On the Use of URL Objects in Directory-Level Metadata

K. A. Shein and R.T. Northcutt
Global Change Master Directory, Greenbelt, MD 20770 USA
Shein: (301) 441-4214 shein@gcmd.nasa.gov
Northcutt: (301) 441-4190 tnorth@gcmd.nasa.gov

Abstract:

With the proliferation of online resources and the increase of interoperability efforts between data providers and data locators, there is a need for Uniform Resource Locator (URL) objects in metadata records. Using the Directory Interchange Format for metadata as an example, we focus on three areas where a URL metadata object or group would enhance a user's ability to obtain and utilize the described data, as well as assisting their discovery of additional, relevant data and information. As more data become directly available online, hypertext links to their locations via a Data_Set_URL object provides an efficient means for direct data access. In cases where data are archived or grouped with other, related data sets, the inclusion of a hypertext URL to the data center or data server is important for the linkage of the user to the data. Finally, a Related_URL group that is able to direct a user to information and resources would assist them in making better use of the data or finding additional relevant information. A Related_URL group would link the directory-level metadata record to project home pages, software sites, metadata extensions and calibration/validation information.

If URL objects are to be included in directory-level metadata, it is crucial they be well defined. Several guidelines for the optimal use of the URL objects and methods in which they might be implemented are presented. Beyond the use and purpose of URL objects lies the way in which they will be constructed and maintained. We explore the use of internal URL registration versus external URN databases, PURL servers, or HANDLE systems for optimal management of hyperlinks. These various frameworks are evaluated on the basis of maximization of functionality with a minimization of maintenance.

1. Introduction

In under a decade the Internet has grown from what was initially a high-speed link between the United States government and research institutions to a worldwide network encompassing nearly every sector of society. The Internet has gone far beyond its intent as a data transfer tool to become an information access tool. In such a capacity, it must be recognized that a lack of access to information means a limitation to the usefulness of an internet site. This is especially true for directory-level metadata providers. No longer is it sufficient to describe a data set and provide data access instructions. Rather, directory-level metadata must, wherever possible, provide a direct and interactive data access method. Provided the data or data provider is on-line, direct access

of data can be most readily accomplished via a hypertext link.

Directory-level metadata serves to provide descriptive information about a data set so that a user may determine whether the data set is of use without needing to examine the data directly. In many cases a user may be unfamiliar with a particular data set and would benefit from additional information too detailed to be included in the metadata. In such instances, links to other on-line locations are provided in a Related_URL group. Related_URL serves to link to sites with information relevant to the data being described. Sites containing useful supplemental information to the metadata include project home pages, other metadata records (e.g., calibration/validation data, sensor technical specifications or archival information) and general knowledge sites. Providing hyperlinks to these related sites increases the ability for a user to receive information which may assist them in more effectively utilizing the data.

One hurdle, however that must be addressed is the implementation and maintenance of dynamic markup (i.e., hypertext links) in the metadata database. As computer systems move, resources change names, datasets are archived and databases are taken off line, the URLs in the metadata database become "broken" leading to user frustration and a sense of unreliability in the database. Several schemes have been developed in order to facilitate more reliable and robust maintenance of hypertext links based on the current technology available. These include the development of a Persistent Uniform Resource Locator (PURL) system by the Online Computer Library Center, Inc. (OCLC)(Shafer et al., 1996) and the development of the Handle System by The Corporation of National Research Initiatives (CNRI) (Sun, 1997). Additionally, the development of a Uniform Resource Names (URN) syntax is in progress (Moates, 1997).

2. URLs in directory-level metadata

A library card catalog is the most basic of searchable metadata directories. Each card in the card catalog is a directory-level metadata record of a holding of the library. By searching a card catalog, a library patron can quickly determine which books they require and where those books can be found in the library's cataloging system. Although in principle an on-line metadata directory is identical, it differs in one important aspect, it has the ability to transport the data or related information directly to the user.

A simple search of the internet for "climate data" using Digital's Altavista yielded nearly 13,000 sites. An on-line metadata directory can be used to narrow this search to data of interest to the user. However, if a directory including these data is to be effective, each metadata record must contain the location of the described data. In the case of data available on-line, the "location" is the address, or Uniform Resource Locator (URL) of the data in cyberspace.

So why is the inclusion of URLs important to the metadata provider? The most important question a metadata directory provider must answer when determining what information to present

to a user is whether that information will simplify the data discovery process. Including direct access to on-line data is a convenient method of reducing the time required for a user to obtain the data they seek. Even more efficient is to present URL information in a dedicated Data_Set_URL field. With a field for the URL of a data set a user can, without having to search the entire metadata record, easily determine if they are able retrieve the data from the internet.

Secondly, the dual purpose of directory-level metadata is to provide both locational and descriptive information about the data. Since users of an on-line metadata directory are likely to be of varying academic backgrounds, it may be important to provide some background regarding the subject about which the data were collected. Also, all users may not be aware of the technical aspects associated with the data such as calibration or validation. To include all such important information in the directory-level metadata would make the record too large to navigate effectively with the user being swamped by information they may not want or need. In order to provide this type of related information without dramatically increasing the size of the metadata record, including fields containing URLs to on-line sites of related and potentially useful information would be beneficial. With a Related_URL object in the metadata, users could selectively visit these sites based on the levels of information they require to make more effective use of the data while those not requiring additional information would not have to navigate too much extraneous information. Additionally, providing the locations of sites that may assist a user in understanding the data may cut down on user inquiries to the data or metadata provider.

3. The need for hyperlinks

Including the URL of data in the metadata record fulfills the purpose of telling the user where they can find the data. Including URLs to related and potentially useful sites adds to the ability of metadata to provide descriptive information about data. However, if URLs are to be included in metadata, the provider must determine whether to use static or dynamic URLs. If a URL provided in a metadata record is static (i.e., not hyperlinked), the user must select the URL and copy it into the OPEN window of their browser for access. Although not very time consuming, this process is a mild annoyance and may lead a user to perceive the system as being behind-the-times.

The alternative to the aforementioned static URL is a dynamic URL. Dynamic URLs are URLs which, using Hypertext Markup Language (HTML), can be identified as active links that will, upon activation, direct a user's browser to the new location. HTML markup of a URL to a hyperlink is done by placing a hyper-reference anchor (A HREF) around the URL (Fig.1).

```
<A HREF=http://www.datacenter.gov/data.html>
http://www.datacenter.gov/data.html</A>
```

Fig.1 *A hypertext link to a data location.*

Because the syntax within the "< >" is hidden, the user sees only the URL, but by changing its color and in some cases underlining, the URL is identified as a hyperlink. A user wishing to access a hyperlinked URL merely moves the mouse cursor over the link and presses the mouse button. Additional functionality is available to users having a two- or three-button mouse, where clicking the non-dominant mouse button (e.g., right button on a right-handed mouse) allows a user options such as opening the URL in a new browser window or saving the data without visually visiting the page. If static URLs already exist in the metadata, they may be made "active" by running the metadata through a search and replace program. The Global Change Master Directory, for example, stores the URL information as static text, but when retrieved, the metadata passes through a filter that "tags" the URLs with reference anchors. Although static URLs provide a great deal of added functionality to metadata, having dynamic URLs linked directly to on-line data and information is one simple method of decreasing a user's data discovery time.

4. Use of URLs in the DIF

The previous section discussed the rationale for including URLs in a directory-level metadata. Equally important to adding URLs to metadata is the determination of what objects should be added to the record and what their functionality should be. This section discusses the implementation of URL objects within the Directory Interchange Format (DIF) of the Global Change Master Directory (GCMD).

Clearly, the most important URL object to include in metadata is one that links a user directly to the data, or, if the data are not directly accessible, then to the data provider. The DIF currently contains a group called Data_Center (Fig. 2)

```
Group: Data_Center
   Data_Center_Name:
   Data_Center_URL:
   Data_Set_ID:
   Group: Data_Center_Contact
      First_Name:
      Middle_Name:
      Last_Name:
      Phone:
      Phone: FAX
      Email: Network > Address
        Group: Address
           Multiple text lines allowed.
        End_Group
   End_Group
End_Group
```

Fig. 2 *The Data_Center group in the DIF*

Within the Data_Center group exists the Data_Center_URL object. Here the URL of the data center, if it exists, resides. The presence of the Data_Center group and URL object allow users to immediately determine the identity of the data center and whether it may be accessed via the internet. Without such a group that information would not be standardized and might be scattered throughout the DIF in formats varying from one record to the next.

For an on-line data set, the DIF contains the Data_Set_Citation group (Fig. 3). This group allows the inclusion of the URL of the on-line data as well as other objects for the user to properly credit the data set creator (Vogel, 1997). However, the URL object in the

Data_Set_Citation group does not necessarily point to data which may be archived on-line at a data center. For this reason, the GCMD has sought to add a Data_Set_URL to the Data_Center group as well. In addition to providing the user instant access to data, this object is placed in a location within the DIF structure that may be more logical to users viewing the metadata.

```
Group: Data_Set_Citation
        Originator(s):
        Title:
        Publication:
        Issue_Identification:
        Publication_Date:
        Publication_Place:
        Publisher:
        Edition:
        Data_Presentation_Form:
        URL:
End_Group
```

Fig. 3 *The Data_Set_Citation group in the DIF.*

In addition to providing URLs that give a user direct access to the data or data provider, other on-line information may also be of use in the data discovery process. In the past few years, for example, there has been a great increase in the number of articles published on-line. In instances where an on-line article either uses or discusses data described in a metadata record, it may be beneficial for the user to be able to access these articles prior to working with the data. In the DIF, on-line articles that deal with the described data may be included in the Reference group (Vogel, 1997). This group is however, a free-text group with no distinct fields. A URL may be entered into this group and the GCMD software will transform it into a hyperlink upon retrieval, however, maintenance of URLs outside defined metadata objects presents some difficulties as will be discussed in the next section.

Other on-line information also can be of use and importance to users of a metadata directory. In an internet environment, a metadata provider such as the GCMD experiences queries from users with diverse backgrounds, from school children to emeritus researchers. The amount of information they require from the DIF greatly varies. To include general scientific information about the subject on which the data was collected would mean including a great deal of information that may be useful to a lay person but extraneous and annoying to a senior scientist. Conversely, providing only a brief summary of the data may not provide enough information for someone less familiar with the data to determine if they have done a successful search. In addition, even experienced researchers would benefit from knowing about project home pages and calibration/validation information.

Thus the GCMD has proposed adding a Related_URL group to the DIF (Fig. 4).

```
Group: Related_URL
        URL_Type:
        URL:
        Group: Description
        End_Group
End_Group
```

Fig. 4 *The proposed Related_URL group*

The Related_URL group is designed to provide a group where links to sites containing related or relevant data or information may be placed. By including such a group in the DIF, it eliminates the need to include a great

deal of information that is of use only to select users. Related sites include project home pages, search engines, data servers, software links, calibration/ validation data or general scientific sites related to the described data. While these URLs could be placed in the References or Summary groups, locating them there would detract from the purpose of those groups as well as increase the difficulty of maintenance on URLs in the DIF. In addition, with the presence of a Related_URL group, a user would not have to search the entire DIF for this information nor would they be distracted by finding them in other groups.

By including several strategic URLs in directory-level metadata, a provider such as the GCMD can dramatically increase the effectiveness of the database while at the same time keep the size of metadata records relatively small. One URL link in a metadata can grant a user direct access to data. Another URL can lead a user to volumes of potentially useful and even critically important information which can in turn result in their more effective use of the data.

5. Maintenance

Currently, the maintenance and verification of the URL links in the GCMD is a resource intensive and exhaustive process. The steps include:

1. extraction of the 5000+ database entries to flat ASCII text files,
2. automatic conversion of flat text files to HTML documents,
3. link verification via URL checking software (MOMspider).
4. Manual URL correction upon finding a broken link:
   a) discover new URL resource location manually (if possible),
   b) manually update database entries individually,
   c) reload the individual database entries, or
   d) update the database manually using SQL.

While automated to the greatest extent possible, the process of verifying the integrity of URL links in the database is still primarily a manually performed exercise. The most difficult of the manually performed steps is tracking down the new location of the resource in question. This requires that the operator:

1. determine whether the network connection is intact,
2. check that the resource truly is missing,
3. attempt to identify the new location of the resource (typically using only limited
4. http access on a network and file system about which they do not have access or knowledge),
5. contact the metadata author for information about the new location of the resource,
6. contact the system or web administrator for information about the new location of the resource.

If a valid new location is determined, the operator must then update each occurrence of the URL in the database. If the URL is contained in a database field, the database operator can update the entire database with one SQL statement. If the URL occurs

randomly through out the free-text block fields in the database, it may necessitate the extraction, manual updating, and reloading of each of the entries which contain the URL.

The problem of link maintenance and verification is common among many organizations responsible for a variety of online resources. The Internet Engineering Task Force (IETF) has been addressing this issue for several years and several proposals have been put forth for the adoption of a Uniform Resource Names (URN) scheme. The work done in the area of URNs has been directed toward the development of a mechanism for resolving information (resource, location) about a data object given a unique, persistent name for that object. A wide variety of online services can be utilized in the resolution process, including DNS, z39.50, THTTP, HDL, etc. (Daniel, 1997). Not only will these resources provide information about the location of the resource, but additional information, such as a description could be provided by the services as well.

As the IETF has worked on the details of the URN specification, several systems have been developed which incorporate many of the same concepts using technology and standards currently available. The Persistent Uniform Resource Locator (PURL) service, developed at the Online Computer Library Center (OCLC), incorporates many of the important features of URNs and should be fully compatible with the URN architecture once it is established. The main purpose for the PURL service is to establish and maintain a persistent name for a given resource (as opposed to incorporating the location of the object into the name, i.e. a URL). The service utilizes a database to store information about the actual location of the PURL. Figure 5 depicts the path from client (browser) to PURL server to the actual online resource:

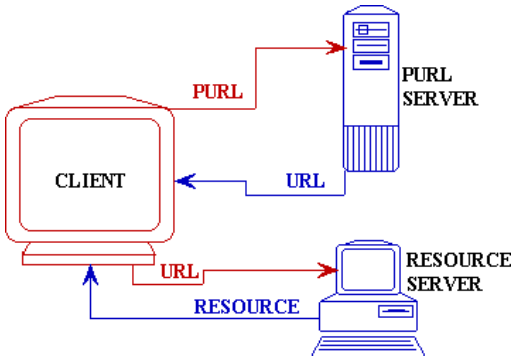

Fig. 5 *Client requests PURL resolution from PURL server which returns valid URL. Client can then use URL to retrieve the resource.*

The path from client request to PURL server to actual resource appears seamless to the user. The PURL server returns to the browser information about the actual location of the resource (i.e., a redirect), which is automatically requested and downloads to the client. To do this, the PURL server maintains a simple database which relates the object name to the object location. The name and location are supplied by the owner of the resource when they register the resource with the PURL server. The location can be changed in the PURL database at any time but name the remains constant. The benefit is that a resource registered with a PURL service can be referred to by its PURL, which, if properly maintained, will always be valid no matter how often the actual location of the resource has changed (Shafer et al., 1996). The OCLC software provides a simple user interface for easy registration and updating of PURL references. The GCMD has

installed the software on several development machines for testing and plans to offer this resource to metadata authors, data archivers and web page maintainers who interact with the GCMD system. If successful, the GCMD will install and offer the PURL server from the main GCMD database/ web server (http://gcmd.nasa.gov). PURLs are specified by a "Protocol://ResolverAddress/Name" syntax. By offering the PURL server on the same system as the actual metadata database: a) the PURL server should always be accessible when the GCMD database is available, and b) the resolver address will be stable for a very long time.

The Handle System, developed at CNRI, provides a service similar to the PURL server with a primary function of providing users with a consistent and persistent way to name internet resources. While the Handle syntax (Protocol:NamingAuthority/ItemID) is relatively similar to the PURL syntax, the Handle System is made up of several different software packages each of which offers a different resolution architecture depending on the users needs. The Proxy Server resolves URLs from handles over the internet; the Caching Handle Server caches handle resolution requests; the CNRI Handle System Resolver is a web browser extension for direct resolution using Internet Explorer or Netscape software; and there also is a client library and API for software development purposes (Sun, 1997). The CNRI is also involved with the Digital Object Identifiers (DOI) project (http://www.doi.org/). The GCMD plans to monitor the ongoing efforts of these projects for possible future use by the GCMD user population. For now, metadata authors are able to specify handles in their metadata documents with the understanding that a resolution step will need to be performed by either a "handle capable" browser or a proxy service able to resolve the handle format. The GCMD plans to compile into the PURL server code the necessary extensions to for resolution of handles as well as PURLs.

Rather than relying on an external database to handle the maintenance of URL references in the GCMD database, the new Related_URL group will provide users with a mechanism to directly update, modify, describe or delete any URLs recorded in that group. Figure 6 shows a simple form that can be filled in by the user to modify any related URL values in the database:

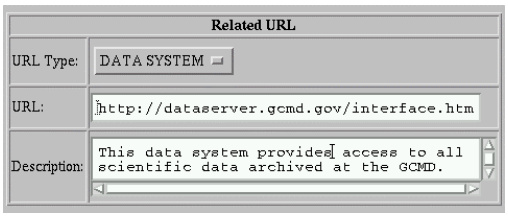

Fig. 6 *The Related_URL group can be updated using a simple online form.*

By submitting information through this form, the metadata fields described in figure 4 will be updated automatically in the metadata database. PURLs are URLs and can be included in the Related_URL group and specified with a URL_Type of PURL. Although not technically URLs, URNs and Handles can be included in this group, but there may not be a need to do so since much of the information included in the Related_URL group can also be

conveyed using the URN syntax and repository information.

6. Conclusions

The use of URLs in directory-level metadata is a simple way to increase functionality to the user without sacrificing the compact nature of the metadata record. Inclusion of URL objects such as Data_Set_URL and Related_URL allow the user instant access to the described data, as well as potentially useful and important information and tools related to the data. Overall, the selected use of URL objects enhances a user's ability to discover data and make effective use of it. Dynamic linking of URLs further increases the efficiency of directory-level metadata in that it allows even faster access to information than a static textual URL. In addition to the increase in efficiency, a database presenting dynamic content to a user will be viewed as more current than one with only static content. This perception may impact the confidence a user has in the information they receive from the database.

In terms of maintenance required for including URLs in metadata, several options have been presented. A Related_URL group has been proposed for addition to the Directory Interchange Format (DIF) for maintenance and descriptive purposes. While this will provide users with a direct mechanism for modification, creation and removal of URL references and descriptions, there are still many locations in the DIF where URLs are not easily maintained (i.e., in the summary and reference sections) and alternative methods must be used. For this reason, the GCMD will use a PURL server to allow users to name and reference any online resources which they describe in the GCMD database. As the URI, URN, URL architecture is further developed and refined, investigation will continue in order to determine if and how these changes can benefit the GCMD user population. This will be particularly important considering the GCMD's role as a coordinating node in the Committee on Earth Observation Satellites International Directory Network (CEOS IDN). The benefit of using URN/URC architectures in distributed metadata networks such as the CEOSIDN for high level data collectioin and information transfer have been demonstrated (van Gulik, 1997) and further research in this area is merited.

In conclusion, URLs should be considered a useful and important aspect of any directory-level metadata. They provide instant access to data, improve on the amount of information a user can discover about the data, do not result in a large increase in record size and, when placed in defined fields and referenced to a URN database are relatively simple to maintain.

Acknowledgments

The authors of this paper would like to thank Lola Olsen, GCMD Project Manager, at NASA Goddard Space Flight Center, for her support and guidance in the area of directory level metadata access.

References:

Daniel, R.: *A Trivial Convention for using HTTP in URN Resolution. IETF Network Working Group Request for Comments Number 2169*. June 1997 Internet Publication: http://ds.internic.net/rfc/rfc2141.txt (note: this document is a work in progress).

Moates, R.: *URN Syntax. IETF Network Working Group Request for Comments Number 2141*. May 1997. Internet Publication: http://ds.internic.net/rfc/rfc2141.txt (note: this document is a work in progress).

NASA Global Change Master Directory: *Directory Interchange Format (DIF) Formal Syntax Specification v6.0a*.
Internet Publication: http://gcmd.gsfc.nasa.gov/software_docs/dif_syntax_spec_6.0a.html

Shafer, Keith, Stuart Weibel, Erik Jul, and Jon Fausey: *Introduction to Persistent Uniform Resource Locators*. OCLC Online Computer Library Center, Inc. 1996. Internet Publication: http://purl.oclc.org/OCLC/PURL/INET96.

Sun, Sam X: *Handle System: A Persistent Global Naming Service Overview and Syntax*. IETF Network Working Group Internet-Draft, November 14, 1997. Internet Publication: ftp://ietf.org/internet-drafts/draft-sun-handle-system-00.txt (note: Expires May 19, 1998).

van Gulik, Dirk-Willem: *Meaningful Resource Location in Disjunct Metadata from a multitude of Sources*. To be presented at Interop-Sweden. European Wide Service Exchange, Ispra, Italy. 1997
Interop Publication: http://elect2.jrc.it/dirkx/sweden/Overview.html

Vogel, R.: *Directory Interchange Format (DIF) Writer's Guide, Version 5.0a.* NASA Global Change Master Directory, January 21, 1997.
Internet Resource: http://gcmd.gsfc.nasa.gov/difguide/difman.html