

**NISO Metasearch Strategy Meeting:
Standards to Support Metasearching
May 7-8, 2003
Four Points Sheraton Cherry Creek
Denver, Colorado**

The day and one-half meeting was attended by content or data service providers, metasearch engine developers and marketers, and several people representing end users or end user organizations. One of the concerns that became obvious to me was that the data service providers, metasearch vendors and the users, while having a huge overlap of concerns, had other areas where the priorities were vastly different and in a few areas, diametrically opposed.

A few of the conflicting areas included the merging and de-duplication of records, algorithms for determining relevance ranking and the solutions for handling server overload. Specifically, the DSPs were in favor of merging and de-duplication only if “their” records were not omitted and felt that relevance ranking should be done only if “their” records were not relegated to “page two” of the display. Both the DSP folks and the metasearch vendors were in favor of omitting the “all databases” button from the display to prevent “frivolous” searches. The DSP vendors felt that the best solution to reduce server overload was to provide only a partial search at times when the traffic is heavy.

A variety of short and long term solutions were suggested by all six of the discussion groups, including new and revised standards, “best practice” documents which could be registered by NISO, and general improvements to search and reporting parameters passed between DSPs and metasearch engines.

Overview

The general organization of the meeting included an introduction by Pat Stevens, OCLC and Oliver Pesch, EBSCO. They described the history of the project and introduced the Planning Committee. The challenge is that users want better search engines to “be like Google” while the resources with which we deal in libraries are not “Google-like”. The stated goals of the meeting were four-fold:

1. To advance the state-of-the-art of metasearching
2. To work within the environment of the World-Wide Web and the Internet
3. To assure that the standards and best practices are well implemented
4. To assure that the solutions should benefit all providers (users? sic!)

The deliverables were to be actions that might result in the establishment of new standards committees, research into a topic (white paper) or advocacy within W3C

The meeting then included presentations from five different points of view which recognize “we are all service providers”:

1. End users and their organizations: Brenda Baily, Colorado State Library and George Machovec, Colorado Alliance of Research Libraries (CARL)
2. Integrated Library System Vendors: Jenny Walker, Ex Libris
3. Metasearch Systems: Peter Noerr, MuseGlobal
4. Content Aggregators: Ed Moura, Gale Research
5. Primary Publishers: Marc Krellenstein, Elsevier.

Each attendee had been asked to identify a preferred discussion topic and assignments were made to balance each group to represent all stakeholders. The six discussion areas were:

1. Access Management
2. Statistics
3. Searching Options
4. Metasearch Definitions
5. Resource Description
6. Result Set Management

The rest of the first day consisted of small group discussions with interim reports from each group at the end of the day. The second day began with small group meetings to finalize recommendations with report to the whole group and general discussion before concluding remarks and adjournment.

Opening Panel – “We are all Service Providers”

Brenda Baily from CSL provided the view of metasearching through a public (library) portal. She believes that users should be able to search across a wide variety of targets, including community information and GIS databases. The metasearch engine should be interactive with such systems as “virtual reference”. A portal should be easy, fast, reliable, customizable, flexible and easily maintainable.

She believes that an 18 month goal should include transparent federated searching with robust statistics. The three-year goal ought to include access to “digital government”, GIS and other non-standard and non-bibliographic databases. She felt the greatest challenge would be speed and system performance.

George Machvec from CARL represented the academic library users. He felt that a portal should be easy to implement, be flexible in target selection, should search across multiple protocols, be ADA compliant, provide automatic selection of content and databases based on user profiles, provide service alerts (new resources added, etc.), should provide access to many targets (300-500), have set timeouts to protect user privacy, a search should “show the good stuff first”, provide links to courseware, and provide the ability to link to the native interface of the resource.

Jenny Walker from Ex Libris, representing both an ILS vendor and a metasearch system provider, reported on the Ex Libris products: Metalib and SFX. She reported that of the 200 or more sites using Metalib, about 60% are not using Ex Libris's Aleph500 ILS system. She said that one of the challenges has been to figure out where the functionality of the ILS-based OPAC ends and the metasearch engine (in their case, Metalib) begins. The role of the union catalog vs. the metasearch engine also provides some challenges.

Marc Kellenstein from Elsevier reported that Elsevier does not have a well articulated strategy to support metasearching. He believes that the challenges are primarily around access control and the need to limit searches (metasearches) to those from Elsevier-licensed sites. Along with the other DSPs, he is concerned with "inappropriate" searches generated by metasearch engines (i.e. searching in a nursing database for an English literature citation). Elsevier is developing an XML gateway and is moving toward a SOAP (Single Object Access Protocol) solution using XQuery language. They plan to provide a "query wrapper" to be sent to the metasearch engine for result set metadata and will transfer the result set itself in XML.

The challenge for Elsevier (in addition to the ones he had already mentioned) is how to handle "unsupported" features, updates, etc. They are also concerned with "branding" issues, whether de-duplication will be fair, who defines "relevance", etc.

Small Group Discussions

We then broke up into the six discussion groups for the remainder of the meeting. I was assigned to the "Searching Options" Group, which I viewed as the real "core" of the metasearch problem.

Searching Options Discussion

We spent time discussing the scope of resources to be searched. It was clear that the group lacked user perspective and was surprised to hear that users wanted to search non-bibliographic targets such as learning systems, museum databases, GIS systems, etc. along with bibliographic targets with the metasearch tools.

There was extensive discussion on the "ethics" of searching and the need to filter out "frivolous" searches. It was not clear how "frivolous" searches would be defined or who would define them. This is an area of huge disagreement between the various stakeholder perspectives.

There was considerable discussion about the problem of server overload (related in part to the "frivolous" searches, but not limited to those occasions). The DSP representatives on the group felt that the best solution would be to completely eliminate the "all databases" button on the search screen. In addition the DSP folks proposed that they would provide "limited results" when the traffic was heavy. Clearly, the user perspective would be for the DSP to purchase more server power. This was another case where there is potential for disagreement between the stakeholder categories.

Summary Discussion

All of the groups provided an interim report at the end of the first day and agreed to refine their recommendations for a presentation to the whole group at the end of the second morning. Some groups continued discussion after the summary session and all groups reassembled after breakfast to resume their discussions. Each group was meant to come up with recommendations for the short term (18 months) and the long term (3 years). A summary of the recommendations of each group follows:

Access Management Group

1. Develop formal “use cases models” to define the problem space.
2. Inventory existing services (Shibboleth, etc.)
3. Make certification separate from authentication (this group clearly understood the difference between authentication and authorization, but introduced a third term “certification” which seemed to get used for both concepts at times).
4. Define whether a search is coming from a metasearch provider or directly from an end user.

Statistics Group

1. Determine customer needs
 - a. Have NISO sponsor a discussion group in conjunction with a national meeting (e.g. ALA)
 - b. Conduct a web survey of end user needs (presumably they mean libraries as “end users”).
2. Review Z39.7 – Metasearch tools on Statistics and Terminology.
3. Promote the writing of articles on metasearching statistics
4. Promote research (i.e. – library school faculty and graduate students)
5. Examine standards initiatives for NISO involvement.

Searching Options Group

This group defined several best practice scenarios:

1. We need to define preferred search language protocols. XML is preferred, followed by Z39.50, with HTTP (Native interface) the least preferable. There was considerable

discussion as to whether new vendors were likely to develop Z39.50 interfaces. Most felt they would not. A great deal of dismay was expressed that NISO has decommissioned CCL as a standard. XQuery may be a reasonable replacement, but should be examined for comprehensiveness.

2. Systems should have the ability for the users to switch to the native interface to the database if they choose.
3. There should be methods put in place to limit “frivolous” searching. While the group discussed a number of scenarios, no single method was cited in the final recommendations.
4. There should be a better way to pass session information back to the metasearch engine in a way that it could be included in the next query where appropriate.
5. We need to make better use of and provision of resource descriptions.

Potential standards work might include

1. A shared vocabulary for users, metasearch providers and information providers.
2. Standards for session management and authentication
3. Work with SRW/SRU or other web service models.

Long-term goals –

1. Define a request attribute set
2. Move toward less stateful models
3. Provide and manage server load statistics
4. Define standards for relevance ranking
5. Develop a standard XML gateway option for all resources
6. Develop templates to help all three stakeholders define information exchange.
 - a. Short term – eye-readable, long term – machine-readable

Metasearch Identification

This group seemed to identify and conclude their work with little discussion. They recommended the development of guidelines for registration of all metasearch engines (providers) with NISO as a short-term solution.

Two methodologies were suggested

1. Providing a special port on the DSP server for all metasearch requests
2. The addition of a protocol element to the search to identify a metasearch request.

Resource Description Group

This group redefined their group as the Collection Description Group.

1. They defined their goal as helping the users find the best resource. This would require both a good “collection description” and a “service access” description.
2. The metadata should define the schemas for exchange and the services for exchange. The task is to define and then develop semantic descriptions in both of these categories.
3. They defined other groups that are working in this area: JISC, ISO, DC (collection description group), ISO-ILL, DLF, EAD, ZING, etc. The group recommended the development of a “Joint Working Party” and hosting a joint meeting with JISC to build on the work that JISC is doing.

Result Set Group

This group looked at both single record metadata and result set metadata. They expressed the need for a “next generation search protocol”.

Result set metadata recommendations

1. Short term – metadata needs to be added to the HTML describing the result set.
2. Longer term – We need to define a data element set that could be used as part of the search protocol.

Single record metadata

1. Short term – We need to define the “best practice” for a basic level of result metadata when a single record is retrieved.
2. Long term – We need to define the element set for administrative control metadata (relevance ranking, deduplication, etc.)

Strategy – Review other schemas: METS, MOPS, etc. Possibly develop new schema.

Open Issues – Need to allow for extensions to core metadata. We need to find the best balance between being “simple to implement” and having rich metadata.

Next Steps

The summaries and recommendations of all six groups will be sent to the Planning Committee to determine the next steps.