

Metasearch Infrastructure Project: Recent Achievements & Current Status

Roy Tennant, roy.tennant@ucop.edu • August 11, 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.

For appropriate background to this update, see the previous updates available from <<http://www.cdlib.org/inside/projects/metasearch/overview.html>>.

CDL Developments

In partial completion of the National Science Digital Library grant CDL received in 2003, the report *Integrating Information Resources: Principles, Technologies, and Approaches* was written to discuss integration strategies (including, but not limited to, metasearching) and the strengths, weaknesses, and deployment decisions to be made about each. It will be available from the Metasearch Infrastructure Project web site when it is released (before the end of the month).

Metasearch Software

Following the decision to not use the MetaLib native interface, CDL began working to integrate MetaLib with the CDL Common Framework — a set of interoperable applications that are destined to provide support for many of CDL's services. Mike McKenna developed a plan to revise existing Common Framework components, or write new ones as needed, to integrate MetaLib with the Common Framework. This development work is in full swing, with an initial release of this functionality scheduled for late Fall.

Concurrent with this work has been the ongoing specification, prioritization, installation, and testing of additional MetaLib X-Server capabilities. CDL has taken the lead in working with MetaLib customers to specify a common set of prioritized enhancements, the first fruits of which have come with the 3.13 version of MetaLib installed earlier this month. Version 3.13 substantially expanded the X-Server capabilities and also vastly improved the documentation. Testing is on-going, and future bug fix releases are expected.

Present plans are to deploy the first prototype portals for testing in late Fall or early Winter. These prototype portals will be drawn from existing metasearch projects such as the search portal for the American West Project, the NSDL Project, the UCLA European Studies project, and the undergraduate "SmartStart" portal. 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.

Metadata Harvesting

Our work with developing a set of tools to support harvesting of metadata from remote repositories using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-

PMH) is progressing. Records have been harvested from a variety of repositories (including CDL) and ingested into a Common Framework repository. Additionally, utilities for normalizing dates, validating URLs, and creating image thumbnails are in various stages of progression. Bill Landis has been leading the Harvesting Core Team in the specification and development of these capabilities.

Targeted Web Crawling

Work has begun on specifying an infrastructure to perform targeted crawls of librarian-specified web sites for exposure through a metasearch portal. This is not to be confused with CDL work to crawl web sites for preservation purposes. Heather Christenson is leading this work.

Recommendation Services

We are continuing to explore a possible recommendation service that could recommend useful databases in response to a specific query.

Campus Developments

Although we receive advice and assistance from staff at nearly all the campuses, we so far have worked most closely with staff from UCLA and UCSC. As we leave the prototype/testing stage we will be looking for partners at more campuses to further refine and test our infrastructure. Also, we may request feedback on specific components from a broad group of campus librarians as needed.

In June, CDL staff collaborated with UCLA staff to perform metasearching needs assessment with graduate students and faculty in European Studies. The final report is available from the Metasearch Infrastructure Project web site.

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

Outside Developments

Mike McKenna and Roy Tennant continue to serve on the NISO Metasearch Initiative, Task Group 3, which is finalizing a standard metasearch protocol for database vendors. Roy also serves on the group planning the NISO Workshop on OpenURL and Metasearching, September 19-21, 2005.

Mike McKenna and Roy Tennant presented our X-Server work at the SFX/MetaLib Users Group meeting in June.

Bill Landis continues to work with the Digital Library Federation and NSDL OAI and Shareable Metadata Best Practices Working Group (see <<http://oai-best.comm.nsdl.org/>> for more information) to develop best practices for OAI data and service providers.