Chris Testa
Mary Ambrosio
Library of Congress

Contact: ctes@loc.gov
Phone: 202-707-2801

Improving Search for Government Resources:
Open Standards and Government, Industry Collaboration

The Library of Congress is the nation's oldest federal cultural institution and serves as the research arm of Congress. It is also the largest library in the world, with millions of books, recordings, photographs, maps and manuscripts in its collections – over 10 million of which are available in digital form via the Library's web site.

The Library's mission is to "make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations." As such, resource discovery and search are central to the Library's mission to the public and Congress. Last year, the Library convened a group of project managers, data specialists, information architects, and programmers to develop strategies and implement product in support of these business objectives. Since inception, this "Project Discovery" group has launched several initiatives that have directly impacted the findability of and access to our digital collections and items. In this paper, we will discuss 3 of these initiatives to improve access to our content via the Library's retrieval systems as well as external tools, services, and applications.

In October of 2006, the Library launched a new federated search beta (http://www.loc.gov/search/new/) to the public. Given the diversity of our content *and* the number of distinct content areas and repositories (American Memory, the Library Catalog, the Prints and Photographs Collection, THOMAS, and the thousands of static HTML pages) available from our web sites, it was becoming increasingly difficult for site visitors to find, access, and use our materials. In addition, though it is a long term goal of the Library to centralize and standardize our data sources, in reality it is a formidable task. As a short term solution designed to improve searching and learn more about our underlying systems and data (as well as learn from visitor usage), we decided to centralize our search interface. Using open source products (Index Data's Keystone Retriever, Apache) and leveraging open standards such as Z39.50 and SRU/W, the Library was able to implement this single search interface to our key content and digital repositories. As a result, we have increased the ability for our users to access our content. In addition to the positive reactions to the product during our pre-launch usability testing, we have received a high percentage of positive user feedback from our visitors.

In addition to the user interface and discovery challenges inherent in the applications the Library has developed internally to accomplish our goals, another key challenge has been making our content and items searchable and accessible via the multitude of external services and search engines. By employing Open Standards and

working with industry on emerging web technologies, we have made great strides in making our content available from these external services.

First, the Library participates in the Open Archives Initiative (OAI) whereby we make our content available for harvesting. The Library provides parameters as to format but they all adhere to an existing protocol/standard for bibliographic data (marc, Dublin Core, marc xml). This ensures public access to our content for another institutions' use, as opposed to limiting them to searching at our site. The use of open standards allows applications to remain flexible and allows for the upgrading of a component without rewriting the whole application. If an application relies on existing protocols, it will be able to communicate with other applications. In terms of architecture, if the presentation/interface is separate from the application, it allows for various interfaces for different audiences.

Another initiative designed to make our resources externally available involves the use of XML sitemaps. Approximately one year ago, the Library began a collaborative effort with Google to implement their sitemaps in order to provide better digital content for their crawls. Prior to this, their crawls only picked up static html pages and missed the dynamically generated metadata and actual content for many of our digital items. The sitemap enumerates not only static pages for a given digital collection, but also includes links which resolve to a display of the individual items in the collection. The use of sitemaps as evolved into a sitemap protocol also used by Microsoft and Yahoo (**www.sitemaps.org**).

Through these initiatives and product releases, the Library has made a great step forward in increasing the accessibility of our resources. By leveraging open source tools and standards, and by employing best practices of information architecture and visual design, we have improved the user experience of our applications, services, and digital discovery tools for our constituencies. By working with standards bodies and industry to make our resources available through services outside the Library, we have improved discovery and enabled further outreach to the American public and Congress.

The session, if the paper is accepted, will discuss the strategies, opportunities, and issues, with attempting to provide access and discovery of legacy data systems and applications in the new Web 2.0 world.