

Metasearch Infrastructure Project: Recent Achievements & Current Status

Roy Tennant, roy.tennant@ucop.edu • May 5, 2005

The Metasearch Infrastructure Project seeks to develop a robust set of tools for crafting tailored search interfaces to diverse information resources for specific audiences and/or purposes. For additional background information, see the project web site at <http://www.cdlib.org/inside/projects/metasearch/>. Roy Tennant is the Project Manager and Mike McKenna, michael.mckenna@ucop.edu, is the Technical Lead.

CDL Developments

Following the installation of the MetaLib application and staff training in Fall 2004, CDL staff investigated methods for customizing the user interface. In so doing we discovered that some of what we wished to accomplish would be very difficult, and some would not be possible at all. Meanwhile, CSU San Marcos unveiled an in-house user interface to MetaLib created using an XML gateway or API (called X-Server) to the MetaLib application. The Metasearch Infrastructure Team decided to stop working with the native interface and instead use the emerging CDL Common Framework as the means to create a new set of tools for crafting specialized interfaces to MetaLib.

After making the decision to go with the X-Server, which has only about 30% of MetaLib functionality available, Roy Tennant marshaled ExLibris customers to come up with a list of desired X-Server enhancements that was prioritized and sent to ExLibris. Most of these enhancements will make it into the next two releases of the X-Server software. CDL now hosts the METALIB-XSERVER-L discussion for ExLibris customers interested in the X-Server, and is one of three institutions beta testing the v. 3.13 release of X-Server, due out in mid-June.

Mike McKenna has been drafting the architecture and development timeline of the Common Framework enhancements required to support metasearching, in association with Stu Sugarman and other CDL staff. Present plans are to be able to support the development of a few prototype portals for Fall 2005. At this time, interface development will not yet be simple, as some level of Java Server Pages and/or XSLT knowledge will be required, in addition to XHTML and CSS experience. Therefore, CDL staff will be more involved with portal deployment at this stage than we hope to be in the future, when interface changes will be easier to achieve.

Mike McKenna installed updates to Metalib that add a newly developed Search Server Architecture that is dedicated to handling user searches with more effective and efficient use of system resources and improves search response times. Improvements were also made in the de-duplication mechanism to achieve enhanced performance in the creation of merged and de-duplicated lists. As a result, MetaLib users should experience a significant improvement in performance.

Sherry Willhite and Mike McKenna worked on fixing problems with resource connection packages surfaced by the UCLA team (see separate interim report). Mike has also successfully experimented with using the Google software gateway (XML API) to make a

resource searchable that is not yet available for searching in the MetaLib Central KnowledgeBase. This solution may provide at least a temporary solution for searching some resources that are either not in the CKB, or are “search and link” resources that report only the number of hits from a search.

Following the principle that fewer targets to search is desirable when possible, we are developing two other tools for making digital library collections available for metasearching. Since the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is considered a necessary tool for making digital library collections available for metasearching, Roy Tennant performed a prototype harvest of several dozen repositories and more than 350,000 metadata records using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). He also constructed some prototype tools for metadata analysis, and created initial specifications for a set of metadata normalization and transformation tools. The Harvesting Core Team, under the leadership of Bill Landis, has now taken over this work and is integrating the harvesting function within the CDL Common Framework.

Targeted web crawling is another capability we are developing for the metasearch infrastructure, and toward that end we have been collaborating with the NDIIPP grant-funded project at CDL that is developing a set of web harvesting tools. Heather Christenson is taking the lead in making sure metasearching requirements are reflected in the design documents, and we have performed some test web crawling to help determine the best method of integrating web search results with licensed content. By harvesting metadata or crawling web sites, we gain more control over searching these targets, just as Google does with the sites they crawl.

To demonstrate how these various components could be used within a metasearch portal, we developed a partially functional prototype of an Earth Sciences portal at <http://dali.cdlib.org:8080/metasearch/nsdl/> that demonstrates the use of RSS feeds on the opening screen, and searching OAI-harvested metadata and crawled web sites once a search is performed.

We are also thinking about how users can get to the appropriate portal once we have a number of tailored portals available. We believe users will come to these portals in at least a few different ways: 1) direct links from library and campus web sites, 2) integrated search boxes on course web pages, etc., and 3) eventually through a single location such as a “portal finder” which could select an appropriate portal based on a specific query.

A similar problem is how can we recommend additional useful resources to search that are not included in a particular metasearch portal. We believe both problems might be solved by creating a “recommendation engine.” In one scenario, the recommendation engine would automatically redirect a user to the best portal for their particular search. In another scenario, once a user is at a particular portal, the recommendation engine may suggest other specific resources to search that are not part of the metasearch, again based on the user’s query terms.

Campus Developments

UCLA Library staff, through a project partially supported by CDL, has helped surface a number of issues regarding how resource connectors are configured by default, leading to a number of changes in those connection packages to improve or fix how they are searched.

Campus staff are also being consulted on the search portal for the grant-funded National Science Digital Library project, as well as for the undergraduate “SmartStart” search portal.

CDL developed a set of “talking points” at http://www.cdlib.org/inside/projects/metasearch/metasearch_talking_points.pdf for librarians to use.

The UCB Interactive University worked with the X-Server, uncovering bugs and enhancement needs.

Outside Developments

Mike McKenna has participated in the work of the NISO Metasearch Initiative since its inception, and Roy Tennant recently joined the group as well. They have worked on the team charged with developing a standard metasearch gateway that database publishers can use to offer a standard machine interface to their databases.

First Roy Tennant, then Bill Landis, have worked with the Digital Library Federation and NSDL OAI and Shareable Metadata Best Practices Working Group (see <<http://oai-best.comm.nsdlib.org/>> for more information) to develop best practices for OAI data and service providers.

Findings

- MetaLib and the CDL Common Framework will remain installed at CDL, with no current plans to install either on campus machines
- The native MetaLib administrative interface will continue to be used for managing resource connection packages, creating QuickSets, etc.
- Resource connection packages in the Central Knowledge Base (CKB) will be co-managed centrally (for Tier 1 resources) as well as locally (for all other resources) from the central CDL installation
- An EZproxy proxy server should be installed locally for authentication purposes; CDL staff can advise campus staff on how to accomplish this