Naming Conventions for Digital Resources

Network Development and MARC Standards Office
Library of Congress

This paper contains a general summary of developments concerning naming conventions for digital resources. It was originally written as a result of a request by the National Digital Library Federation Planning Task Force for a summary of initiatives related to naming conventions both at the Library of Congress and elsewhere. The author works on the development of MARC standards and has followed discussions of the Internet Engineering Task Force's Uniform Resource Identification (URI) working groups, has participated in OCLC's Internet resources projects, and is a member of the Digital Naming Committee at LC, which is developing requirements for naming digital objects in LC's repository. Most of the information is from documents and discussions produced from these initiatives. Developments in the area of naming digital objects is rapidly changing, and the information provided below is only as current as late 1995.

Naming documents are a crucial aspect of design, because it involves the syntax of a name by which a document is referenced. A name is a logical way of referring to an object in some abstract name space. Standards are being developed in the IETF for naming conventions, called the Uniform Resource Name (URN). A URN is the name of a resource within the context of the larger Internet information architecture, or Uniform Resource Identification (URI). Other elements in this architecture include location (URL) and description/metadata (URC, or Uniform Resource Citation). A URN resolution service would be used to retrieve information about the named resource. Within this architecture, URN's are used for identification, URC's for meta-information, and URLs for locating or finding resources.

URNs improve upon URLs because they are intended to provide a globally unique, location independent identifier that can be used for identification of the resource, and to thus facilitiate access to both metadata about it and to the resource itself. Persistence is desirable, and must be provided by naming authorities. In the document RFC 1737: Functional Requirements for Uniform Resource Names by K. Sollins and L. Masinter, requirements are: global scope; global uniqueness; persistence; scalability; legacy support; extensibility; independence; resolution. URN requirements include requirements on their functional capabilities, requirements on URN schemes, and requirements on the way they are encoded.

Uniform Resource Locators (URL) have been widely used and accepted as a method for locating resources on the Internet. However, the problem of resources moving from one location to another, with locations themselves changing names or becoming obsolete, etc. has been recognized as a major problem. Everyone agrees that the concept of the URN, a persistent, unique name that can be used to provide a location for a resource, is what is needed for the future viability of information retrieval. However, that standard is still under development and clear consensus has not been reached.

The IETF URI Working Group was recently divided into subgroups, one of which is considering URN proposals. Currently there are four competing proposals which have some aspects in common. The major differences seem to lie in the resolution mechanisms used, and this is an area in which participants have failed to come to consensus. These schemes are: 1) path-URN scheme, by Daniel LaLiberte and Michael Shapiro; 2) X-DNS-2 URN scheme, by Paul Hoffman and Ron Daniel; 2) Handle scheme, by William Arms and David Ely; 4) OCLC scheme, by Keith Shafer, Eric Miller, Vincent Tkac and Stuart Weibel. Two new ones are also emerging.

Although no one scheme has been endorsed by all interested parties, there has been some consensus on components of the URN. Some differences have were ironed out in late 1995, resulting in some convergence in philosophy. One important part of the URN is the Naming Authority (NA), and this authority would have responsibility of maintaining the names in its domain. Thus, the naming authority determines naming conventions that satisfy the URN requirements mentioned above.

Reaching consensus has been slow. For URNs to work, not only is consensus on definitions and functions needed, but developments are needed for Web browser implementors and new code needs to be written by the community of network system managers who administer the Domain Name System for the Internet.

Because of the difficulty of moving the URN standard forward, institutions have had to try to find solutions for the URL changability problem. With the lack of a general resolution mechanism widely agreed upon, institutions are developing naming conventions and resolution techniques that make sense locally. Usually a unique name serves as an identifier to locate the item using locally developed software (i.e. the resolver). The following detail a few of those efforts.

Library of Congress. LC has been digitizing historical material as part of the American Memory Project for years. Early efforts determined the importance of naming conventions for retrieval of items. It is desirable to provide a link from the bibliographic record (sometimes in LC's mainframe bibliographic system, MUMS, sometimes in a special collection's local bibliographic database). In addition, because of the huge scope of this effort consisting of many thousands of files and the hierarchical file management system adopted at LC, it was decided that the pointer from the bibliographic record to the item itself could not be a URL because of the likelihood of the location frequently changing. Thus, a Digital Naming Committee explored requirements for file naming in this environment. It determined that the file name must be unique within its aggregate (a directory name which groups together digital objects). Rules for naming both aggregates and files have been developed. Together the file name (encoded in MARC field 856 subfield $f) and its aggregate name (in MARC field 856 subfield $d) provide a unique identifier that then uses a locator table to determine where the item is stored and how it is to be retrieved. The bibliographic record functions as a URC and resolves something like a URN (the aggregate and file name) into a location or URL (from the locator table). LC expected that the local software providing the locator table would not scale when the number of digital objects was multiplied thousands of times, and has two efforts working on resolving names to locations. (CNRI is providing a handle server; IBM is taking a different approach for LC's Federal Theater Project.) See attachment for excerpts from the file naming conventions document.

OCLC Online Computer Library Center. OCLC has been experimenting with Internet resource discovery and retrieval first through its Internet Resources Project (which, with other participants, developed

field 856 and the Internet cataloging guidelines) and in its Internet Resources Cataloging Experiment, and through a follow-up Internet Cataloging project. Libraries have contributed records for Internet resources containing a field 856, often with a URL. Because location-specific (URL) information is subject to failure and such information is subject to change, catalog maintenance liability is created for libraries. When records for Internet resources are distributed and included in local catalogs, this maintenance burden multiplies. In an effort to diminish the maintenance of URL information, OCLC is developing a new form of locator, Persistent Uniform Uniform Resource Locators (PURLs).

With PURLs OCLC will use a naming system and resolution service that assures reliable access to the resources. Instead of using a URL in a bibliographic record to link to a resource, a PURL instead is used. It will have the accepted syntax of a URL, but it will point to an external administrative component that then resolves the location and allows for the updating of URLs. OCLC will assign PURLs to records cataloged in the Internet Resources Cataloging Project and NetFirst as an experiment. PURLs will exploit widely accepted URL protocols in combination with unique identifiers such as OCLC control numbers. This mechanism is intended to be an intermediate step towards the time when URNs are an integral part of the Internet information architecture.

OCLC is also an active participant and supporter of the IETF working groups on URIs. As mentioned above, one of the proposed URN schemes was developed at OCLC.

Corporation for National Research Initiatives (CNRI). CNRI has developed the handle system as mentioned above. This is another of the proposed URN schemes, and is being used at the Library of Congress for some of its digitization projects.

Coalition for Networked Information (CNI). The Coalition for Networked Information has sponsored various projects in the area of Uniform Resource Identification and is also an active participant.

Research Libraries Group (RLG). RLG's Digital Image Access Project (DIAP) used a local MARC field to provide descriptive information about items and to provide a mechanism for linking the written description with a representation of the graphic image in digital form.

RLG has developed a naming and linking plan for its Studies in Scarlett project very similar to the one developed at LC. The plan includes the use of a document locator file for actual location information and collection and item identifiers to link from the cataloging (bibliographic records, finding aids, or other descriptive text) to the digitized item. Different collections may require different strategies for assigning file and directory names. As with LC's naming scheme, the collection identifier, or highest level directory name, combined with the item stem provide for a unique identification.

In the aforementioned projects (LC, OCLC and RLG) the names assigned to digital objects are determined according to local considerations. For the most part these names consist of numbers and alone do not carry a lot of meaning. In the case of LC, the file name is often a one-up number with additional elements for file use and sequencing. The aggregate name carries more meaning, representing a grouping together of items. OCLC's proposed mechanism is the control number of the bibliographic record for OCLC objects with a link to an intermediary service, and RLG's scheme is similar to LC's.

It is important that the library community keep abreast of these developments and become an active participant in the developing standards. The concerns and problems of libraries have not always been considered in the deliberations of these Internet groups. Libraries will need to come up with their own solutions in the meantime until URNs are well established.

Librarians have wrestled with many of the difficult issues now being considered in this arena and have experience from which these groups could benefit. One argument that consumed a lot of time in the URI working group concerns when a new URN is assigned, i.e. when are two digital items the same. The library community has established rules for determining when an item is a new edition or a copy, and this determines the assignment of a new ISBN. These rules may or may not hold up in the digital world, but the experience of scrutinizing the implications could benefit the current discussion. The only resolution in the IETF on this issue was that it was up to the "publisher" or naming authority. Hopefully the emerging URN standard will give some guidance, and it would be useful if librarians could contribute.

The problem of metadata, or URCs, is also under discussion. OCLC sponsored a workshop to discuss this issue, which was held in March 1995. It included a number of librarians, as well as computer scientists, image specialists, publishers, etc. Since the URC plays a role in resolving a URN into a URL, this development is important in the Internet architecture. Certainly librarians have been creating URCs for a long time, since that is essentially what bibliographic records are (albeit more detailed than what is currently being considered for general URCs). Library catalogs could serve as URN resolvers, and some of the IETF players are interested in exploring that, perhaps using Z39.50.

Excerpts from LC naming conventions documents

OVERVIEW

This document presents the rules for identifying data representing materials from LC's collections that are part of the LC digital archive. The emphasis is on unique identification rather than meaningful or expressive names. A fundamental assumption for those building this archive is that information concerning the archive's content will be made available through descriptions, finding aids, or other similar techniques. These descriptions will include directive devices [logical pointers?] to the digital texts, images, and sound materials they describe. Software will be invoked to find and display the digital materials when requested using the directive device. The operation of this software will depend on the organizational structure of the data stored in the archive. The organizing schema for the data and any file naming conventions that define meaning into character positions within the name will be made explicit and made available to those tasked with responding to requests for digital data from the LC archive.

(Source: Digital Naming 7 November 1995)

ASSUMPTIONS

1. The end user/searcher will use cataloging records, finding aids, et.al. to identify content that is of interest. The logical name that is associated with the digital materials will be recorded in MARC tag 856, subfields d, f, and g, when the Library's cataloging records are used to identify needs. When using the Library's finding aids that have been formatted with SGML, the logical name will be recorded as TAG 856d, TAG 856f, and TAG 856g. (Note: the SGML approach is still under development and may change.)

2. The information provided to the Document Locator function will be an AGGREGATE ID and an ITEM ID or a RANGE of ITEM IDs. This is NOT a complete path name. Each piece of information has a maximum length of 8 characters which will be compatible with DOS and UNIX. Alpha characters will all be lower case.

3. Each AGGREGATE ID is unique within the digital archive at LC. Each ITEM ID is unique within its AGGREGATE. An ITEM may comprise many FILEs.

4. Because each digital file will not have its own header, some necessary information has been incorporated into the FILE ID.

5. Permissions and restrictions on use may be granted at the AGGREGATE, ITEM, or FILE level.

6. An hierarchical storage management system will permit the LC digital archive to store materials on expensive/immediately accessible media, on moderate/promptly accessible media, or on inexpensive/less immediately accessible media.

(Source: Requirements for Document Locator Function at LC Prepared 26 October 1995 JLO/ITS/DA)

Prepared by: Rebecca Guenther (rgue@loc.gov)

Go to: MARC Home Page | Library of Congress Home Page

Library of Congress
Library of Congress Help Desk (10/13/2006)