they assign a persistent identifier, somehow permanence and persistence is assured. Persistence will always be a managed process, both at the global level in terms of the resolver services used to route requests based on identifiers to the custodians of resources and metadata, and at the local level to ensure that the identifier is associated with the resource and the request can be met. There is no system that can avoid a serious and sustained commitment to persistence by those organisations that wish to preserve public access to electronic documents.

Name persistence is not a technological problem. Some schemes will provide more reliable tools to achieve such persistence, but whatever system is used, persistence of citation will be achieved through the commitment of institutions responsible for assigning the names and maintaining the resolution services, not by the inherent qualities of the names themselves.

It is possible, and indeed desirable, in the absence of URN or name style identifiers to manage URLs better. They can be managed with a resolver database, with redirects, with PURLs, and this can be done now. The organisation and discipline required to do this is a prerequisite for implementing any other system of identifiers at a later time.

3.1 What categories of material need to be managed for persistence?

There are many categories of material that need to be managed for persistence. Some of these are obvious, such as material which is archived for the long term, others are less obvious, such as a web page which describes the opening hours of a library which may have been bookmarked by many users of that library.

The best rule of thumb is that any item which has been made publicly available online and which is likely to have been cited or referenced in other digital objects or online resources should be organised in such a way that citations to it persist. This should be done regardless of whether the item in question is subject to a formal archiving policy. Persistent and unambiguous identification for online digital resources currently made available on the web is essential, and for archival materials, this identification should also be location and protocol independent.

This is crucial for a number of applications:

Citation of online documents.

An increasing number of scholarly papers, research reports and significant documents and reports, including government documents, are available only online, mainly on the web. Many more are made available in both hardcopy and online form. These digital documents often cite other online documents in the form of links to those documents. This has two implications, firstly it is necessary to unambiguously identify the item referred to, secondly there is an expectation that the link will continue to be actionable while the document is available online.

There are many examples of documents cited in lists, references in papers and resources indexed in search engine databases which can no longer be accessed at the address cited. The researcher coming across such a dead end is faced with a number of questions Does the resource still exist?, If it does, where is it?, and finally, If I find something which looks like it, is it the same?

Archiving of digital materials

Persistent identification is also an essential tool in the archiving of digital materials. There is a strong association between the issue of identification and the preservation issue in the sense that persistence of identification is essential for reliable long-term access and is one of the main reasons libraries and archives have such a strong interest in persistent identification schemes. The value of preserving material is reduced if it cannot be identified reliably, found, and accessed when cited or referenced. It is therefore important that we have functioning identification scheme(s) that leave no doubt as to the commitment to persistence implied by the use of such an identifier.

Multiple copies

There are occasions when multiple copies of a resource exist at different locations. These may be identical and they need to be identified as such. This is not possible with an identifier based on location. Examples are mirror sites and archival copies and the original work, or a resource which may be made available via a web site and an ftp server.

Rights management/access management

Rights management and access management need unambiguous identification of a resource for machine processing, hence the publishing industry and those holding intellectual property rights in online resources also have an interest in persistent identification. Because automated, commercial

transactions may be involved, industrial strength persistent identification is critical. This is the reason the publishing industry have developed the their own identifier system, the DOI (Digital Object Identifier)

Catalogues, indexes and resource discovery services

Persistent identifiers are also essential for general linking in catalogues, indexes and gateways. In addition to being an irritation for users trying to access material, constant change and movements of material which change the identifiers of that material can impose an unacceptable maintenance burden on the maintainers of lists, indexes, gateways or catalogues of WWW materials. Libraries and archives therefore must have a keen interest in the development and implementation of permanent identifier systems that are independent of location. There is a general expectation that links in online documents and resources will work.

Bookmarks

Although all of the categories mentioned above may also be bookmarked in personal web browsers, there is a category of material which is frequently bookmarked but which may not be of long-term significance, e.g. the opening hours page of a library website, the What's new or events calendar of an organisation. While these may not have long-term value and will not usually be accorded archival treatment, any organisation with a service orientation would ensure that they are managed for persistence so that clients who consult them on a regular basis are not inconvenienced.

There is no doubt that an effective system of persistent identifiers will assist in the task of ensuring the long-term reliability of links. However the absence of a persistent identifier system does not mean that nothing can be done. Good data management and an understanding of the issues is a prerequisite for the introduction of a persistent identifier system. There are a number of positive steps that can be undertaken to prepare the ground whilst a system is being developed, including:

- Promotion of the concept of persistence of naming and the need to manage for persistence.
- Developing and promoting guidelines for managing for persistence
- Leading by example by developing a data management plan for the NLA website.
- Working with others to promote adoption and trial systems.

We have little control over what others do but there are many ways we can assist them in the development of best practice in the way they treat their information/resources.

3.2 Promote the concept of persistence

3.2.1. Encourage awareness

There is not yet a general awareness of the value of managing for persistence and the important role that can be played by the individual institution with a relatively small commitment of resources.

The NLA can encourage an awareness of the importance of managing for persistence as a precursor to archiving. Archiving can seem daunting because of the newness of the field and the complexity of the issues, many of which have yet to be fully appreciated and addressed. This can be discouraging to those who may feel that they do not have the resources or expertise to undertake the task. It is important to emphasise that there are many things that can be done now to achieve persistence of citation and to ensure that materials are identified for retention online. This will at least ensure that valuable information is not lost while the digital preservation field matures.

3.2.2 Develop guidelines for best practice management for persistence

Developing a set of guidelines for best practice management for persistence can assist in a variety of ways:

- Many people in large institutions do not have control over the organisation of the institutional website, although they may be the owners of some of the content and be responsible for other parts of it. Others may have outsourced the development and management of the site. The guidelines envisaged would assist those responsible for the content of the site to state their requirements for persistence and negotiate strategies for achieving it.
- Many of those responsible for their own sites do not have the time or resources to research and document the actions they might take. A set of guidelines that sets out the options may help this group of users to take the first steps.

These guidelines should cover consideration of the issues mentioned above, including data management policy, as well as hints on organising the site, file structures, etc. This is work that would have to be undertaken prior to the implementation of a persistent identifier system. It is likely that those already accustomed to managing for persistence will take the next step and use persistent identifiers. *[This has already been implemented in the form of Managing web resources for persistent access brochure]*

Recommendation 6

That the Library continues to promote awareness of the value of good information management practices, especially managing for persistence.

3.2.3 Lead by example

One of the tools for managing for persistence is the development of a data management strategy for online resources. In order to manage for persistence in the long-term, we need to be able to identify clearly the materials we wish to preserve access to, and which are also therefore candidates for persistent identifiers.

The library can lead by example in this process by developing an integrated data retention and access policy for our own resources, including not only Pandora and the digital surrogates, but also the website and TRIM. It is likely that a policy for the latter categories will be of greatest relevance to the

wider community. The policy for digital surrogates and the material in Pandora is fairly obvious and clear cut, but there are grey areas for materials which exist in multiple instantiations, e.g. items which exist in the records management system (TRIM), hard copy and online on the website. This is a situation which is common to many sites such as government and academic institutions which have significant information that is frequently cited.

It has been argued that material on the NLA website does not need to be regarded as permanent because the site is just a delivery mechanism and the preservation aspect is taken care of by the copying of the website mandated by Australian Archives and the routine electronic records management process, also mandated by Australian Archives. However, neither of these processes preserves persistent access to material that has been made publicly available online. Much of this material, especially position papers, staff papers and policy documents will be cited in other online documents, gateways and collections of resources online. These citations are made to the web version and if that version is subsequently removed from the site, links to it are broken. For this reason, such resources must be managed for persistence and are candidates for persistent identifiers.

Material cited on the website will fall into two categories:

- static resources
- dynamic resources (note that databases, e.g. RAAM, ALG, are not considered here as they are not candidates for persistent identifiers, except at the service entry points and help pages.)

Static resources include items which have a definitive version and will not be altered, e.g. a policy document, an annual report, a conference paper, all of which reflect the point of time at which they were created. Most of this category will be managed for long-term persistence of access.

Dynamic resources are materials that undergo change, either incremental change by addition of content or change by modification, deletion and insertion of content. They can be either ephemeral or archival.

The ephemeral category includes What’s new, What’s on, opening hours, contact information, lists of current publications, charging policies, etc. These are not likely to be cited in other objects, but are frequently bookmarked for regular use and users want the latest version. Old versions are of no particular interest and the normal records management and archival policies can take care of the information needs of the future researcher and policy maker. This category needs persistence of citation, but only for the current version.

The archival category includes material that will be preserved and guaranteed persistent access. The decision about what level of persistence of access to provide will be a matter of judgement and will depend on the nature of the material. E.g. for an electronic journal such as Gateways that operates in an incremental fashion, it may be necessary to provide consistent access to both the home page and to individual issues and articles for citation purposes; whilst for a publication which may be modified by change, addition and deletion, decisions will be needed on whether to archive and provide public access to successive versions so that little information is lost and the evolving nature of the publication is documented, or whether to rely on the routine archival copying to provide future

access.

Other types of material can be left to the normal business process and require no management for persistence, e.g. Registration page for a conference, entry point for a service no longer offered (although while it is alive, it should be managed for persistence of access), temporary alerts such as the recent V8 car race information.

A detailed analysis and typology of the information on the site and policies on the identification and organisation of the material will enable those responsible for managing it to decide at the point of creation of a document what level of persistence is required and it can be located, flagged and named accordingly. To all intents and purposes, some of these access copies of material will be treated as de facto archival material even though the official preservation copy may reside in the records management system.

Management for persistence of access is an important service to the community served by a website.

Recommendation 7

That the Library develops a data management and archival strategy for its own online resources, addressing specifically the relationship between the routine management and archiving of electronic records as part of the business process and the requirement to ensure persistent access to certain categories of resources which have been made publicly available on the website.

3.2 4 Work with others to promote management for persistence and development of PI systems.

There are a number of organisations that have a common interest in the achievement of persistence of citation for significant information sources. The most obvious of these are archives and libraries, particularly, but not limited to, state and academic libraries.

National Archives of Australia has responsibility for the preservation of government records. It discharges this responsibility by requiring agencies to preserve electronic records and by periodically copying their websites. There is, however, no requirement that documents that have been made publicly available online remain accessible so that links to them remain actionable. This is a serious deficiency, given the significance of many government reports and papers and the value of much government information to researchers and policy makers.

The old AGPS deposit scheme operated by the Commonwealth government deposited copies of Commonwealth government publications in National and State libraries to ensure widespread access to government information. It could be argued that in the online environment, this aim can be achieved by requiring not only the preservation of documents which have appeared on websites, but by mandating that websites be managed to ensure that important reports and documents remain accessible online and be managed for persistence so that citations to them are remain actionable. The NLA and National Archives of Australia should work together on this. The National Archives are responsible for the preservation of electronic resources that form the record of the business processes

of Commonwealth government agencies and the National Library for the preservation of the published record. The convergence of the two functions in the electronic environment has blurred the distinction between them. There is a need to address and agree upon processes and strategies for achieving persistence of access to government electronic resources published on websites.

The state libraries and academic libraries also have a vested interest in persistent access to online digital resources and may be willing to work with the Library. CASL and CAUL may be vehicles for publicising and addressing the issues within their respective communities of interest.

Recommendation 8

That the Library work with other interested bodies, in particular custodial institutions and publishers, to promote the use of persistent identifiers for online resources.

4. Summary of recommendations contained in the report

Recommendation 1

That the Library postpone a decision on implementation of an external resolution mechanism for persistent identifiers and that, for the time being, persistence of our own resources be achieved through a combination of managed URLs and a resolver system based on directing query strings to our application delivery mechanisms for digital materials. (See Appendix 2)

That this decision be re-examined in the first quarter of 2002 in the light of positive developments that may have taken place in the intervening period. (Section 2.3)

Recommendation 2

That the Library formally adopt the naming guidelines outlined scheme in Appendix 1

That the Library develop a naming scheme for archival resources in consultation with the state libraries and archives. The scheme should be designed for distributed assignment based on a pattern of naming authority and local string to guarantee uniqueness and that the syntax of the identifier separate the naming authority component from the local string in a way that will allow machine identification of the two components. This will facilitate migration to a global system at a later date if one of them is implemented. In the short term, it will be accessed using standard URLs (Section 2.4)

Recommendation 3

That the Library actively monitor developments relating to existing identifier schemes, (Section 2.5). In particular:

- Possible establishment of a W3C working group on URI issues. The Library should join W3C working group if one is established and support rationalisation of URI and URN procedures to

assist acceptance of standards and development of URN enabled browsers.

- The progress of DOI deployment and the establishment of Registration Authorities. It would also be worthwhile canvassing the views of publishers, libraries and archives on the issue of use of DOI.
- The progress and possible deployment of the latest IETF proposals on the Dynamic Delegation Discovery System (DDDS) issued as an Internet draft in February 2001. The DDDS approach is to use the DNS to locate resolvers that can provide information on individual resources, potentially including the resource itself. If this becomes a standard, it may be possible for an interested group of custodial institutions to support resolution of their identifiers using the domain name system as a resolver discovery service
- Monitor ARK developments. The ARK has the potential to be a significant development for libraries, archives and other custodial institutions.

Recommendation 4

That the Library revisit discussions in CDNL with a view to i) affirming support for a number of identifier systems used by national libraries and. ii) exploring the possibilities of developing and trialling shared resolution systems for identifiers used by national libraries and other custodial institutions. (Section 2.5.2)

Recommendation 5

That the Library explores the outstanding issues of metadata, publishers identifiers and their relationship with archival URLs (Section 2.6)

Recommendation 6

That the Library continue to promote awareness of the value of good information management practices, especially managing for persistence. (Section 3.2)

Recommendation 7

That the Library develops a data management and archival strategy for its own online resources, addressing specifically the relationship between the routine management and archiving of electronic records as part of the business process and the requirement to ensure persistent access to certain categories of resources which have been made publicly available on the website. (Section 3.2.3)

Recommendation 8

That the Library work with other interested bodies, in particular custodial institutions and publishers, to promote the use of persistent identifiers for online resources. (Section 3.2.4)

Return to Table of Contents | Proceed to Appendix 1

APPENDIX 1
NLA Guidelines for the Development and Application of a Persistent Identifier Scheme for Digital Resources

This document looks at

- the general considerations which need to be addressed when considering persistent identifiers for online digital materials.
- general principles guiding an NLA persistent identifier scheme
- specific application of those principles to the various categories of NLA digital materials and digital surrogates

1. Introduction

All resources housed in a computer system have one or more identifiers. A resource will have a system identifier which may be transparent to the user, and it may have a public identifier which may be available to a user to access the resource. In some cases, the system identifier will also serve as the public identifier. Public identifiers are used on the www to access materials directly via links in other objects or databases.

Identifier systems designed for digital materials in global distributed systems such as the www generally follow a pattern consisting of two parts, a naming authority which is registered as globally unique, and a local string which must be unique to the naming authority, but may take any form. Such systems are based on a recognition of the need for maximum flexibility in the naming of resources whilst at the same time standardising to facilitate interoperability. The only way to guarantee uniqueness in a federated distributed system is to include some form of unique name or identifier for the naming authority to supply that uniqueness. (This is the pattern recommended for URNs in RFC 1737 *Functional requirements for a Uniform Resource Identifier (December 1994)*

The patterns for some of the currently available naming systems are:

Scheme	**Pattern**

URI	most common pattern used is <scheme>://<authority>/<path> e.g. http://www.nla.gov.au/somewhere/something.html ftp://www.nla.gov.au/wherever/whatever.txt
URN	urn:<Namespace Identifier>:<Namespace Specific String> e.g. urn:issn:1035753X urn:nbn:fi-234567
DOI	<Directory code.Registrant code>/<DOI Suffix String> e.g. 10.1045/january2000-levy 10.1001/PUBS.JAMA(278)3,JOC7055-ABST:
Handle	<Naming Authority>/<Name> e.g. nstrl.cornell/TR93-1335 loc.pnp/12345
OAI	<scheme (always oai)>:<repository id>:<local string> e.g. oai:nla:123454321

This pattern is based on the premise that there is no right way to identify all materials under all circumstances and for all applications. As long as the requirement that an identifier be globally unique and persistent is met, any locally defined name will work. The issue for an organisation or individual wishing to adopt a system of persistent identifiers for their resources will be what else is required of that system and what other needs of the organisation the identifier is intended to serve.

2. General Principles

2.1 General requirements

The kind of structure described above allows maximum flexibility to an organisation in the local identification of its resources, whilst allowing the identifiers to be incorporated into a global system with the addition of an appropriate naming authority component.

At this stage, only URI schemes based on the DNS (Domain Name System) can be resolved through standard web browsers.

There are a number of qualities that will be required of a persistent identifier system regardless of the type of material being described. These are:

- Uniqueness
- Commitment to persistence
- Extensibility
- Flexibility
- Ease of use

Uniqueness

The needs of global uniqueness will be met by the naming authority component of the system adopted. However, identifiers must be unique within the context of the organisation's digital resources and therefore globally unique when associated with the name of the organisation or authority. Many organisations use hierarchical naming systems to achieve uniqueness.

Commitment to persistence

An organisation must have a commitment to maintain the association of the current location of the resource with the persistent identifier. It is important that a resource with a persistent identifier should not be moved, or removed without updating the location information associated with the identifier.

Extensibility

The system must be extensible and capable of accommodating all resources which require an identifier.

Flexibility

An identifier system will be more effective if it is able to accommodate the special requirements of different types of material or collections. A one size fits all approach is not always the most sensible. The appropriate level of intelligence to support processes and systems for particular applications may be incorporated where it is useful and omitted in cases where it serves no purpose.

Ease of use

Although not absolutely critical, and not essential for machine generated persistent identifiers, a system will generally be more successful if it is easy to understand and apply, and if it lends itself to short and easy to use citations. It is thus advisable to avoid long and unintelligible identifiers where possible unless they are intended purely for internal use or machine processing.

2.2 Form of the persistent identifier

Within the framework of the general requirements listed above, an organisation may to use almost any form of identifier that meets its business needs and still maintain compatibility with standard naming schemes. Decisions about the form of identifier will depend on the types of materials described and the processes and system requirements associated with their use and display.

The following are some additional considerations that will need to inform the decision on the choice of the form of a persistent identifier:

2.2.1 Dumb number or intelligent identifier

A persistent identifier may be a random collection of characters with no associated semantics that contains no information about the object it identifies – a dumb number. On the other hand, it may have some intelligence, or meaning, built into it. This meaning may have varying degrees of complexity and specificity and varying degrees of risk associated with it. As a general rule, it is easier for humans to remember and use identifiers with built in mnemonics than a sequence of meaningless characters, although for machine processing this is irrelevant.

There are many reasons to build intelligence into an identifier:

- The identifier may be easier to cite and remember e.g. DOI: 10.1045/june2000-payette is easier to remember and understand than a meaningless string of digits.
- The identifier may assist in the checking of work from a digitisation contractor.
- The identifier may be parsed by presentation software to display related files.
- Intelligent identifiers are easier to make unique, especially amongst a variety of organisational units. Running numbers need to be carefully controlled and monitored across an organisation to ensure uniqueness.
- The identifier may include relational information whereby smaller components are identified by reference to larger entities or collections of which they form a part (See Section 2.2.2)

It is unwise to use information that is not static in intelligent identifiers. Use only information that is **forever associated** with the item, e.g. collection ID, ISSN, hierarchical information relating to a **permanent** collection, date of publication, name of author etc.

2.2.2 Relational information in a persistent identifier

Hierarchy and structure of collections

One of the most common uses of intelligent identifier systems in the library world is to incorporate relational information that reflects the organisation and hierarchies of digital collections or aggregates. Smaller component such as digital page images or lower level aggregates are identified by reference to the larger collections or entities to which they belong. This form of identification is especially useful for digital surrogates of physical collections, although it can also be used to reflect the hierarchy of born digital collections such as serial issues and articles.

In the identification of digital surrogates of physical collections, relational information in an identifier can have advantages:

- it preserves the relationship with the original collection or aggregate
- it assists the digitisation process by facilitating the tracking of files, checking for completeness and quality control. This checking is more difficult for large collections if the files cannot easily be linked back to the original.
- such identifiers can be interpreted by humans which makes for easier grouping and management of the files
- the hierarchies expressed in the identifiers may be parsable by software designed to

manipulate the files and generate displays for the user
- the system provides obvious identifiers for logical aggregates which have no physical manifestation, e.g. for citation of a series, or groupings of correspondence, or the files which make up a diary
- the collection can be navigated without software or an associated database

Relationship with the original

It is often useful to use existing identifiers associated with the original of a digitised work to derive identifiers for the digitised version. In addition to enabling unambiguous identification of the original work used to create the resource, it can facilitate the linking of finding aids and metadata inherited from the original source.

2.2.3 Support for machine processing

An intelligent identifier can facilitate processing, assist in management, or support applications such as presentation software, e.g. html pages can be generated dynamically using information encoded in the identifiers to retrieve and organise the files or, in the absence of metadata or a structural map, such information could enable all files belonging to a particular collection to be identified and listed.

2.3 Other considerations

Granularity

Decisions concerning the level of detail at which persistent identifiers will need to be assigned will depend on the perceived needs of the material. The granularity will be different for different applications and materials. For many purposes, needs can be met by citing a top level web page which serves as an entry point to a self referencing collection of web files, or by citing at the journal article, item or chapter level. However some applications such as rights management may require a finer level of detail.

All files may have a system identifier, but a simpler, easier to use identifier may be assigned as a public identifier.

Versions

Each version of an object will require a separate persistent identifier. A version may be different in a variety of ways. It may have different content, a different physical format or simply a different resolution in the same format. Whether and how to reflect the relationships between versions will be decided by the organisation. The relationship can be reflected in the identifier using some kind of version code or date or type code or in metadata.

Existing naming schemes

It may be necessary to accommodate an existing scheme. If this is the case, the existing scheme may be encapsulated in the persistent identifier.

Automatic assignment

If persistent identifiers need to be system assigned, there may be constraints that take precedence over other design considerations.

3. Naming scheme for the NLA

General principles

The NLA will apply these general principles:

- Identifiers are designed to be globally unique
- Identifiers will be compatible with a national system of identifiers
- Identifiers must be unique within the context of the NLA collections and systems
- Identifiers will be managed within the context of collection groupings, e.g. maps, web server, manuscripts, archival collection, music, both to ensure uniqueness and to facilitate processing.
- In the case of digitised collections, codes indicating derivative types will be consistent across collections and will be derived from a standard list e.g. (thumbnail, display, master, structural map, sgml file)(See Appendix 1a)
- Identifiers will be based on the individual requirements of specific categories of material

Collection Identifier

It is suggested that the identifier for each category of material commence with a collection identifier consisting of text **nla** followed by a subgroup, or collection identifier, e.g. **nla.pic**, **nla.ms**, **nla.map**, **nla.gen**, **nla.mus**, **nla.aus**, **nla.arc**, **nla.web**(?). Although this part of the identifier broadly accords with the type of material, it is primarily a designation of collection ownership and responsibility for the management of the collection. This will guarantee that the identifiers are unique both globally and within NLA collections and systems.

This identifier can be incorporated into one or more other schemes at a later date, but does not require such incorporation to make it unique.

4. Identification for specific categories of material

Broadly speaking, there are three main categories of material that the NLA makes available online and which require persistent identification:

- Digital surrogates of hardcopy materials
- Materials created digitally and hosted on a website
- Materials hosted on behalf of others – digital archive

4.1 Digitised materials or digital surrogates

Each digital file created needs to be individually identified for processing and management.

The general principles for creating identifiers for digitised materials or digital surrogates are as follows. Identifiers should:

- be globally unique even in the absence of a national resolving system.
- incorporate aggregate and hierarchical relationship information (see section 2.2.2)
- be based on an existing local or public identifier for the original if one exists (e.g. local identifiers such as accession number for maps or manuscripts; public identifiers such as Ferguson number, ISBN or ISSN or Kinetica number for published materials)
- Identifies the purpose of a file (display, unstructured transcript, intermediate object, generation, etc.)

4.1.1 Collections of original, unpublished and rare materials

This category includes manuscript collections, oral history collections and some groupings of pictorial material.

Although individual items held in collections may have intrinsic value, many items derive their research value from their context within the collection to which they belong. Researchers may also cite them in the context of the collection. It is therefore important that any digitisation of collections of this kind be undertaken in such a way that the organisation and relationships that are present in the original collection can be reproduced for browsing and display purposes. (Many major digitisation projects use systems of this kind e.g. the National Digital Library Program of Library of Congress and RLG Studies in Scarlet).

For the reasons outlined in Section 2.2.2, the form of the identifier for materials within a specific grouping of material is based on an identifier for the group as a whole. If the collection already has a well-established unique identifier, it will be used as the collection identifier. This will assist in the relating of bibliographic metadata and finding aids for the original work to the digital collection.

This identifier can be extended by adding a representation of the particular aggregate(s) within that collection, if applicable, e.g. the series number and item number in a manuscript collection or sub-collection identifiers in a pictorial collection, or a series in the Oral History recordings. The identifier can be further extended by representation of further aggregates, if appropriate (item within a series)

At the lowest level, the identifier for the item can be extended by sequential numbering for the page images themselves and an extension indicating purpose of the resource, e.g. thumbnail, master copy, display copy, sgml text version.

For numbered pages, a sequence number and a page number and coded information for features such as contents lists, title pages, etc. should be added.

4.1.1.1 Manuscript collections

The naming of files in the manuscript collection will mirror the organisation of the physical collection as exemplified in the finding aid. It will reflect the logical rather than the physical collection, i.e. it will be based on the series and items within the series rather than the physical arrangement in boxes or folders.

There is a need to preserve the relationships which exist in the collections in order to allow the user of the files to navigate across, up and down the hierarchies of the series, folders and items that make up the collection structure. There is also a requirement to be able to go directly to an individual file or an individual group of files making up a digital object or sub-object that represents a particular item or grouping.

Each file will have a persistent identifier. The pattern for manuscripts identifiers will be:

<collection id>-<collection no.>-<series no.>-<item no.>-<sequence no.>-< role code>-<generation code>

<collection id>: The collection id will be nla.ms for items in the manuscript collections

<collection no.>: The top level aggregate, is the collection or accession number, e.g. MS8822 which represents the Mabo papers. This number is used as the collection location and the identifier for the finding aid.

<series no.>: The series number will be represented by a 3 digit number, with leading zeroes.

<item no.>: The item number within the series, represented by a 4 digit number with leading zeroes.

<sequence no.>: The sequence of the image within the item, represented by a 3 digit number with leading zeroes

<role code>: a code drawn from a standard list indicating the role of the file

<generation code>: A two digit code representing the version of a resource that has been migrated from its original format.

e.g. nla.ms-ms8822-001-0001-001-m
(The master file for the first page of the first item in series 1 of the Mabo papers.)

nla.ms-ms8822-001-0001-002-d
(The file for the display image of the second page of the first item in series 1 of the Mabo papers.)

In addition, an identifier may also be assigned to each logical aggregate in the hierarchy. (e.g. the

collection, the series, the letter, the diary …). This aggregate may have no tangible existence as a digital object, and the use of such an identifier will depend on the application used for the presentation of the digital objects that make up the aggregate. For example it may be used to initiate the generation of a page or display of the items making up the item or it may be used to access metadata in the form of a finding aid to locate and display its component files.

e.g. nla.ms-ms8822-001-0001 refers to the first item in series one of the collection, for which there is no corresponding discrete file.

4.1.1.2 Oral History sound recordings

Sound files are generally cited and requested by the TRC number. The TRC number is an old and well-established number based on the original tape recording number. This identification number has been carried over into the naming system for the digital files as the collection is converted to digital form. TRC numbers are used as control numbers in the sound recordings database. As unique and persistent numbers already in use for digital as well as analog recordings, they are suitable for use in the online environment.

The TRC number consists of 12 digits organised into three groups of four in the pattern 0000/0000/0000. These groupings represent a hierarchy consisting of the collection number, the series within the collection and the tape number within the series.

The identifier for oral history recordings will take the form:

<collection id>-<collection no.>-<series no.>-<tape no.>-<fragment identifier>-<role code>

<collection id>: The collection id will be nla.oh for items in the oral history collections

<collection no.>: the TRC number of the recording.

<series no.>: a four digit number representing the series number within the collection, left justified with leading zeroes.

<tape no.>: a four digit number representing the tape number within the series, left justified with leading zeroes.

<fragment identifier>: formalised representation of a sound byte from the recording (format to be determined)

Note: When the details of the pilot for the delivery of sound bytes over the web are finalised, it will be necessary to examine how the fragment identifier will be further divided. Possibilities include notation indicating the starting minute and ending minute of the recording fragment within the larger recording, or the starting minute and duration of the fragment.

<role code>: a code drawn from a standard list indicating the role of the file, e.g. Transcript, master file.

<generation code>: A two digit code representing the version of a resource which has been migrated from its original format

e.g. nla.oh-trc0234-0003-0001-m
(Master recording of the first tape in series 3 of TRC0234)

4.1.2 Single items of original and rare material

This category includes maps, pictures, papers, unpublished books, and single manuscripts

If the item already has a well-established unique identifier, this will be used as the basis of the identifier. If the item does not have a suitable identifier, a running number or some other unique identifier will be used as the basis of the identifier.

4.1.2.1 Maps

The physical Rare Map collection is largely organised and named by collection or aggregate, e.g. Ferguson, Nan Kivell and Tooley. The pre 1900 Petherick and the remainder of the Library's Rare maps are designated simply the Rare Map Collection and are assigned a running number within that collection.

The items are numbered within the collections using a unique number prefixed by a collection identifier. The collection identifiers are:

F (Ferguson)
NK (Nan Kivell)
T (Tooley) and
RM (Rare Maps)

The collection identifier is prefixed by a general collection identifier MAP, (e.g. MAPF) to further identify the material as cartographic, since some of these large formed collections also contain textual and pictorial materials.

The identifier for maps consisting of a single image will take the form:

<collection id>-<item no.>-<role code>

<collection id>: The collection id for maps is nla.map

<item no.>: The item number for maps will be the map number used in the physical collection as described above.

<role code>: a code drawn from a standard list indicating the role of the file.

e.g. nla.map-f1234-d
(a file representing a display version of the Ferguson map 1234)

The identifiers for single maps consisting of multiple segments will take the form:

<collection id>-<item no.>-<segment position id>-<role code>-<generation code>

<collection id>: The collection id for maps is nla.map

<**item no.>:-**The item number for maps will be the map number used in the physical collection as for single image maps.

<segment position id>: A two character code indicating the position of the item in a table representing the whole item, as follows

cp 1 horizontal co-ordinate, alphabetic beginning with **a** for row 1

cp 2 vertical co-ordinate, numeric beginning with **1** for column 1

e.g

nla.map-t2345-a1	nla.map-t2345-a2	nla.map-t2345-a3
nla.map-t2345-b1	nla.map-t2345-b2	nla.map-t2345-b3
nla.map-t2345-c1	nla.map-t2345-c2	nla.map-t2345-c3

Identifiers for a set of nine images that make up a digital representation of map T2345.

<role code>: drawn from a standard list

<generation code>: A two digit code representing the version of a resource which has been migrated from its original format

4.1.2.2 Pictorial materials

The identifier for pictorial images will be based on the immutable, or Kinetica record number for the bibliographic description. Identifiers will take the form:

<collection id>-<item no.>-<role code>-<generation code

<collection id>: The collection id for pictorial items is nla.pic

<item no.>: the Kinetica number of the bibliographic record for the original work. In the case of a Bibliographic record which represents a number of images in a collection, the item no. will be made unique by the addition of e suffix,

e.g nla.pic-an1234567-t. (for a thumbnail of a single image)

or:

nla.pic-an456789-1-v
nla.pic-an456789-2-v
nla.pic-an456789-3-v
etc. (for view copies of multiple items represented by a single bibliographic record)

<role code>: derived from a standard list.

<generation code>: A two digit code representing the version of a resource which has been migrated from its original format

4.1.3 Copies of published textual and music materials

4.1.3.1 Music

Identifiers for published music will follow the pattern:

<collection id>-<work no.>-<sequence no.>-<page no.>-<role code><generation code>

<collection id>: The collection id for printed music will be nla.mus.

<work no.>:- If a published work has a unique number in the public domain that can assist in the identification of the manifestation represented by the digital surrogate (e.g. a Ferguson number, Kinetica number or an ISMN), the unique number will be used as the work no. If no such number exists, a running number, or dumb identifier will be used.

<sequence no.>A three digit number with leading zeroes representing sequence of the image in the work

<page no.>- A group of four characters, the first being an alphabetic character indicating a special feature of the page, e.g. title page contents listing or index, the last three being a three digit number with leading zeroes representing the page number of the image if it has one.

If a page is unnumbered, the page no. will contain zeroes.

If there is no feature, the first character will be zero

<role code> derived from standard list

<generation code>: A two digit code representing the version of a resource which has been migrated from its original format

e.g. nla.mus-12345-002-t000-v
(A display copy of an unnumbered title page that is the second image in the group representing music item 12345)

nla.mus-45678-007-0005-v
(A display copy of page 5 which is image no. 7 in the group representing music item 45678)

4.1.3.2 Monographs

Identifiers for published monographs will follow the pattern:

<collection id>-<work no.>-<chapter or section no.>-<sequence no.>-<page no.>-<role code>-<generation code>

<collection id>: The collection id for printed monographs will be nla.aus for Australian material or nla.gen for overseas material.

<work no.>:- If a published work has a unique number in the public domain that can assist in the identification of the manifestation represented by the digital surrogate (e.g. a Ferguson number or an ISBN), the unique number will be used as the work no. If no such number exists, a running number or dumb identifier will be used.

<chapter or section no.> A three digit number with leading zeroes representing the chapter or section number

<sequence no.>A three digit number with leading zeroes representing the sequence of the image in the work

<page no.>- A group of four characters; the first three being a three digit number with leading zeroes representing page number of the image, the fourth being an alphabetic character indicating a special feature of the page, e.g. title page contents listing or index. If the page is unnumbered, the page number will contain zeroes.

<role code>: a code drawn from a standard list indicating the role of the file

<generation code>: A two digit code representing the version of a resource which has been migrated from its original format

e.g. nla.gen-0457362546-002-043-039-d

(A display copy of page 39, which is part of chapter 2 and is image no. 43 in the group representing the book)

nla.gen-0457362546-002-044-000-d
(A display copy of the following page that is an unnumbered plate which is also contained in chapter 2 and is image no. 44 in the group representing the book)

4.1.3.3 Serials

As with other digitised collections, there is, for printed serials, a need to be able to preserve the relationships which exist in the printed product in order to allow the user of the files to navigate across, up and down the hierarchies of the individual issues and volumes which make up the serial title. There is also a requirement to be able to go directly to an individual file or an individual group of files making up a digital object or sub-object that represents a particular article or issue.

The persistent identifier for a serial file will reflect the place of the file in the structure of the serial and will include date and/or issue/vol numbering and page numbering.

The pattern for serial item identifiers will be:

<collection id>-<serial identifier>-<issue no.>-<issue date>-<sequence no.>-<page no.>-<role code>-<generation code>

<collection id> The collection id for published serials will be nla.aus for Australian material or nla.gen for overseas material.

<serial identifier>If a published work has a unique number in the public domain which can assist in the identification of the manifestation which is represented by the digital surrogate, (e.g. a Ferguson number or an ISSN), use that number as the item number.

If no such number exists, obtain an ISSN and register the title with the ISDS Centre. In the case of an overseas serial, ask the ISSN Agency to obtain an ISSN from the appropriate ISDS National Centre.

<issue no.> A six digit number, the first three digits representing the volume number and the second three digits representing the numbering of the issue

<issue date.> An 8 digit number in the form YYYYMMDD representing the issue date. A cumulative index will be identified by the year of issue and the code **in**

<sequence no.> A three digit number with leading zeroes representing the sequence of the image within the issue

<page no.> A group of four characters; the first three being a three digit number with leading zeroes representing page number of the image, the fourth being an alphabetic character indicating a special

feature of the page, e.g. title page contents listing, index. If the page is unnumbered, the page no. will contain zeroes.

<role code> A code indicting the role of the file derived from a standard list.

<generation code>: A two digit code representing the version of a resource which has been migrated from its original format

Comment:

Citation at the article level can be expressed in a number of ways:

- Citing the identifier of the first page of the article within the issue (can be accommodated by the format outlined above
- Assigning an article number within the issue (an addition to the format outlined above)
- Assigning some other identifier for the article such as author's name or initial letters of the words of the title.
- Use of structural mapping

4.2 Materials created digitally and hosted on a website

Persistent identification of materials hosted on a website is potentially a more complex issue than that of specific groups of digitised materials. Digitised materials are likely to be a self contained, homogeneous and static group, whereas the materials on a working and evolving website will be expanding, heterogeneous and dynamic.

The challenge for persistent identification scheme for websites is to provide for flexibility and extensibility without introducing unnecessary complexity.

It is not necessary to assign persistent identifiers to all materials on a website. Materials hosted on a website will generally fall into three broad categories:

- Transient or ephemeral materials that need neither preserving nor persistent identification. Such categories include notices of one-off events, old versions of programs of events, advertisements for materials available for sale, drafts made available for comment, etc.
- Dynamic materials which are amended or updated frequently which are not likely to be cited but which users may bookmark to enable them to consult the most recent version. This requires only the most recent version to be accessible via a the persistent identifier for the resource. On the NLA site, categories would include What's new, descriptions of services offered, resource listings, corporate structure of the library, contact listings, collection descriptions, navigation pages, etc. These items will Dynamic materials which are amended or updated frequently which are not likely to be cited but which users may bookmark to enable them to consult the most recent version. This requires only the most recent version to be accessible via a the persistent identifier for the resource. On the NLA site, categories would include What's new, descriptions of services offered, resource listings, corporate structure of

the library, contact listings, collection descriptions, navigation pages, etc. These items will be interesting to future researchers, but it is not necessary to meticulously save each version as they are modified. The regular snapshots of the website mandated by Australian Archives will suffice for preservation of superseded versions of this material. The library does not have at this stage dynamic files that would require sequential archiving outside this process.

- Static and permanent materials of research value. These are the definitive versions of a resource that reflect a point in time and will not be altered. Such resources include policy papers, annual reports, staff presentations, online exhibitions, publications such as Gateways, etc. This type of material is likely to be cited, or referenced, and it should therefore be persistently identified and its accessibility maintained.

Like most early websites, the material on the NLA website is currently identified using a URL based on its place in the directory structure. Care is taken to move material as little as possible, and the latest version of the website guidelines reflects the guidelines for managing for persistence.

Nevertheless, it would be worthwhile devoting some resource to a systematic analysis and classification of the information on the site and an identification of the files that need to be retained indefinitely and allocated a persistent identifier.

There is no need to take any action at all for the transient information on the web site. Its transitory nature means that once its useful life has gone, it can disappear.

Dynamic information, as described above needs to be accessible for long periods of time at the same location, but not indefinitely. It does not need to have a life beyond the current organisational and technical environment and therefore does not need a location and protocol independent identifier system. It is therefore a good candidate for well managed URL style URIs.

Static and permanent information of research value on the other hand needs to be identified and made accessible indefinitely. A system of persistent identifiers is therefore appropriate to this material. Ideally an identifier should be assigned when the material is loaded to the server so that citations to the material use the identifier as soon as the material is made public.

Much web material is self-referencing and therefore does not need grouping. Nevertheless, it might help to manage the files at a later date if files that are part of the same digital object are grouped or related by a persistent identifier.

Persistent identifiers for the archival component of the website should be assigned at the time they are made available online, according to the principles outlined in Managing web resources for persistence. They should be candidates for use of the identifier to be developed by the NLA in conjunction with the state libraries.

4.3 Persistent identification of digital archival material (Pandora)

Material destined for a digital archive requires persistent identification. By definition, it is valuable material, preserved for posterity and therefore needs to be accessible indefinitely via an actionable

identifier.

NLA's current digital archive was developed as an outcome of the Pandora project and consists largely of material harvested from selected websites. Many of the sites are harvested periodically and when this is done, all files are retained even though the content of many files remains the same.

A large number of files are harvested and the volume of files processed makes it impractical to manually assign identifiers to each component file. It is therefore necessary to automatically assign identifiers to each file.

Identifiers for materials in the NLA digital archive consist of a work identifier, the date of archiving the particular instance of the resource, and the host, path and file name of the original file. A combination of the work identifier and the date harvested relate the individual files that make up the particular version of the resource as it existed at that point in time. The work identifier relates the different versions harvested on different occasions and will be used as the identifier for the Pandora title entry page.

Specifically, the Pandora PIs take the form:

<collection id>-<work identifier>-<archive date>-<publisher's URI>-<generation code>

<collection id>: For the archival collections the collection ID will be nla.arc

<work identifier>: a unique number within the digital archival collections assigned to the parent work of which the resource is a component.

<archive date> the date the file was archived in the format YYYYMMDD.

<publisher's URI> currently the host name, path name and file name of the resource on the publisher's site.

<generation code>: A two digit code representing the version of a resource which has been migrated from its original format

The identifier for the Pandora title entry page will take the form:

<collection id>-<work identifier>

The above pattern results in a persistent unique identifier for each file which has the following properties:

- It provides a unique identifier which meets the essential need for managing files
- It provides the granularity of identification required for facilitating preservation activities
- It contains sufficient intelligence to enable a measure of grouping and relating of versions

without the availability of metadata for structural mapping

However, there remains the problem of specifying and recording the relationship between the original and the archival copy and giving practical expression to that relationship in the ability to translate citations made to the original work into requests for the archival copy.

The main problem facing an archive of resources that have not originated with the archival institution relates to consistent citation of the works archived.

Assuming that the publisher's site is the preferred site while it exists, it follows that the identifier used on the publisher's site will be the one used in references to the resource in other works, etc. Provided the publisher's site is extant, this presents no problem. However it becomes vital that the identifiers for the publisher's version and the archival version are linked, or are the same, if the full purpose of the archiving is to be achieved, i.e. that the resource remain available and recognisable in citations. There are a number of options for achieving this

- Through metadata
- Through multiple resolution capabilities
- Through redirection
- Through contractual arrangements or agreements between publisher and custodial agency

The particular method used to resolve this problem is not critical, however it is critical that the issue be tackled.

APPENDIX 1a
Role codes preliminary list

This is a standardised list intended for use across media types. New roles will be added as necessary

Code	Description
t	thumbnail
v	view
e	examination
m	master
a	archive
s	summary
st	structured transcript
ut	unstructured transcript
sm	structural map
fa	finding aid

APPENDIX 1b
Code list for features recorded for Music items

Highlighted codes will probably be most used for music items

C	**Front cover**
T	**Title page**
L	**List title page**
B	**Back Cover**
U	**Contents page**
I	**Index**
X	**Additional page/s inserted/pasted into item**
A	Additional title page
F	Inside (ie verso) front cover
E	End papers
M	Fly leaf
G	Inside (ie verso) Back Cover
K	Preface
R	Introduction
S	Port/s
Q	Leaf / leaves of Plate/s
W	Verso plate/s (reverse side of plate)
H	Illustration/s
D	Bibliography
N	Biography
J	Discography
V	Various covers (ie. the cover of the 2nd or extra copy of the same bibliographic item that is different to the cover of the item digitised. eg. different performer featured)
1	page 1
2 etc.	page 2 etc.
i	page i
ii etc.	page ii etc.

Music code list in alphabetical order

A	Additional title page
B	Back Cover
C	Front cover
D	Bibliography
E	End papers
F	Inside (ie verso) front cover
G	Inside (ie verso) Back Cover
H	Illustration/s
I	Index
J	Discography
K	Preface
L	List title page
M	Fly leaf
N	Biography
Q	Leaf / leaves of Plate/s
R	Introduction
S	Port/s
T	Title page
U	Contents page
V	Various covers (ie. the cover of the 2nd or extra copy of the same bibliographic item that is different to the cover of the item digitised. eg. different performer featured)
W	Verso plate/s (reverse side of plate)
X	Additional page/s inserted/pasted into item
1	page 1
2 etc.	page 2 etc.
i	page i
ii etc.	page ii etc.

NOTE: Advertisements are not being coded

Return to Table of Contents | Proceed to Appendix 2

APPENDIX 2
Proposed Use of an NLA Re-direct Service

1. Requirements of a Persistent Identifier scheme

The requirements for a long-term fully operable persistent identifier scheme are as follows:

- Naming scheme (including commitment to uniqueness and persistence)
- Location and protocol independence of the naming scheme
- A resolution process to resolve that name into a resource or metadata about a resource

The resolution process needs the following capabilities.

- Recognition of the naming system by standard browsers
- Some method of routing the identifier to the agency/server where the information or resource can be found.
- The resolution of the identifier to return the resource by the agency/server

In addition, identifiers for materials archived for long periods (created by one organisation and copied for preservation by another) must be “transferred” or routed to the new organisation for resolution to the archival copy when a publisher has given up responsibility for the work. There must also be a reverse look-up capability.

Most of the PI systems developed for use on the Internet currently fail one or more of the criteria above and there is no contender sufficiently ahead of the others to rule out of the others as candidates for implementation. Various institutions are taking a variety of measures to deal with the issue of persistence.

The most significant failure of the systems lies in the area of browser recognition and routing. The identifier has to be recognised, a server registered as capable of resolving the identifier has to be located and the identifier routed to that server in order for the server to return the resource or metadata. The first two steps require a high degree of interoperability and the widespread deployment of an infrastructure based on open standards. The method of resolution employed once a request reaches the host server is generally considered to be a local issue and is transparent to the requester.

Most systems which incorporate a resolution capability use a simple database to associate the Identifier with the actual location of a resource when a resolution request is made, e.g. PURL, Handles, DOI.

A proposal for a re-direct using NLA applications

The digital surrogates at the NLA are delivered by applications that are database driven. These applications have indexes in which persistent identifiers are linked to the locations of the resources. An application must always “know” where the files or digital objects are for the delivery of the service. Because they are the basis of service delivery applications, the databases will always be updated when resources move or applications change. They are therefore

also capable of doubling as internal resolver databases for servicing requests based on persistent identifiers.

Tony Boston has developed a proposal that this capability be exploited to resolve URL style identifiers for NLA resources. The proposal involves the development of a general server side program that redirects queries for a particular digital object to the appropriate application based on the type of object being delivered, e.g. if the item were a picture, the request would be directed to the Images1 database, if a manuscript, to whatever delivery mechanism is developed for manuscripts. The program would do this using information in the URL that would indicate the type of material the identifier was describing.

Advantages of the proposal

Low risk

The proposal is a low risk option for digital surrogates. It involves the development of a (relatively) simple program to parse the identifier so that a request can be directed in the form of a query string to the appropriate application. It is essentially a persistent URL using a different form of redirect.

Easy maintenance

The proposal would be easy to maintain. The applications are active delivery mechanisms for the digital resources and therefore databases must be updated when there is any change of application or change of location of the files. Minor changes only would be needed to the redirect programs if there is a change of application software.

Compatibility with future options

A complete unique identifier for the resources will be stored in the URL so that it is not dependent on the domain name of the URL to provide uniqueness. This ensures that it does not rule out later use in the implementation of another system such as DOI, Handles, ARK etc. Indeed, with a slight modification, it could be used to populate a database associated with one of the other options.

Speed of implementation

The proposal would be quick to develop for immediate applications for the materials being produced for the digitisation program.

No new database required

It does not require the creation of a new resolver database as would be necessary for the implementation of Handles or DOI or OCLC’s PURL.

Particularly suitable for digital surrogates

The proposal is particularly useful for digital surrogates for the following reasons:

- Large numbers of files requiring applications to associate files and display the objects. Displays are usually generated “on the fly”
- The service delivery applications are database driven and tightly controlled and managed. The resources are homogenous and created to uniform standards and with the same preservation status, i.e. if they are migrated to a different format, they will all be migrated at the same time
- They have no relationship with other versions outside the NLA

Disadvantages of the proposal

The system is a persistent URL and is therefore subject to the same disadvantages as a PURL. It is not protocol and location independent. The method could be implemented in a similar fashion to a PURL.

The URL is re-directed, so the URL displayed in the browser is the destination URL (a query string) rather than the source URL. This is a problem with all resolving services that are based on a re-direct, eg PURLs, Handles or DOI using a proxy server etc.

It does not work well for current, dynamic websites that are not database driven.

It is of doubtful value for Pandora or archives of resources which do not originate with the archival institution because it does not have a multiple resolution capability. It would be difficult to support one to many resolution, e.g. multiple instances or multiple versions.

It does not offer a basis on which to offer a national persistent identifier system. It is a single site solution dependent on the existence of a database application to access the resource. (Although it could be cited as an example of a resolver database and the NLA could offer to assist those who wished to follow the example.) It should be noted that this method is being proposed as a method of managed URL or persistent URL for NLA resources and is not being proposed as a universal solution to the general issue of persistent identification.

Comparison with other persistent identifier schemes considered

The following table is a rough comparison of the option with other persistent identifier schemes.

PID	Location Independent	Protocol Independent	Browser recognition	Routing	Relative cost	Relative risk	Multiple resolution
Proposal	No	No	Yes	http/DNS	Low	Low	No
PURL	No	No	Yes	http/DNS	Low	Low	No
Handles	Potentially	Potentially	No. Plug in	Handle servers	Medium	Medium	Potentially
DOI	Potentially	Potentially	No. Plug in	Handle servers	High	Medium	Potentially
URN	Yes	Yes	No	No	?	?	Potentially

Recommendation

That a general server side program that redirects queries for particular digital objects to be resolved by the appropriate application software be developed for the NLA's digital surrogates. The program will achieve this by parsing a URL style identifier.

[NB. This proposal was accepted by the Persistent identifier group, April 2001]

Return to Table of Contents | Proceed to Appendix 3

About This Site | Copyright | Privacy | Accessibility | Site Map | Site Search | Content A-Z | Contact Us

APPENDIX 3
Handles Test

1. Timing and background

1.1 Initial difficulties

The handles test took a much longer elapsed time than originally expected for a number of reasons unrelated to the basic functionality of the system. They related mainly to:

- Obtaining and installing desktop software (X-windows and Java RTE)
- Problems associated with the accidental loading of an outdated version of the software.

The Handles software was downloaded at the end of November and installed a few days later with relative ease. At the beginning of the test, X-windows software was required to enable DD to directly access the server the system from her PC. This required the installation of X-vision. The only licences for X-vision in the Library are for CJK use and are installed via PCMan. The version loaded via PC man requires the loading of CJK fonts.

The most time consuming initial hurdle was the inability of DD to authenticate herself on the system to initiate testing. Peter Newman was able to authenticate himself intermittently, but never on Diana's machine. Many frustrating hours were spent on this. Eventually Peter traced the transactions and discovered that whenever requests were going to a particular IP address, authentication failed. When this information was forwarded to CNRI, they realised that we were operating with an old version of the software that was addressing a server that was no longer operational. In the few days between the time the software was downloaded and the time it was installed, a new version of the software had been issued. Unfortunately, this was discovered just before Xmas. After Xmas DD was involved in writing and delivering a paper at the Information Online conference in Sydney. No further work was done on the system until the end of January.

In the meantime, the Helpdesk had installed Java runtime environment on PCman so that the Java client could be run on the desktop. Peter installed the client on Diana's PC and she began functional testing of the administration client at the beginning of February. After familiarisation, Julie Whiting

also had a look at the system.

1.2 The Handle system

A Handle System operation has two basic components, resolution and administration.

- Handle resolution is achieved by a software client submitting a handle query to a handle service and receiving back the set of handle values from the responsible handle server. These can be displayed or used to access a resource if a URI is one of the values returned.
- Handle administration allows users to manage handles, including adding and deleting handles, or updating their values, over the public Internet. It also deals with naming authority administration via naming authority handles.

Both types of operations may require authentication of the user via the handle system authentication protocol

2. Aims of the testing

The aims of the testing were to assess the functionality and the installation, support and maintenance requirements of the system.

Aspects of the functionality which were considered included

- The administration and change function
- Ease of use
- Adequacy of documentation
- Management and reporting functions

Aspects of the installation and support requirements of the system that were considered included:

- The level of experience needed to install
- The adequacy of the documentation to support the process and enable problem resolution
- The level of support required from CNRI
- The level of support that may be required from the NLA if the system is adopted and offered to others.
- Compatibility with existing systems and skills
- Efficiency of the caching software
- Ease of use of the API to facilitate the development of ancillary services such as batch loading procedures
- Performance (including, unusual or excessive resource requirements, platform requirements, disk and memory use)

3. Administration and change functions

The administrative and change functions were carried out using the Java client on a standard NLA desktop PC and using X-windows on the server. Java RTE (Run Time Environment) was incorporated into the PCman suite by the Helpdesk staff and loaded via PCman. Since the Java client is what users would be supplied with in an operating environment, the comments on the functionality that follow are based on use of that client.

3.1 Authentication

Each handle established requires at least one administrator. In order to create, modify or delete a handle, a user must first establish that (s)he is authorised to perform those functions under the naming authority.

The authentication system is not intuitive, and the available documentation on the authentication process is minimal. An element of confusion is introduced by the fact that administrators are identified by a handle. The exact procedures to be undertaken were not therefore immediately obvious. This led to the creation of a batch of handles that could not be changed because they had invalid administrators. The first handles created manually also had invalid administrators for a different reason (See 3.2). Unfortunately, handles with invalid administrators, cannot be deleted or amended, since the authentication process is dependent on the presence of a valid admin type to authenticate against.

Once worked out and understood, however the authentication and administration process is not a difficult one. With improved documentation and a small amount of training, these problems could be avoided in a production implementation. A suggestion will be forwarded to CNRI that a check be made for an invalid handle used as an administrator value.

3.1.1 Creating a vlist

Administration and authentication can be managed by way of “vlists” or groups of people identified as having a given set of privileges for a particular naming authority. This is the most sensible way to manage the administration function, since people can be assigned to the group or removed from the group and immediately acquire or lose the authority to administer the handles for which that group has administrator privileges. This is an extremely useful feature, as the creation of large numbers of handles with individuals designated as administrators would be impossible to manage.

3.1.2 Security

The administration of the system is secure and relies on the use of private keys that remain on the site, on the desktop of the user.

3.2 Creating, Modifying, deleting handles

The tasks associated with the creation, modification and deletion of Handles were straightforward once the administration and authentication issues were sorted out and the syntax of the HS-ADMIN

field understood.

The system prevents the creation of duplicate handles and responds with the message “Handle already exists”. It is possible; however, to recreate a handle after it has been deleted. It would therefore be better not to delete handles that are withdrawn.

It is also possible to use an invalid handle for an administrator and thus create a handle that cannot be changed or deleted. This happened in the case of the first handles created because the high level “1703.test” was used without inserting the “0.NA” prefix required for the top level naming authority when used without the Namespace Specific String (NSS)

3.2 Querying

Querying the attributes and values of handles is straightforward with correct authentication. Public read values can be queried without authentication.

3.4 Batch loading

A batch of handles for Pandora items was loaded with no problems other than that mentioned above relating to authentication. Because of a misinterpretation of the documentation, they were inadvertently created with their own handle as the administrator and could not be modified or deleted. This is not a problem that is likely to be repeated.

3.5 Listing Handles

The administration client has a function for listing handles, but in its present form it is very basic. It allows the listing of handles assigned under a particular naming authority and gives two options for output; output to the window on the screen, or output to a text file. There is no facility to sort the output and the list appears to be a random arrangement neither alphabetical nor arranged in date order. (Possibly the order stored on the disk?)

The only data output is the handle name, with no associated information. When output as a file, it is presented as a text file in a single string.

To be useful as an administrative tool, the listing would require, as a minimum, the ability to extract other associated information from the database (URLs, date, etc.) and the development of some formatting and processing software.

4. Documentation

The documentation is minimal and could be improved. It would be necessary to supplement it and an expanded set of user documentation with examples would have to be written.

5. Ease of use

We were looking at the system with no training and minimal documentation but after a few mistakes with authentication, managed to understand and operate the system. If a decision were made to install the system in house, it would be advisable to provide a training course such as those run by CNRI and IDF for DOI participants. This would avoid a few of the “traps for young players” which could be better documented.

There are a few minor irritants with the user interface which do not affect the functionality of the system but which can be frustrating and detract from the ease of use of the Java client. For example, before entering data it is necessary to click with the mouse in the box; using the browse option to locate the private key wipes out the text input into the handle box and it has to be rekeyed.

Use of the high level naming authority name without prefixing 0.NA/ results in an invalid handle and therefore unchangeable handles when it is used in the administrator field. There is no error message warning that the handle is invalid, even though the syntax could be easily validated on input.

6. Installation, support and maintenance requirements.

6.1 Loading the software

Installation was relatively straightforward with the exception of the initial difficulties relating to the outdated version of the software outlined above. Peter is a very experienced Unix programmer and is familiar with Java. A basic understanding of Unix and Java would be necessary for a system administrator.

6.2 Support

CNRI response to questions was good. They were prompt and informative. The software is straightforward and would be relatively easy to support in organisations with appropriate skills. It would, however, need improved documentation and some training to enable users to avoid or minimise errors and questions when starting out.

6.3 System requirements

An estimate of system requirements is as follows:

- 10Mb of disk storage for the application
- 6Mb of RAM to run the server process plus a further 32Mb in swap space
- Approximately 3Mb of disk storage for every 1000 handles put into the system and tapes needed to backup this data (Pandora currently there are approximately 600,000 to 1 million individual files. 1 million items would use approximately 3Gb of disk storage. Images 1 has 50,000 files requiring 50 Mb of disk storage)
- The java runtime 1.1.7 or 1.1.8 environment would be required

Staff resources needed for batch loading:

Once the standard format of NLA handles are decided it would take 2-5 days to write the batch file to load into the system and plus a further week to load them. For newly archived sites within Pandora it would be feasible to automatically load the details into the handle service as a part of the Pandora Collecting System post gathering process.

Similar times would be required to develop scripts for the batch loading of pictorial images and other digital surrogates.

7. General observations

The system has the basic functionality required for a PI system for a digital library and is perfectly adequate as a building block in a total system architecture. It is the only system available that currently offers resolution and multiple resolution capabilities.

It would however need a number of software developments, and considerable work to integrate it into processes, procedures, and organisational arrangements to make it a complete and useable system. This is especially true if identifiers are to be assigned automatically in batch mode for Pandora items and for digital surrogates. Handles for web documents would be allocated individually as they are created and mounted on the server.

Management reporting is very basic and the development of routines for the extraction of reporting information would be necessary.

Many of the NLA requirements are the same as those required by the IDF for the DOI system and may also be of interest to LC. For this reason, it is possible that NLA will be able to take advantage of these as they are developed, e.g. multiple resolution, association with metadata and reverse look up.

LC has indicated that they would, if the NLA decides to use the Handle system, be interested in cooperation in the development of some of the functionality required and in improving the user interface.

8. Conclusions

The system is straightforward and well supported. There is no doubt that the Library could support it either as a straightforward PI resolver system or as part of the business processes of a DOI registration agency. It is the best option currently available.

However, there are a number of factors that make it inadvisable at this time

- Investment required for the establishment of the systems and procedures and the extent of system modifications, including reporting mechanisms, that would be required.

- Many agencies that might be partners in a distributed system may not be ready at this time. UNIX and Java skills of a high order are needed.
- Still need proxy server to use with standard browsers.

Recommendations

That the Library not proceed with the implementation of the Handles System at this time for the reasons outlined above.

That the developments with DOI's use of handles for multiple resolution and the association of metadata with handles continue to be monitored.

Return to Table of Contents | Proceed to Appendix 4

About This Site | Copyright | Privacy | Accessibility | Site Map | Site Search | Content A-Z | Contact Us

APPENDIX 4
Managing Web Resources for Persistent Access

The success of a distributed information system such as the World Wide Web (www) for research depends on the long-term consistency of the interlinking between online resources and the persistence of the links that are provided in the catalogues, indexes and listings which constitute resource discovery services.

Links between resources can range from casual links inserted in a document because it is possible to do so, (e.g. a link to the site of an organisation because the name of that organisation is mentioned), to the formal citation of references, source documents, research results or to other related information that is an integral part of the resource. The category of link that represents formal citation is the most significant for the integrity of the intellectual heritage and it is especially important that these links continue to identify and provide access to the resources.

The persistence of links between resources and of links in resource discovery services is essential to ensure long-term public access to web based materials and is therefore an important aspect of the archiving and preservation strategies adopted for these resources. For information on how to manage web resources to assist later preservation action, see the companion document *Safeguarding Australia's web resources: guidelines for creators and publishers*

Creators, publishers, libraries and other information repositories all have a vital role to play in ensuring web based information resources can be located and accessed for use in the current environment and into the future. These guidelines are intended to assist those responsible for the management of published online materials to make sure links made to those resources continue to work.

1. Why links fail to work

Broken links are very common on the web and few people will not have come across an error message saying that the resource is no longer available.

Among the most common reasons why links on the web fail to work are:

- The resource may have been removed because it is no longer current or no longer perceived to be of value by its owner – ie. it "disappears"
- The resource may have been moved as a result of a file restructure, a change of domain name or a change of ownership and it has therefore acquired a different URL

Documents which "disappear"

A document may be removed from a website for many different reasons, including.

- it no longer reflects current policy
- the information in it has been superseded
- a hard copy exists for preservation
- there is an official copy in the electronic records management system
- the publication is no longer of commercial value
- "no one will want this any more"
- there is no place for it on the redesigned website

When considering removing a document from a website bear in mind that

- although you may no longer appreciate the value of your older documents, they may be valuable to others for continuing reference, research or may be of historical interest
- keeping the documents available can add to the depth of content on your site
- there could be links to the document in other documents and in catalogues and indexes
- broken links to documents on your site reduce its credibility

Documents which are moved

When considering moving documents around a site, bear in mind that

- unless measures are taken to point to the new location, all existing links to the document are broken once it is moved.
- there could be links to the document in other documents and in catalogues and indexes

2. Managing resources to ensure persistence of citation

The maintenance of links to resources on the web is essentially a task of good management. It involves planning ahead and taking deliberate steps to avoid the unnecessary removal or movement of documents on the site. The following guidelines will assist with developing a data policy for your site that should be used to manage resources for persistent access.

Determine document characteristics

Take a look at the types of document you have on your site, or on your organisation's site, and see what categories they fall into. Resource categories which need persistent access include

- information which is likely to be of long term value
- resources which you know will be linked to by other resources
- resources that are bookmarked by many people

There will also be transient or temporary information, which can be removed.

The information which needs to be managed for persistent access will generally fall into two types

- static resources
- dynamic resources

Static resources include items which have a definitive version and will not be altered, e.g. a published policy document, an annual report, a conference paper, all of which reflect the point of time at which they were created.

Dynamic resources are materials which undergo change, either incremental change by addition of content or change by modification, deletion and insertion of content. They can be either “archival” or ephemeral.

The “archival” category includes material that you wish to preserve and provide persistent access to. The decision about what level of persistence of access to provide will be a matter of judgement and will depend on the nature of the material. E.g. for an electronic journal which operates in an incremental fashion, it may be necessary to provide consistent access to both the “home page” and to individual issues and articles for citation purposes; for a publication which may be modified by change, addition and deletion, you may wish to consider archiving and providing access to successive versions so that little information is lost and the evolving nature of the publication is documented.

The ephemeral category includes “What’s new”, “What’s on”, opening hours, contact information, current publications, etc. These are not likely to be cited in other resources, but are frequently bookmarked for regular use and users want the latest version. Old versions are generally of no particular interest and this category needs persistence of citation only for the current version.

Once you have determined what categories and type of material you have, decide how you will ensure that citations to the material in the form of links remain active and enable the user following the link to access the resource.

Maintain active links

There are many ways to manage the material so that citations remain active. They fall into three broad categories:

- Use a persistent identifier system
- Organise the website to reduce the need to move material around.
- Keep older material accessible even if it is no longer on the main pages of the website.

Use a persistent identifier system

Most resources currently on the www are described using a Uniform Resource Locator (URL) and the URL as generally used describes a resource in terms of its current location, i.e. it describes the domain name of the server, directory path and the document name of the resource.

The location based nature of the URL is its major weakness for document naming, since the identifier will change if the material is moved either to a different server or a different place in the file structure of the same server. When this happens, all links to this resource embedded in other documents or databases will be broken.

A persistent identifier is a name for a resource which will remain the same regardless of where the document is located. Use of a persistent identifier will ensure that when a document is moved, or its ownership changes, the links to it will remain actionable.

It is important to note that a persistent identifier system will only be effective if it is maintained. When resources are moved, the current location **must** be associated with the persistent identifier in whatever system is chosen. This is usually achieved by a resolver database. There is no system which does not require such management.

Your choice strategies to achieve persistent identification will depend on your circumstances. Some options are:

Use redirects to guide users to the current location of the resource

As a minimum, the standard redirect capabilities of the web server can be used to redirect requests for resources using a persistent identifier to the actual location of the document or resource. This method is hard to manage on a large site and records of redirects in place need to be kept.

Use a resolver database

A resolver database can be used to map the name to a current location. Discuss this option with your webmaster or ISP, as there may be other people who have material on the site who may also wish to use such a facility. One option for achieving this is to mount a PURL server on the site. The PURL server software is made available for downloading from OCLC's PURL site http://purl.oclc.org/.

Use a persistent identifier service offered by another organisation

There are a number of other persistent identifier systems designed for use on the Internet. You may wish to investigate the use of one of these, e.g. DOI (Digital Object Identifier) http://www.doi.org/ or a Handle system http://www.handle.net. OCLC offers a PURL service to organisations that do not wish to mount a PURL Server on their own site. Users can register and apply online for a PURL and maintain their PURL(s) by updating the location of the resource online if it is moved. http://purl.oclc.

org/

Persistent identifier services and the National Library

The National Library is evaluating options for offering a persistent identifier service.

Organise the website to reduce the need to move material around

Even if a persistent identifier system is not used, a website can be organised and managed to reduce the need to move material around.

Organisation of the site

The directory structure in a site can be structured and named in a way which reduces the likelihood of change being necessary. Guidelines for consideration include:

- Avoid long directory paths and try and use logical names rather than addresses in a directory structure so that file names do not need the directory path to make them unique and the temptation to reorganise the drectory structure to accommodate new resources is reduced
- Avoid putting publicly accessible resources in private directories so that when the owner of the directory moves from the organisation or hands over responsibility for the resource, the URL will not have to change
- Avoid the use of transitory organisational names in the directory structure to reduce the temptation to change the directory structure and hence the URLs when the organisation ceases to exist or changes its name
- Consider using a date as far up the directory structure as possible for parts of the website. This can help to organise the website in a way which may assist in the management of static resources or versions of dynamic resources by indicating the date of creation and hence the standards used to create them
- If a query string is used for a known item search, don’t encode system related data in the field, as the system will be replaced in time

Re-organising the website

You do not have to re-organise the directory structure of a website to revamp the site for a “new look”. The navigation and presentation features of a website are completely separate from the underlying organising structure of the site. The site can be given a totally new look by redesigning the graphics, presentation and navigation without moving any of the underlying files on which it is built.

Keep older material accessible

Older documents which have been superseded need not be removed from a website just because they no longer reflect current policy, or because their factual content is dated, or simply because they are no longer “flavour of the month”. Older content is valuable to researchers and policy developers and

can be of interest to others. As stated earlier, there will be links to the documents from other sources which will be broken if the document is removed.

To keep older documents, you do not necessarily have to retain them on the main pages of the site. They can be:

- accessed through a logical archive
- moved to a physical archive on the site
- lodged with an archival organisation such as a library for preservation

Access through a logical archive.

When documents are no longer presented on the main pages of a site, they can be made available through a logical archive. The documents themselves are not moved, but access to them is through some form of "archive listing" on the site. Because they have not moved, the documents remain accessible via the links that have already been made to them in other resources.

Moving to a physical archive on the site

Documents may also be moved physically to a section of the site designated as an archive for management purposes and accessed via an archive listing or a search engine run over the archive. For long-term preservation, this is a safer way to proceed, since it simplifies the management process. If this method is chosen, a pointer or some other form of mapping will be needed to direct users to the new location.

Ideally, items would be identified for long-term preservation at the time of creation and would be located in such an archive when initially mounted on the server.

Lodging with an archival institution

If you cannot maintain access to the resources on your own site, then consider asking another body such as a library or an archive to take over management of the document for you. The National Library, National Archives of Australia, your local State library or archive or university library may be willing to preserve it if it is considered to be within the scope of their archiving policies. Again, if this is done, a pointer to the new location will need to be maintained.

If an older document is "archived" using any of the above methods, you may wish to display a notice with the resource stating that it no longer represents current policy, or advising that the content has been superseded and directing the user to more up to date information.

For further information on steps to take which will assist the preservation process for resources published on the web, see the companion document *"Safeguarding Australia's web resources: guidelines for creators and publishers"* http://www.nla.gov.au/guidelines/2000/webresources.html>

Contact

For more information or assistance with any matter raised in these guidelines, please contact:

Manager
Electronic Unit
National library of Australia
Canberra ACT 2600

Email: elecunit@nla.gov.au
Fax; 02 6273 4322
Phone: 02 62621140

March 2001

[Note: this document has been published as a brochure as a companion brochure to *Safeguarding Australia's web resources: guidelines for creators and publishers*]

Return to Table of Contents | Proceed to Appendix 5

About This Site | Copyright | Privacy | Accessibility | Site Map | Site Search | Content A-Z | Contact Us

APPENDIX 5
Select Reading List on Persistent Identifiers and Naming Schemes

1. General – Digital libraries and architecture

Engelbart, Douglas C. *Knowledge-Domain Interoperability and an Open Hyperdocument System*, June 1990 (AUGMENT,132082,)
http://www.bootstrap.org/augdocs/augment-132082.htm

Arms, William. *Key Concepts in the Architecture of the Digital Library* in D-Lib Magazine, July 1995
http://www.dlib.org/dlib/July95/07arms.html

Kahn, Robert and Wilensky, Robert. *A framework for distributed digital object services*. May, 1995.
http://www.CNRI.Reston.VA.US/home/cstr/arch/k-w.html

Sollins, Karen R. and Jeffrey R. Van Dyke. *Linking in a Global Information Architecture*
http://www.w3.org/Conferences/WWW4/Papers/193/

2. General naming issues

Causton, Laurie. *Identifying and describing Web resources*. (European Commission DGXIII/E-4) [2004 update: items on this site (www.elpub.org) are no longer available for download. Requests for specific docs can be sent to the webmaster at webmaster@elpub.org]

Green, Brian. *Unique Identifiers: a brief introduction* / by Brian Green and Mark Bide (undated, probably 1997)
http://www.bic.org.uk/uniquid.html

Miller, James S. *W3C and Digital Libraries*. In D-Lib Magazine, November 1996.
http://www.dlib.org/dlib/november96/11miller.html

Payette, Sandra. *Persistent Identifiers on the Digital Terrain* / by Sandra Payette in RLG Diginews Vol. 2, no. 2 (April 15, 1998) ISSN 1093-5371
http://www.rlg.org/preserv/diginews/diginews22.html#Identifiers

Powell, Andy. *Unique identifiers in a digital world*. in Ariadne, Issue 8 (9 April 1997)
http://www.ariadne.ac.uk/issue8/unique-identifiers

Requirements for an Identification Scheme for BIBLINK May 1977 in BIBLINK - LB 4034, Section D2.1 Identification
http://hosted.ukoln.ac.uk/biblink/wp2/d2.1/doc0004.htm

Van der Werf-Davelaar, Titia. *Identification, location and versioning of web-resources* URI Discussion paper. DONOR 1999
http://www.kb.nl/coop/donor/rapporten/URI.html

3. Citation linking

Atkins, Helen. *Reference Linking with DOIs: a Case Study* / Atkins et al. in D-Lib Magazine, February 2000
http://www.dlib.org/dlib/february00/02risher.html

Caplan, Priscilla and Flecker, D. *Choosing the Appropriate Copy* . Caplan & D. Flecker. Digital Library Federation Architecture Committee. September 1999
http://www.niso.org/news/reports/DLFarch.html

CrossRef - the central source for reference linking
http://www.crossref.org/

Doyle, Mark. *Citing and Linking in Electronic Scholarly Publishing: a Pragmatic Approach* / Mark Doyle. American Physical Society, [1999?]
http://www.bth.se/elpub99/ap.nsf/0/e7afb3280e8bfe47c12566ff0037b4f8/$FILE/51-59.pdf

Flecker, Dale. *Citation - E-Journal Links from the (Academic) Library Viewpoint : a presentation to the NISO/NFAIS/SSP Linking Workshop 11 February 1999*, Washington, DC
http://www.niso.org//presentations/flecker.html

Hellman, Eric S. *Scholarly Link Specification Framework S-Link-S Outline*. Openly Informatics, Inc (March 20, 2001)
http://www.openly.com/SLinkS/SLinkS.html

Hitchcock, Steve. *Towards Universal Linking for Electronic Journals* / Steve Hitchcock et al. Open Journal Project in Serials Review, Vol. 24, No. 1 (Spring 1998) pp. 21-33.

http://journals.ecs.soton.ac.uk/IFIP-SerRev98.html

Integrating and navigating eprint archives through citation-linking. NSF / JISC - eLib Collaborative Project: International Digital Libraries Research Programme
http://www.princeton.edu/~harnad/citation.html

Paskin, Norman. *E-Citations: actionable identifiers and scholarly referencing*. Version 1.2, December 17, 1999.
http://dx.doi.org/10.1000/170

The Scholarly Link Specification Framework, S-Link-S Home page
http://www.openly.com/SLinkS/

4. Naming systems – general issues

Information Object Numbering Systems (IONS) (JTAP report 42), Division of Learning Development De Monfort University
http://www.jtap.ac.uk/reports/htm/jtap-042.html#_Toc454868885

Library of Congress Network Development and MARC Standards Office. *Naming Conventions for Digital Resources* / prepared by Rebecca Guenther
http://lcweb.loc.gov/marc/naming.html

Naming and linking strategy RLG Digital Collections Project Studies in Scarlet Project
http://www.rlg.org/scarlet/. [2004 update: Information page only is still available at this address]

Paskin, Norman. Information Identifiers / by Norman Paskin (Originally published in Learned publishing, Vol. 10 No. 2, (April 1997) pp 135-156)
http://www.elsevier.com/inca/homepage/about/infoident/Menu.shtml

The Relationship between URNs, Handles, and PURLs. Library of Congress National Digital Library Program 1997.
http://lcweb2.loc.gov/ammem/award/docs/PURL-handle.html

Universal Document Naming Systems: Links
http://elib.cs.sfu.ca/USIN/naming.html

5. Specific naming schemes

5.1 Handles

Introduction to the Handle System. CNRI
http://www.handle.net/introduction.html

Handle System Namespace and Service Definition, Sam X. Sun, Sean Reilly, Larry Lannom (updated Feb. 2000)
http://www.handle.net/draft-sun-handle-system-def-02.html

Handle System Overview. Sam X. Sun, Laurence Lannom (updated Feb. 2000)
http://www.handle.net/draft-sun-handle-system-04.html

Library of Congress, NDLP (National Digital Library Program) Internal documentation. Handle Server: Overview
http://lcweb2.loc.gov/ammem/award/docs/handle-server.html

Library of Congress, NDLP (National Digital Library Program) Internal documentation. A global system of Handle Servers.
http://lcweb2.loc.gov/ammem/award/docs/h-s3.html

Library of Congress, NDLP (National Digital Library Program) Internal documentation. More on handles and naming authorities.
http://lcweb2.loc.gov/ammem/award/docs/h-s2.html

[See also the descriptions of DOI's use of the Handle system by Norman Paskin.]

5.2 DOI

About the DOI system. IDF April 2001
http://www.doi.org/about_the_doi.html

Bide, Mark. *In Search of the Unicorn: the Digital Object Identifier from a User Perspective* 1998 (British Library Research and Innovation Report; 89).
http://www.bic.org.uk/unicorn2.pdf

Davidson, Lloyd A. *Digital Object Identifiers: Promise and Problems for Scholarly Publishing* / by Lloyd A. Davidson and Kimberly Douglas in Journal of Electronic Publishing, Volume 4, Issue 2 (Dec. 1998) ISSN 1080-2711
http://www.press.umich.edu/jep/04-02/davidson.html

DOI Deployment IDF Discussion paper:
http://www.doi.org/deployment2.pdf

DOI Home page
http://www.doi.org/

Erickson, John S. *The DOI and Rights Management: Tying Up Loose Ends*. in Trialogue No. 11,

Summer 1999.
[2004 update: Trialogue not available]

The INDECS metadata framework : principles, model and data dictionary, June 2000 / Godfrey Rust, Mark Bide
http://www.indecs.org/pdf/schema.pdf

Lynch, Clifford, *Identifiers and Their Role in Networked Information Applications* in ARL: A Bimonthly newsletter of research library issues and actions. No. 194 (October 13, 1997)
http://www.arl.org/newsltr/194/identifier.html

Arms, William. *Digital Object Identifiers (DOIs) and Clifford Lynch's five questions on identifiers* / by William Y. Arms, Corporation for National Research Initiatives in ARL: A Bimonthly newsletter of research library issues and actions. No. 194 (October 13, 1997)
http://www.arl.org/newsltr/194/arms.html

Miller, Paul. *I am a name and a number : with JISC joining the International DOI Foundation, explores the use of systems such as the Digital Object Identifier*. In Ariadne, Issue 24, Jun-2000
http://www.ariadne.ac.uk /issue24/metadata/

National Information Standards Organization (US). *Syntax for the Digital Object Identifier* : an American national standard developed by the National Information Standards Organization. Bethesda : NISO, 2000. ISBN 1-880124-47-5. ANSI/NISO Z39.84-2000

Paskin, Norman. *The DOI Handbook*. Version 1.0.0 February 2001 [doi:10.1000/186]

International DOI Foundation
http://www.doi.org/handbook_2000/index.html

Paskin, Norman. *Digital Object Identifier : implementing a standard digital identifier as the key to effective rights management*. April 2000
http://www.doi.org/doi_presentations/aprilpaper.pdf

Paskin, Norman. *DOI: Current Status and Outlook* in D-Lib Magazine May 1999
http://www.dlib.org/dlib/may99/05paskin.html

Paskin, Norman. *Position paper for W3C Workshop on Digital Rights Management for the Web (22/23 January, 2000)*
http://www.doi.org/2000/12/drm-ws/pp/doi-paskin.html

Paskin, Norman. *From one to many : the next stage in DOI functionality /* Norman

Paskin (DOI) and Laurence Lannom (CNRI) August 2000

http://dx.doi.org/10.1000/190/

Powell, Andy. *Resolving DOI Based URNs Using Squid* in D-Lib Magazine, June 1998 ISSN: 1082-9873
http://www.dlib.org/dlib/dlib/june98/06powell.html

Requirements for DOI-Based Applications and Services: A DOI Discussion Paper: Revised January 30, 1998
http://www.doi.org/topics/Erickson-Services/DOIServices_jse98.htm

Rosenblatt, Bill. *The digital object identifier: solving the dilemma of Copyright protection online* / by Bill Rosenblatt in Journal of Electronic Publishing Dec. 1997, Volume 3, Issue 2, ISSN 1080-2711

Rust, Godfrey. *Metadata: the Right Approach. An Integrated Model for Descriptive and Rights Metadata in E-commerce* in D-Lib Magazine July/August 1998
http://www.dlib.org/dlib/july98/rust/07rust.html

5.3 PURL

Introduction to Persistent Uniform Resource Locators / Keith Shafer, Stuart Weibel, Erik Jul, Jon Fausey OCLC, 1996
http://purl.oclc.org/OCLC/PURL/INET96

Persistent URL Home page
http://purl.oclc.org/

PURL Frequently Asked Questions
http://purl.oclc.org/OCLC/PURL/FAQ

5.4 OAI

The Open Archives Initiative Protocol for Metadata Harvesting. 2.3. Unique identifiers - Document Version 2001-04-24
http://www.openarchives.org/OAI/openarchivesprotocol.htm#Unique Identifier

6. Publishing Industry schemes

SICI

SICI : Serial Item and Contribution Identifier Standard. ANSI/NISO Z39.56-1996 Version 2
http://sunsite.berkeley.edu/SICI/

BICI

*Book Item and Component Identifier: BICI (*Committee AP)/NISO
http://www.niso.org/bicidrft.html. Draft standard at http://www.niso.org/pdfs/BICI-DS.pdf

PII

Publisher Item Identifier as a means of document identification. American Institute of Physics
[2004 update: no longer available]

Publisher Item Identifier Adopted. American Institute of Physics
[2004 update: no longer available]

7. International Standard Numbering Schemes

ISSN

ISSN-URN
http://www.issn.org/urn/

ISSN International Centre. *Information concerning the ISSN (International Standard Serial Number).*
http://www.issn.org/

ISRN

International standard technical report number (ISRN) ISO 10444:1994
http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=18506

ISMN

International standard music number (ISMN) ISO 10957:1993
http://www.collectionscanada.ca/iso/tc46sc9/standard/10957e.htm

ISBN

International standard book numbering (ISBN) ISO 2108:1992
http://www.collectionscanada.ca/iso/tc46sc9/standard/2108e.htm

8. URI/URN

URI

Peacock, Ian. *What is...a URI?* in: Ariadne, Issue 18 (1998)
http://www.ariadne.ac.uk/issue18/what-is/

RFC 2396 Uniform Resource Identifiers (URI): Generic Syntax. T. Berners-Lee R. Fielding L. Masinter (August 1998)
http://www.ietf.org/rfc/rfc2396.txt

RFC 1630 Uniform Resource Identifiers in WWW. T. Berners-Lee. (June 1994)
http://www.ietf.org/rfc/rfc1630.txt

RFC 1738 Uniform Resource Locators (URL) T. Berners-Lee, L. Masinter, M. McCahill (December 1994)
http://www.ietf.org/rfc/rfc1738.txt

RFC 1736 Functional recommendations for Internet Resource Locators. J. Kunze (February 1995)
http://www.ietf.org/rfc/rfc1736.txt

RFC 2468 Resolution of Uniform Resource Identifiers using the Domain Name System., Ron Daniel, M. Mealling (June 1997)
http://www.ietf.org/rfc/rfc2468.txt

RFC 2717 Registration Procedures for URL Scheme Names. R. Petke, I. King. (November 1999)
http://www.ietf.org/rfc/rfc2717.txt

RFC 2718 Guidelines for new URL Schemes. L. Masinter, H. Alvestrand, D. Zigmond, R. Petke, (November 1999)
http://www.ietf.org/rfc/rfc2718.txt

URN

Uniform Resource Names : A Progress Report / The URN Implementors. in D-Lib Magazine, February 1996.
http://www.dlib.org/dlib/february96/02arms.html

Uniform Resource Names (URN) Working Group Charter:
http://www.ietf.org/html.charters/OLD/urn-charter.html

RFC 1737 Functional requirements for Uniform Resource Names. K. Sollins, L. Masinter (December 1994)
http://www.ietf.org/rfc/rfc1737.txt

RFC 2276 Architectural Principles for Uniform Resource Name Resolution. K. Sollins (January 1998)
http://www.ietf.org/rfc/rfc2276.txt

RFC 2141 URN Syntax. R. Moats (May 1997)
http://www.ietf.org/rfc/rfc2141.txt

RFC 2483 URI Resolution Services necessary for URN resolution. M. Mealling, Ron Daniel (January 1999)
http://www.ietf.org/rfc/rfc2483.txt

RFC 2611 URN Namespace definition mechanisms. L. Daigle, D. van Gulik, R. Iannella, P. Falstrom (June 1999)
http://www.ietf.org/rfc/rfc2611.txt

URN implementation in National Libraries Hakli, Esko; Hakala, Juha; Helsinki University Library 1998
http://linnea.helsinki.fi/urn/urnimp.html

Hakala, Juha. *Uniform Resource Names* In: Tietolinja News, 1/1999
http://hul.helsinki.fi/tietolinja/0199/urnart.html

URN/URI resolution

RFC 2168 Resolution of Uniform Resource Identifiers using the Domain Name System (Experimental) June 1997 / R. Daniel and M. Mealling
http://www.ietf.org/rfc/rfc2168.txt

RFC 2169 A Trivial Convention for using HTTP in URN Resolution Experimental June 1997 / R. Daniel
http://www.ietf.org/rfc/rfc2169.txt

Note: The following two documents are Internet drafts, and there is no guarantee that they will continue to exist beyond the expiry date.[2004 update: these items no longer accessible.]

Mealling, Michael. *Dynamic Delegation Discovery System (DDDS)* Internet-Draft, February 8, 2001 (Expires: August 9, 2001)
http://www.ietf.org/internet-drafts/draft-ietf-urn-ddds-04.txt

Mealling, Michael. *URI Resolution using the Dynamic Delegation Discovery System* Internet-Draft, February 8, 2001 (Expires: August 9, 2001)
http://www.ietf.org/internet-drafts/draft-ietf-urn-uri-res-ddds-03.txt

NBN

Hakala, Juha. *Using National Bibliography Numbers as Uniform Resource Names* Internet-Draft February 2000
Slightly later version still accessible at http://rfc3188.x42.com/

Hakala, Juha *URN implementation in national libraries*
[2004 update: no longer accessible]

URN Namespace ID Registration for the National Bibliography Number (NBN)

9. IETF standardisation process

RFC: 2026. Bradner, S. *The Internet Standards Process* -- Revision 3 BCP: 9. October 1996
http://www.ietf.org/rfc/rfc2026.txt

10. Organisations

Internet Corporation for Assigned Names and Numbers (ICANN)
http://www.icann.org/

Internet Engineering Task Force
http://www.ietf.org/overview.html

World Wide Web Consortium (W3C)
http://www.w3.org/

11. Mailing lists

Discuss-DOI: The DOI Mailing List
http://www.doi.org/mailman/listinfo/discuss-doi

Archives of uri@w3.org Mail Archives
http://lists.W3.org/Archives/Public/uri

Archives of xml-uri@w3.org Mail Archives
http://lists.w3.org/Archives/Public/xml-uri/

Archives of urn-ietf@lists.netsol.com
http://LISTS.NETSOL.COM/archives/urn-ietf.html

12. Persistence

Web Pages Must Live Forever. Jakob Nielsen's Alertbox for November 29, 1998:
http://www.useit.com/alertbox/981129.html

Cool URIs don't change / W3C

http://www.w3.org/Provider/Style/URI

13. Open URLs

Metadata OpenURL demonstrator. UKOLN
http://www.ukoln.ac.uk/metadata/openurl/

NISO Creates OpenURL Standard Committee NISO, March 2001
http://www.niso.org/news/releases/PR-OpenURL.html

Open Name Services OCLC
http://names.oclc.org/openurl/

Powell, Andy. *Encoding OpenURLs in Dublin Core metadata* / Andy Powell and Ann Apps. in: Ariadne, Issue 27 (23-Mar-2001)
http://www.ariadne.ac.uk/issue27/metadata/

Van de Sompel, *Open URL syntax description* / Herbert Van de Sompel; Patrick Hochstenbach; Oren Beit-Arie .
http://www.sfxit.com/openurl/openurl.html [2005 note: no longer accessible]

Return to Table of Contents

About This Site | Copyright | Privacy | Accessibility | Site Map | Site Search | Content A-Z | Contact Us