CENDI

PERSISTENT IDENTIFICATION: A KEY COMPONENT OF AN E-GOVERNMENT INFRASTRUCTURE

CENDI Persistent Identification Task Group

Final

March 23, 2004

Members of the CENDI Persistent Identification Task Group

George Barnum (US Government Printing Office)
Ardie Bausenbach (Library of Congress)
Charles Bradsher (Defense Technical Information Center)
Robert Chadduck (National Archives and Records Administration)
Sherry Davids (National Agricultural Library)
Walter Finch (National Technical Information Service)
Evelyn Frangakis (National Agricultural Library)
Glenn Gardner (Library of Congress)
Melanie Gardner (National Agricultural Library)
John Garrett (NASA Goddard Space Flight Center)
Jeff Given (Dept. of Energy, Office of Scientific and Technical Information)
Gail Hodge (CENDI Secretariat)
Lawrence Lannom (Corporation for National Research Initiatives)
William LeFurgy (Library of Congress)
Kurt Molholm (Defense Technical Information Center), Chair
Barbara Nekoba (Defense Technical Information Center)
David Pachter (Federal Library and Information Center Committee)
Karen Spence (Dept. of Energy, Office of Scientific and Technical Information)

CENDI is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Agriculture, Commerce, Energy, Education, Defense, the Environmental Protection Agency, Health and Human Services, Interior, the National Aeronautics and Space Administration, the U.S. Government Printing Office, and the National Archives and Records Administration. CENDI's mission is to help improve the productivity of federal science- and technology-based programs through the development and management of effective scientific and technical information support systems. In fulfilling its mission, CENDI member agencies play an important role in helping to strengthen U.S. competitiveness and address science- and technology-based national priorities.

EXECUTIVE SUMMARY

As more U.S. federal government information is generated in digital form, it is increasingly important to develop a digital information infrastructure to ensure effective management and access. A key component of this infrastructure is persistent identification of digital information resources. Currently, many government resources do not have any uniform type of identification; individual agencies instead devise their own methods for naming sources such as internal reports, presentations, and other documents. One exception is for information posted on agency World Wide Web sites. These sites make use of Uniform Resource Locators (URLs) to identify specific Web pages and objects. Unfortunately, this approach directly associates the name of the digital object with a physical location. When the object is removed from its original location, the association between the name and the location of the object is "broken" and accessing the original name yields and error message. Broken links are a major barrier to expanding electronic government, since citizens require consistent, reliable, and accurate access to government information on the Web. Current methods ensuring the association between the object and the name require maintenance, and if this is not performed consistently the association remains broken. Addressing this problem requires incorporating methods for creating and maintaining persistent identification as a key component of the Federal Enterprise Architecture.

Two primary persistent identifier applications have emerged: the Persistent URL (PURL) and the Handle System®. Both systems are in use in the government and private sectors to enable Web applications to redirect users from the "persistent URL" to the current location of the object. Handles and PURLs are globally unique and can support mechanisms such as OpenURL, which associates critical descriptive information (metadata) with identification to enable context-sensitive linking. Handles are also supportive of a federated implementation, are independent of any physical location, and can resolve to multiple locations or multiple versions of an object. The Handle System has been adopted by major publishers for persistent identification of commercially traded content through its implementation with the Digital Object Identifier (DOI) system. While most of today's implementations of persistent identifiers use PURLs or Handles for document-like objects, there are a variety of other object types from events to agreements to data sets that could be managed using persistent identification schemes.

Establishing methods for persistent identification of government resources requires extensive analysis of issues such as preferred identifier approaches, core metadata, identifier maintenance, and relationships with existing information management systems. Consideration must be given to all aspects of the government information life cycle from creation to long-term management and access to ultimate disposition, including permanent preservation. There is also potential impact on key federal information management requirements and directives such as A-130 and A-110. For these reasons, a logical next step is the formation of a group under the E-government Interagency Committee on Government Information representing a variety of stakeholder groups to study the implementation issues, analyze costs and present recommendations.

The Opportunity: Ensuring Persistent Location of Digital Objects in E-Government Services

The President's Management Agenda and the E-Government Act of 2002 emphasize the development of electronic services for citizens and industry and the efficient and effective sharing of information between and among federal government entities and other government levels. As more of these initiatives use the Internet, and in particular the Web, as their information dissemination platform, the ability to consistently locate digital objects over time becomes increasingly important.

While there are no specific studies that address the percentage of broken links ("linkrot") for government information objects, it can be assumed to be significant. While broken links on federal government Web sites may occur less frequently than on the Web at large, the realignment of responsibilities within agencies, fewer resources, changes in contractors, the ephemeral nature of some projects, and the closure of government programs and units all increase the possibility of resources being moved without warning.

Persistent identification is a key component of a reliable digital information infrastructure and is essential for providing e-government services and information. Without such persistence, citizens who have bookmarked URLs or who try to access government services from an outdated reference will receive the ever-annoying 404 message rather than valuable government information and services. References to government Web sites will result in an increasing number of broken links over time; this will frustrate citizens, increase maintenance costs, and potentially result in the withholding of services or information from the public.

As Is

What is Persistent Identification?

Bits and bytes not only need to be displayed, but to be labeled or referenced in such a way that they can be reliably found over time in a dynamic information environment. The current addressing structure for the Web is based on the Uniform Resource Locator (URL). This technology uses a physical location (IP address/server/path/file name) to identify and locate digital objects. While the URL provides direct, efficient access, URL-only naming fails whenever the resources are moved or reorganized. In addition, the URL may stay the same, but the object addressed by the URL may change significantly or be replaced with completely different content. Thus, while the URL permits interoperability in assigning an initial address for an object, it offers no assurance that this address will follow the object as it moves among locations. The lack of persistence or "linkrot" leads to 404 errors (file not found), inhibiting access to digital objects and causing problems when archiving material for long-term preservation and permanent access. A recent study by researchers at the University of Nebraska found a half-life of 55 months for what was considered to be stable information appropriate for inclusion in a Web-based curriculum in biochemistry and molecular biology (Markwell & Brooks, 2003). These statistics are of particular concern when URLs are used as links, citations or references, or bookmarks.

Unlike the URL, a persistent identifier tracks a specific object regardless of its physical location or current ownership. It is similar in function to a Social Security number which is

assigned to an individual and does not change when that person's address changes. Likewise, in the digital environment, digital object identifiers must be unique, persistent, independent of specific Web domain names, resolvable using standard Web protocols, and flexible enough to allow efficient management of digital information and accommodation of technological changes.

Persistent Identifier Approaches

Among the most commonly used persistent identifier applications are Persistent URLs (PURLs) and the Handle System. Both approaches provide registration and resolution services (similar to the resolver concept used in the Domain Name System (DNS) for URLs) to map the persistent identifier to the current physical location of the digital object. The PURL approach retains the URL construct, which can be used directly by today's Web browsers, using the Web's indirection techniques for resolving the old URL to a new one.

PURL software, developed by OCLC, a third-party provider of library services, resolves Persistent URL identifiers using servers identified by their Web domain address. The PURL is structured as:

http://purl.[resolver name, e.g. oclc.org]/[specific resource identifier]

A PURL contains the URL for the PURL Resolver Service (in the example above, the resolver at OCLC is used) followed by an identifier for the resource. The PURL assigned to the document-like object points to the PURL resolver record, which contains information to redirect the PURL to the current URL of the object. Of course, the resolver table must be updated when the actual URL location changes, but the document's PURL does not change. A PURL server can resolve only the PURLs it maintains (e.g., OCLC's PURL server cannot resolve PURLs assigned by other PURL servers). The PURL Resolver software is available free from OCLC, or PURLs can be deposited on OCLC's Resolver under an agreement with OCLC.

The Handle System, developed by the Corporation for National Research Initiatives (CNRI) under contract to several U.S. government agencies, is an interoperable network of distributed resolver servers, linked through a Global Resolver currently maintained by CNRI. The Global Handle Server registers, maintains, and resolves the naming authorities of locally-maintained Handle Servers. Any local Handle Server can, therefore, resolve any Handle through the Global Resolver. Handles, as most commonly used, resolve to the current URL of a digital object. A Handle is structured as:

[unique persistent naming authority for the assigning agency]/[unique, persistent identifier for the resource]

The Global Handle Server assures that each naming authority is unique; local Handle Servers assure that each resource identifier assigned by a naming authority is unique within that naming authority. The resource identifier portion of a Handle can be an intelligent string or an unintelligent "dumb string." Many organizations use identifiers already developed for their internal systems in this portion of the Handle.

The Handle System also supports the resolution of one Handle to multiple targets, and priorities can be established for the order in which the multiple resolutions will be used. Handles can, therefore, resolve to different digital versions of the same content, to mirror sites, or to different business models (pay versus free). They can also resolve to different digital versions of differing content, such as a mix of resources required for a distance-learning course. For example, one Handle could provide the capability to access all of the digital materials for a course. In addition to URLs, Handles can resolve to email accounts or to other Handles (supporting various Web services applications). Each of these various target categories has a unique data type. Because current Web browsers cannot support the Handle resolution directly, it is necessary to use intervening software. The software can be downloaded as an add-on client or hosted on a proxy server.

Therefore, both PURLS and Handles are redirection mechanisms that can be addressed through normal looking URLs, e.g.,

http://dx.doi.org/10.123/456

http://purl.oclc.org/some-doc-name

In addition there are other persistent identification schemes under development. For example, the XRI (eXtensible Resource Identifier) is a proposed scheme for distributed directory services to enable identification of resources and the sharing of data across disparate computer systems. The ARK (Archive Resource Key), developed by the California Digital Library, emphasizes archived digital objects. An ARK requires a link from the object to a promise statement regarding the degree of persistent maintenance, a link from the object to its metadata, and a link to the object itself.

Who Uses Persistent Identifiers?

The need for persistent identification for document-like objects has been recognized by a number of organizations and there are several implementations in the public and private sectors, in the United States and in other countries. Within the U.S. Government, for example, the Government Printing Office and the Department of Energy's Office of Scientific and Technical Information use PURLs and their own installations of the PURL Resolver to manage their connections to the full text of documents.

The Defense Technical Information Center uses the Handle System to control the identification and location of digital objects it receives from throughout the DoD. DTIC is a Handle Naming Authority. At the present time, Handles are assigned to DTIC's full text, publicly-releasable technical reports. Resolvable Handles are displayed on citations in DTIC's Scientific and Technical Information Networks (Public STINET and Private STINET). In addition, a separate DTIC Central Handle Service Directory stored in an Oracle database, contains searchable key metadata for each Handle resource. A search of key metadata (i.e., Title, Corporate Author, Personal Author, Report No., DTIC AD No., Publication Year) returns a results list from which a Handle can be selected. The central directory serves two purposes: 1) To provide handle resolution for any known Handle when a Handle prefix and suffix are known; 2) To 'discover' a Handle when some information is known about a resource, but not its Handle. The Handle provides a direct link to the resource. This directory also provides the interface that will enable DTIC to manage

resources held by outside DoD agencies. DTIC is increasing the functionality of its Handle Service and will soon provide secure access to unclassified but protected digital assets, support remote access and management of Handles and affiliated data (as part of its partnership building effort), and extend Handles to fit different digital models (e.g., distance learning objects).

Additionally, DTIC is exploring new information technologies to make a variety of digital materials available to its user communities through a Defense Virtual Information Architecture (DVIA). These materials may include textual materials such as technical reports and electronic journals, plus maps, videos, photographs, sound, spatial data, architectural drawings, computer programs, instructional materials and possibly medical imagery. DTIC will store some of the materials in its own repository and also provide links to remote sites when linking is the best way to deliver the information. Seamless searching across diverse resources will be offered. The Handle System is an essential component of this application.

The Library of Congress has naming authorities under CNRI's Global Handle Resolver for its major units. Over 400,000 Handles have been assigned since 1995 to digital objects maintained by the Library. Handles are used as identifiers in the Library's American Memory Collection (a large digital library collection), as identifiers for electronic finding aids for the Library's archival collections, and as persistent links to digital content described by the Library's distributed cataloging records and electronic finding aids. There is a current project to investigate the assignment of persistent identifiers to the Library's XML schemas.

The National Agricultural Library also uses Handles, but they are registered with CNRI. NAL has created a metadata element set for describing digital publications produced by NAL, including both original documents and digitized versions of publications formerly available only in print. To facilitate the creation of metadata for NAL-produced digital objects, an online fill-in form has been developed (the NAL Metadata Template). In addition to facilitating the creation of metadata for digital objects, use of the template also assigns a unique identifier to the digital object, registers the object with CNRI, adds the digital object to a MySQL database maintained at NAL, and alerts the Technical Services Division that a digital object has been created and that a catalog record may need to be created or modified.

The largest single supplier of persistent identification is the International DOI Foundation (IDF). IDF developed the Handle-based Digital Object Identifier to support e-commerce. CrossRef is the largest user of the DOI system. CrossRef is a collaborative reference linking service that functions as a sort of digital switchboard. It holds no full text content itself, but rather links users to digital content through the DOIs, which are tagged to object-level metadata supplied by the participating publishers, including the URL for the digital content. The end result is an efficient, scalable linking system through which a researcher can click on a reference citation in a citation database, for example, and access the cited article. More than 200 publishers deposit and maintain DOI persistent identifiers in the CrossRef System to link citations and references to the full text of journal articles and books. To date over 9 million identifiers have been registered in the CrossRef system and more are being added each day.

CrossRef is a DOI Registration Agency. The primary role of Registration Agencies is to provide services to Registrants - allocating DOI prefixes, registering DOIs, and providing the

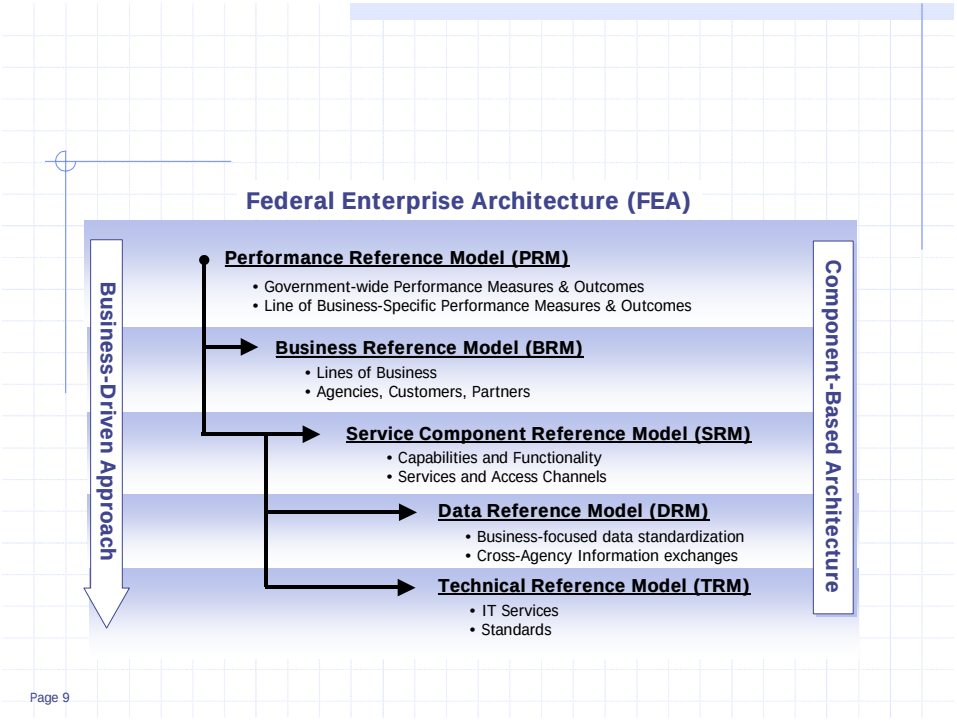
necessary infrastructure to allow Registrants to declare and maintain metadata and state data. The Registration Agency concept allows the Handle System to be a distributed system.

The TSO (The Stationery Office) in the UK is also a DOI Registration Agency. TSO publishes on behalf of UK's Parliaments and assemblies and is the UK's definitive source of official and regulatory information. The TSO recently joined the International DOI Foundation in order to help its government and private clients better manage digital resources. With a major emphasis in the UK on E-government and with support from the E-government Envoy, the TSO is seeking to provide a series of services that use the DOI to support management of government information over time.

What Could Be

Persistent Identifiers and the Federal Enterprise Architecture

How does this technology fit with the Federal Enterprise Architecture (represented by the figure below) being developed by OMB under the authority of the E-Government Act of 2002?



Persistent identification is a core infrastructure component that touches every model of the FEA above. Persistent identifiers specifically support the Data Reference Model of the FEA by enabling information exchange and use within and even across FEA business lines. The technology also provides the Technical Reference Model with both an IT service (the resolution) and a standard identifier syntax (the structure of the identifier and the core metadata).

The FEA's emphasis on standards and best practices can be addressed by development of a Persistent Identifier framework. A federal-wide design, if not implementation, of a Persistent Identification Resolver would permit data types used in the resolver to be standardized. This would result in a common look and feel for users. In addition, a central approach can produce shared best practices, additional services and plug-ins, and non-redundant submission for services and products that are produced by multiple agencies.

The government examples described above outline the use of persistent identifiers to link to the current location for document-like objects. However, persistent identifiers can also be used in "non standard" ways to integrate and link digital content. Identifiers, for example, can help track events, such as conferences (to manage conference proceedings as either a single resource or separately as independent entities) or meetings, track the flow of correspondence, link versions and updates of departmental regulations, monitor the impact of training on its employees, bring together information about a terrorist from a variety of agencies, manage agreements, identify data sets, or manage electronic records (including appraisal, retention and disposition activities).

Project Management and Maintenance Issues

The successful implementation of persistent identifiers may well face more social and economic challenges than technical ones. The implementation of a Federal Persistent Identification Resolver requires ongoing maintenance and, therefore, ongoing resources. At the individual agency level, the resolver must be kept up-to-date with the current URLs for the locations of the government digital objects. The resolution provided by the system is only as up-to-date as the physical locations to which the persistent identifiers point. While some of this updating can be automated, responsibility for this updating and ensuring its reliability must be assigned within each agency, program or office or to a trusted third-party. It is not sufficient to create identifiers and leave them without maintenance; active management is needed in order to gain the benefits of such a system.

The adoption of persistent identifiers will require the introduction of new technology, although the specifics of the identifier and the use of that identifier will determine exactly what new technology will be needed. In the case of both PURLs and Handles, an organization will run its own server, either a PURL server or a Handle server, or arrange to register its identifiers on another organization's server. In the case of Handles the server containing the Handles and the proxy server mapping the URL requests to the id server can be one in the same or they can be separate while in the case of PURLs they are always the same. On the client side a normal Web browser can be used if the identifier is in the form of a URL. However, use of the native Handle protocol, which allows added functionality, would require special plug-in software or a separate client application.

Successful implementation of a Federal Persistent Identification Resolver may impact federal information policies, including A-130, A-10, the American Technology Preeminence Act, the data clauses of the DFARS, the FAR, and subsequent agency directives and regulations. These areas need further discussion and investigation.

Specific Next Steps

- The E-government Interagency Committee on Government Information should establish a group representing various stakeholders to address persistent identification as a component of the e-government infrastructure. This should include discussions of centralized versus distributed implementations, a framework for development of a government-wide persistent identification scheme or schemes, and discussions of the gateway and user interfaces. The group must analyze the costs of such a system based on the specific implementation decisions.
- Once criteria have been established, the FEA model should be modified to specifically reference the need for each agency to consider what information should be managed using persistent identifiers and the specifics of the scheme to be used.
- Policies and guidelines should be established for the creation and maintenance of persistent identifiers and the management of related services.
- A key area for discussion and consensus is the determination of a core set(s) of persistent identification metadata (and metadata formats) necessary to assist in discovery, digital rights management, and the provision of associated services for different user communities.
- All the above discussions should consider the life cycle of information from creator to initial and secondary dissemination to preservation, long-term records management and archiving.
- Since the Web crosses all sectors and national boundaries, it will be important to work with other organizations that are involved in the discussion of persistent identifiers on a national and international scale. This includes not only organizations such as the International DOI Foundation, CNRI and OCLC which are directly involved in these technologies, but also groups such as the World Wide Web Consortium (Web protocols and other technologies), the International Federation of Library and Museum Associations (libraries and museums), the International Publishers Association (publishers of all kinds), the International Standards Organization, the National Information Standards Organization, and the Dublin Core Metadata Initiative (which has created a special interest group on Persistent Identification within the Dublin Core Metadata Initiative).

References

Handles System. [http://www.handle.net/]

International DOI Foundation. [http://www.doi.org]

Kunze, J. and R. P. C. Rogers. "The ARK Persistent Identifier Scheme." Internet Draft January 31, 2003. [http://www.cdlib.org/inside/diglib/ark/arkspec.pdf]

Markwell, J. and D. Brooke. "Broken Links: Just How Rapidly Do Science Education Hyperlinks Go Extinct?" 2003. [http://www-class.unl.edu/biochem/url/broken_links.html]

OASIS. "eXtensible Resource Identifier TC." [http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xri]

Persistent URLs. [http://www.purl.oclc.org]

TSO. "Information Management & Interoperability Strategies: The Case of Digital Identifiers." March 31, 2003.