Integrating Information Resources: Principles, Technologies, and Approaches

In partial fulfillment of National Science Foundation Award #0333710
August 15, 2005

Heather Christenson and Roy Tennant

http://www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_report2.pdf

Introduction

The California Digital Library (CDL) was awarded a two-year National Science Foundation grant in 2003 to build on and enhance the National Science Digital Library (NSDL)[1]. As part of this research project, CDL is building a prototype service that will demonstrate how NSDL collections can be integrated with other library science and technology collections (e.g., licensed commercial databases).

A stated deliverable of this project is a "formal review of products, technologies, and approaches available to the CDL to build the integration service as specified." This review is to "provide a broad analytical overview that will be valuable to anyone interested in the current state-of-the-art in the market for federation, integration, and customization of networked information resources."

As a first step toward building the integration service prototype, CDL conducted a review of products, technologies and approaches for building an integration service. Findings from those activities that are of general interest are included here. To enhance the lifespan and applicability of this report to the experiences of others, information regarding specific products is not included. Given the rapid development of products within this market space, readers of this report are encouraged to perform their own review of potential software solutions at the time of need, using many of the criteria outlined herein and in other reports referenced here.

User Needs

Any project that attempts to serve user needs must begin there. What do the prospective users of your system wish to do? How do they think about it? What words would they use to describe it?

Needs assessment activities should be undertaken prior to the development of a service. With needs assessment, your goal is to determine what your target user populations (many services have more than one target audience/purpose) wish to do, whether you can deliver

[1] CDL's NSDL project, <http://www.cdlib.org/inside/projects/metasearch/nsdl/>.

such a system or not. To accomplish this, it may be necessary to use sample services as illustrations of specific concepts, but the goal is not to get feedback on a specific system.[2]

Focus groups in particular are useful in early stages of development. By conducting a carefully planned discussion designed to obtain participants' perceptions and points of view about a particular area of interest, the scope of a project can be explored and perhaps narrowed down. For example, the following can be learned: what features and functions users would like to see, how and why they would use the tool/service, how they feel about the tool/service, and what their needs and expectations are.

Interviews are also a particularly useful way to solicit feedback on services under consideration or development. For projects with many development stages, an *advisory group* of potential users can be useful to expand the interview concept into an ongoing dialogue that can be initiated when necessary.

Usability testing is the next appropriate step, to garner feedback on how easy or difficult it is to understand and use a specific system. Therefore, once a system is initially created (typically as a prototype), usability testing can provide focused feedback on aspects of such a system that succeed or fail. The ARL Portal Applications Working Group's 2004 report[3] provides a good survey of assessment methods used by ARL member libraries to evaluate recently built integrated access "portal[4]" implementations.

As the stages of CDL's Metasearch Infrastructure Project[5] progress, an advisory group of campus librarians will provide feedback at critical points. Once development of a prototype integrated search service has reached a point where users are able to interact with it, usability testing will be conducted[6].

Integrated Search Needs Assessment

CDL conducted needs assessment research to determine specific user needs related to integrated search tools. The research leveraged and contributed to a broader assessment program being carried out by the CDL to re-evaluate the value proposition that digital

[2] *CDL Assessment Program Toolkit*,
<http://www.cdlib.org/inside/assess/evaluation_activities/docs/2005/toolkit_2005.pdf>.

[3] Jackson, Mary E., *The Current State of Portal Applications in ARL Libraries*, 2004,
<http://www.arl.org/access/portal/PAWGfinalrpt.pdf>.

[4] The survey acknowledges a "wide range of interpretation and understanding of the word 'portal'", but goes on to describe "cross-resource searching" as generally understood to be one of the component services of a portal. (p.1)

[5] *CDL Metasearch Infrastructure Project*, <http://www.cdlib.org/inside/projects/metasearch/>.

[6] Lee, Jane, California Digital Library, Metasearch Infrastructure Project, *Core Collection Search Portal Usability Report*, November 2004 <http://www.cdlib.org/inside/projects/metasearch/core_ucsc_oct2004usability.pdf>.

libraries bring into the scholarly information space. Still further context for CDL's findings is offered by several additional UC focus groups[7].

CDL's focus groups were oriented towards three different kinds of digital libraries so as to control for any idiosyncrasies that might be associated with a particular user community or discipline. The three digital libraries were the NSDL; a virtual collection of openly accessible information pertaining to the history, culture, society, and ecology of the American West; and a collection of peer-reviewed publications assembled to support undergraduate science education. The focus groups were conducted between April and June 2004. There were 5 in all involving 35 librarians and teachers, drawn from educational institutions across California. Participants included academic librarians, public librarians, community college librarians, and K-12 teachers and media specialists[8].

A number of overall themes emerged from the focus groups, which can inform development of integrated search tools. Most importantly, where educational information is concerned, users want:

- *Speed and simplicity* of the Internet search engines (Google). Participants spoke of "a simple, uncluttered search interface." "[Users'] heartfelt and oft-repeated advice to us was 'simplify, simplify, simplify.'" "Participants would prefer to run one search and get results back from a variety of sources."

- *Convenience* of e-commerce (Amazon). Participants' Internet usage has set high expectations for a service-rich environment.

- *Reliability, authority, and integrity* of information resources that are trusted because of the brand they carry (whether imparted by a prestigious library, academic institution, professional society, or even a state education curriculum).

Additional observations can be made from the focus groups, with regards to features to be offered and desirable content types to be accommodated within an integrated search tool:

Collection building and integration features: The academic librarians we surveyed agreed about the value of well-curated online collections. Librarians want to have editorial control over the selection of what resources will constitute an integrated search universe. Tools that enable virtual collection building and federated access to those virtual collections are desirable.

[7] *Report on UC Berkeley Library Web Site Focus Groups*, Spring 2003 <http://www.lib.berkeley.edu/Staff/wag/focus_groups_report_spring2003.html>.

[8] *California Digital Library, National Science Digital Library, Focus Group and Market Assessment, Final Report*, July 1, 2004 <http://www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_assessmentfindings.pdf>.

California Digital Library, Documenting the American West, User Interviews, July 1, 2004 <http://www.cdlib.org/inside/projects/amwest/americanwest_assessmentfindings.pdf>.

California Digital Library, Core Collection Interviews, Final Report, Jul 1, 2004 <http://www.cdlib.org/inside/projects/metasearch/corecollection_assessmentfindings.pdf>.

Advanced search features: Although, overall, simplicity was desired, there were voices in favor of supplementing a simple interface with close-at-hand advanced search features. Some examples that were mentioned: keep all additional non-keyword search parameters on an advanced search page, make search history terms visible, allow iterative searching, offer ability to filter results by resource type, combine searching by subject and content type.

Personalization features: A number of focus group participants viewed their research activities as "transactions." They are interested in such features as a "personal library manager", shopping basket style saving and exporting, emailing and exporting citations in popular formats, and saving search histories for future reference and repurposing.

Interactivity: Participants expressed a desire to have "hooks" where instructional technology can connect, and permanent URLs to refer to. Linking capabilities (especially to full-text) were considered an essential feature. Users also expect to have interaction with a local library catalog within the context of an integrated search tool.

Desirable content: Although CDL's focus groups sought to gather opinions across a wide range of disciplines; several particular content types were repeatedly mentioned as potentially "more discoverable" within an integrated search environment — images, and primary source materials.

Content Discovery Principles

CDL first deployed a metasearch service in January 2000. Dubbed *Searchlight*[9]*,* the system provided UC librarians with early experience with the benefits and problems of metasearching. In addition, CDL has frequent, practical engagement with the products of most major vendors of both databases and electronic journals, and has developed many other collections and services aimed at helping users locate the information they need. Our experience has indicated that the following content discovery principles should be considered when deploying new user services.

Searching should be as pain-free as possible for users; that is, we should strive to build systems and services that join users with what they need in the simplest, easiest, and most effective ways possible. We should not expect users to learn the idiosyncrasies of our systems unless there is no practical alternative. *Content Discovery Principle #1: Only librarians like to search, everyone else likes to find.*

Most of our users do not require 100% recall for any given topic search — in many cases they will be well served by finding a reasonable number of relevant items. Therefore, we should strive to design our systems to meet the majority requirements (e.g., the 80% rule), not the minority. Serving minority needs may be met by an unobtrusive "advanced search" option, which is now a common element on most search engines. *Content Discovery Principle #2: "Good enough" is just that.*

[9] *Searchlight*, <http://searchlight.cdlib.org/cgi-bin/searchlight>.

If a given user's needs can be adequately met by searching in one place, they will generally go to that one location rather than to two or more better locations — even if those locations would result in a greater number of retrievals and/or more relevant retrievals. *Content Discovery Principle #3: All things being equal, one place to search is better than two or more.*

Finding good information means narrowing in on the particular slice of the information universe that best matches your information need. This means that information that is clearly out of scope and yet is retrieved due to the presence of the search words — perhaps within an entirely different context and with an entirely different meaning — is simply "noise" that gets in the way of the information that is sought. Therefore, selecting an appropriate slice of the information universe to search for a given need is an essential part of effective information retrieval. *Content Discover Principle #4: What is not searched is as important as what is.*

When users need help, they should travel the shortest possible distance. For example, a student at a given UC campus has a right to expect to be served by their local reference librarians when assistance is needed. Therefore, centralized services should decentralize user support by any method available. Some methods that immediately come to mind are that a central metadata store could be the back-end for a campus designed and maintained front-end; alternatively, campuses can be afforded the opportunity to "skin" (layer their own branding and navigation on top of) a central service. *Content Discovery Principle #5: Place services as close to the user as possible.*

Integration Principles

The Content Discovery Principles above indicate that the more libraries and information providing organizations can integrate access to the content that a specific user needs at a specific time for a specific purpose, the more we can meet users' needs and expectations. The problem is that this is neither easy nor even possible in all cases. But there are nonetheless some principles regarding content integration that can be identified.

Integrating access to disparate information resources is a continuum ranging from completely integrated to geographically dispersed and technically divergent. All things being equal, the best situation for the end user is a completely integrated system, with all appropriate metadata stored internally in a common format uniformly applied. This provides the greatest flexibility to meet the needs of a specific user group or purpose, since total control (both just-in-case and just-in-time) can be exerted on the system. *Integration Principle #1: Integrate metadata whenever possible.*

All solutions other than complete integration are a compromise, in which the control you have over all aspects of the system may be considerably lessened, from user search options to how results can be manipulated and displayed. Order begins to be replaced by chaos. As chaos increases, users are less well served.

One way to reduce metadata chaos is to exploit similarities between disparate metadata. For example, one record's "title" field may be similar enough to another record's "caption" field to be unified for the purposes of searching. It should be acknowledged that this process might, at the extreme, reduce the number of fields shared by all records to one, which nonetheless may be sufficient for basic discovery purposes. *Integration Principle #2: Exploit metadata similarities.*

But the reduction of metadata chaos should not come at the expense of the metadata itself. That is, rich metadata, in terms of both content and granularity, should not be reduced irretrievably for the sake of simplicity. Rather, it is better to use strategies that allow metadata to be stored in its richest, most granular form for future purposes while mapping terms and values to a common format for indexing and display. Similar to the way in which libraries create digital master files and lesser-quality display copies, libraries will need to build and retain richer forms of metadata than what may be used for indexing and/or display within any single user interface. In the end, the first law of conservation of metadata should prevail: metadata should only be created, not destroyed. *Integration Principle #3: Honor metadata differences.*

Federating metadata relating to disparate collections provides a method to present unified searching of a vast array of content. This is a good thing, for the right need and purpose. But other needs are not well served by searching across borders that the user might wish be established or preserved. For example, if the user is only interested in images they will not want to see text objects. Therefore, establishing appropriate "slices" of the federated metadata data store will be essential. Some obvious slices include by material type, contributing institution, and subject. *Integration Principle #4: Offer appropriate methods to narrow the scope.*

It may not be possible to centralize in one location all of the metadata you wish to search. Libraries today are in many cases unable or unwilling to do so, either from contractual limitations imposed by database vendors or from lack of resources or experience to manage the data locally. In these cases, which admittedly is almost universally the case today, the hapless user must identify a set of databases to search (an often daunting task in itself), travel to each in succession, learn how to use it, search it, and back out for the next one in line.

If the metadata to be searched cannot be centrally integrated in one search interface, the only remaining solution may be to use metasearch software to virtually integrate searching. Although this should be considered a solution of last resort, it is a solution that should be seriously considered anytime the user would be better served by searching multiple sources. *Integration Principle #5: If you can't centralize metadata, centralize searching.*

Integration Methods and Practices

In practice, a number of potential paths to integration (e.g., ingesting, harvesting, metasearching) offer different strengths to support these principles. For any organization seeking to implement an integrated search service, it is worthwhile to explore the tradeoffs between the various techniques that can underpin integrated search. Determinations based on user needs and usage patterns, nature of the content, and the organization's level of technical competency and resources, will all play a role. The following chart offers a summary of some of these techniques:

	Enable Content Submission (Ingest)	**Harvest Metadata (OAI-PMH)**	**Crawl Web Sites**	**Enable Content Syndication (RSS)**	**Enable Federated Queries (Metasearch)**
Relevant integration principle(s)	• All appropriate metadata stored internally in a common format uniformly applied	• All appropriate metadata stored internally in a common format uniformly applied • Honor metadata differences • Offer appropriate methods to narrow the scope	• Integrate metadata whenever possible	• Integrate metadata whenever possible	• If you can't centralize metadata, centralize searching
When is this method appropriate?	• Local collection that will be locally accessed • Content is relatively stable • Resources available to provide rich native interface	• Need access to large collections you don't want to have in-house • Need a fast search	• To provide search access to a targeted collection of web sites	• Provide access to frequently updated content or news – current awareness	• When metadata cannot be centralized • When it is too time consuming for users to access multiple resources separately • Resource discovery • When users will need to find "just a few good things" • When content is frequently updated
What are the obstacles?	• May not want to have "ownership" responsibilities • Storage space (at a very large scale)	Mostly obstacles related to providing access: • Normalization of metadata • Duplication of records –aggregate providers • Varying levels of granularity amongst digital objects • Contextualizing results And • Accounting for XML validation errors	Mostly obstacles related to providing access: • How should search results be presented? By individual web page? By web site, then by page?	• At this point in time, still a limited number of resources in this format • Range of options yet to be fully explored	• Lack of standards • Avoiding "lowest common denominator" interface – losing benefits of native interface(s) • Staff training • Maintenance time/costs • De-duping difficulties and vendor concerns about duplicate display • Vendor concerns about server overload (as target) • Contextualizing results • Inadequate or non-existent search result ranking

A suitably developed metasearching infrastructure can be used to provide a common interface to content integrated by any or all of these methods. Thus the standard metasearch application marketed by software vendors is but one piece of a robust metasearching infrastructure. Such an infrastructure must be capable of using each of the integration techniques identified in the above chart while providing a unified user interface to the whole.

Ingesting

Ingesting information resources is "the process by which a digital object or metadata package is absorbed by a different system than the one that produced it."[10] Ingesting often requires procedures that are tailored to a specific data source; for example, translation from one metadata format to another or verification that the submitted package adheres to appropriate standards (e.g., a specific XML schema).

Ingesting is appropriate for collections for which the ingesting institution is willing to be responsible (when ingesting both content and metadata) or for metadata-only collections of particular importance and for which a relationship exists with the contributing institution. Since ingesting is potentially a time-consuming activity, it usually requires particular commitment to the collection and to keeping it current.

An organization that has resources available to provide a rich native interface to specialized and or/local content can be successful in providing integrated search across ingested information resources. Storage space and costs (including staff time) may be a concern, but if resources are available, the drawbacks to this approach are few since once the metadata and/or content is ingested, many enhancement options become available (e.g., metadata normalization and enrichment).

Harvesting

CDL defines *harvesting* as "the process by which software can collect metadata packages from remote locations that describe information resources available at those locations."[11] More specifically, in common parlance the term *harvesting* usually refers to the aggregation of metadata from repositories compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Harvesting is appropriate when access is needed to collections that are not in-house, and for which ingesting is not desired or an option. In harvesting, an organization assumes a *service provider* role by centralizing metadata and offering some form of user access to the metadata provided by one or more *data providers*. Increasingly, libraries are discovering that harvesting is just the first step of a multi-step process.

At minimum, libraries will need to create either a user interface to their harvested metadata, or connect it to a metasearch tool via a protocol such as SRW/U, or both. Also, the harvested metadata itself usually comes with problems that must be addressed if full-featured

[10] CDL Glossary definition, <http://www.cdlib.org/inside/diglib/glossary/?field=any&query=ingest&action=search>.

[11] CDL Glossary definition, <http://www.cdlib.org/inside/diglib/glossary/?field=term&query=harvest+not+metadata+not+full&action=search>.

searching and browsing is desired[12]. CDL is presently creating a model for harvesting that includes post-harvest activities such as metadata normalization, subsetting (selecting a subset of records from the harvested whole), and enrichment (adding elements or element qualifications)[13]. We believe it is likely that others will need to perform such post-harvest activities as well in order to provide good user service.

RSS and similar emerging formats

RSS (often defined as Really Simple Syndication, but other definitions exist) is a method by which a small set of metadata elements (items like title, description, link, etc.) can be exposed for others to dynamically access and use. A foundational technology underlying web logs or "blogs", RSS enables lightweight current awareness services.

RSS is useful when users need access to frequently updated news or other time-sensitive content. It is also a way to provide alerts to new content that fits a user's pre-defined set of needs. Currently there may be a limited amount of content available via RSS in a given subject area, but content available via RSS is rapidly increasing.

Increasingly, RSS or RSS-like technologies are also being used in innovative ways. For example, A9's OpenSearch[14] provides a method by which search results can be syndicated.

Therefore, RSS or RSS-based technologies can offer a method by which libraries can deeply integrate content from other sites. Such implementations can be as trivial as dropping an RSS feed onto a search portal web page, or as advanced as using A9's OpenSearch protocol to integrate search results from other systems.

Web Crawling

Web site crawling provides a method to integrate access to web site content not contained in a database. Selected web sites can be crawled, with the retrieved web pages indexed into a searchable target. A challenge for this approach, however, is exactly how to make this content available via metasearching.

It's unlikely that individual web pages should be integrated with published articles without at least some indication of the type of material and the source. At CDL, we are experimenting with presenting search results from selected web sites in a sidebar separate from results from licensed databases for just this reason.

[12] Tennant, Roy. *Bitter Harvest: Problems and Suggested Solutions for OAI-PMH Data & Service Providers*, 2004, <http://www.cdlib.org/inside/projects/harvesting/bitter_harvest.html>.

[13] Tennant, Roy. *Specifications for Metadata Harvesting Tools*, 2004, <http://www.cdlib.org/inside/projects/harvesting/metadata_tools.htm>.

[14] *OpenSearch RSS 1.0 Specification*, <http://opensearch.a9.com/spec/opensearchrss/1.0/>.

Also, the granularity of the results display is a particular challenge with web sites. Should individual web pages be displayed from multiple web sites in one undifferentiated list? Or should individual pages be subsumed under the site from which they come? There is as yet not enough experience and usability testing available to provide guidance on this point, although we hope to have some evidence on this soon.

We suppose that an initial results display should first indicate the web site as a whole, with the number of matching pages indicated. Once the user click on a web site link, information for each matching page would be displayed, then if the user clicks on a particular search result, they would be sent off to the actual source of the information on the remote web site. We expect that usability testing will eventually indicate if this method is effective, or if another method is better.

Metasearching

Metasearching is the activity of dynamically searching two or more databases and presenting the results in an integrated display. Thus metasearching is a "just-in-time" unification of search targets. Federated searching, often confused with metasearching, is a "just-in-case" integration strategy, wherein the metadata from two or more databases is integrated into one system.

Metasearching is required when the organization providing a search service does not control, nor can federate access to, all the resources for which they wish to provide unified searching. It is, in other words, a compromise, but often a necessary one.

Metasearch software provides a single point of access to a potentially wide array of resources, and thus can be a powerful tool for resource discovery. However, care must be taken in crafting the universe to be searched, as each additional database searched adds to the software processing overhead and the potential of cognitive overload for the user.

Implementing metasearch software requires a major investment in time and resources. Staff must be trained, software installed (except for applications which are hosted by the vendor), and maintenance procedures put into place.

Searching multiple databases simultaneously is both a complicated and time-consuming problem. Multiple network connections must be set up and broken down, results parsed and formatted, and records de-duplicated and sorted. Each resource searched adds significantly to the processing overhead. Meanwhile, the metasearch user is presumably waiting patiently for the system to respond.

In addition, third parties are involved here. In the last few years, database providers have had concerns about server overload (as search targets) and having their citations removed from search results in the deduplication process. A lack of standards has exacerbated this situation. But as libraries, metasearch vendors, and database providers gain more experience and develop standards, this situation will likely improve.

There are also issues to be dealt with such as avoiding too much of a "lowest common denominator" interface — losing additional benefits of native interface, contextualizing results, and inadequate or non-existent search result ranking.

Despite the significant barriers to entry, metasearching may be the most appropriate integration choice for many libraries. As libraries have invested in building large collections of licensed databases and electronic journals, metasearching can provide a single point of access to these collections, or allow the library to create a search point for specific subsets of those collections.

Metasearching can potentially be used to integrate access to not only licensed databases, but also to local collections, OAI harvested metadata, even crawled web pages. The integration methods are not mutually exclusive. Indeed the prototype CDL is building will enable us to evaluate the viability of such options on behalf of the NSDL. To begin this process, CDL first sought to acquire metasearch software. A checklist of considerations was developed, which can be used as a tool by any organization seeking to do the same[15].

Metasearch Software Deployment Options

Purchasing a metasearch application is just the beginning of the long and complicated process required to implement the service. Of primary importance is the decision on how it will be deployed. Although most libraries approach metasearching with the assumption that "one-stop shopping" is the best (and perhaps only) way to deploy these systems, in our experience, it may possibly be the worst of several possibilities — at least with the current state of the software. Each of several options will be discussed in turn.

One-Stop Shopping

With a one-stop shopping type of deployment, a library creates (or uses default) subject categories for search targets, assigning each searchable resource to one or more of these categories. The user is then presented with a list of subject categories from which they must choose the best match for their query. Libraries can either search all resources assigned to a category by default, or require the user to select specific resources within a category before performing a search.

Potential benefits of this type of deployment are the relative ease of setting it up (most metasearch products assume this method of use) and one place for the user to come to search in any of a number of topic areas.

Drawbacks include the lack of ability to tailor the system for any specific purpose or audience, an interface that may be confusing or difficult for a user to understand (i.e., being faced with the need to make multiple decisions before even entering a search), and the inability to effectively present clusters of targets within a subject category (e.g., offering "American", "European", etc. as subcategories under the broad topic of "History").

Integrated With Another Web Site/Service

Another deployment option is to embed metasearch services within another web site or service. For example, a course web page or a library web site. The University of Rochester

[15] Arie, Julie, Kent Weaver, & Roy Tennant. *A Checklist of Considerations for Selecting Metasearch Software (Draft)*, 2004, <http://www.cdlib.org/inside/projects/metasearch/metasearch_checklist.pdf>.

has pioneered implementations of this nature, using a highly tailored Endeavor Encompass-based metasearch service. The "Find Articles" link from the University of Rochester Library web site pulls up a search box that is a metasearch of a few core databases. The user is not required to select any search targets to perform a search, but if they wish to select a topical heading below the search box that will take them to a page tailored for finding information in that subject area.

Audience or Purpose Focused

For those libraries with the available staff time and expertise, one of the best deployment options may be to create metasearch services that are tailored for a particular audience or need. For example, the needs of an undergraduate student who simply needs "a few good things" with which to complete an assignment or write a paper does not need an exhaustive search of topic-focused databases. For them, searching a few general-purpose databases may be sufficient. Also, since the search service is aimed at a particular purpose, additional services such as paper topic selection guides, finding appropriate encyclopedia articles, and suggesting search term synonyms may be useful and appropriate.

A search portal for graduate students and faculty within a specific discipline may have a very different feature set. One could easily imagine the utility of integrating an RSS feed of new resources appropriate to a specific discipline on the front page of such a portal, for example. Likewise, it may make sense to integrate access to selected web sites important to a discipline and/or OAI-harvested metadata appropriate to that topic area. There are many possibilities, and only through needs assessment, experimentation, and usability testing will libraries come to know what works and what doesn't.

Technical Considerations

No matter which deployment option is selected, there are key technical issues with deployment that must be considered. Some metasearch vendors (e.g., WebFeat) expect to host your metasearch portal as an Application Service Provider (ASP). Others will expect you to install and run it on local servers.

Some applications will have user interfaces that are much more malleable (e.g., Endeavor's Encompass system uses XML and XSLT to present the interface, which offers a great deal of control), while others are not so easily changed. Another option for controlling the interface is to take total control through using an Application Program Interface (API) rather than the native interface to access the application. For example, by using the X-Server XML gateway to ExLibris' MetaLib product, libraries are free to create whatever user interface they wish, while leaving the more difficult parts (e.g., simultaneous searching, deduping, etc.) to the application.

The State of the Metasearch Market

CDL's initial evaluation of metasearch products was conducted in 2002-2003.[16] Six products were identified via a survey of literature and UC-wide recommendations. Vendors were asked to demonstrate the products, product literature was collected, and live systems were tested either as installations on the vendor site or at customer sites. Follow-up questions were asked of existing customers.

Since then, several good comparative studies have been done which outline existing and emerging product capabilities. In particular, Daniel Dorner and AnneMarie Curtis' June 2003 report commissioned by the National Library of New Zealand, which characterizes desirable features as established, maturing, emerging, or not yet available.[17]

Martha Brogan's report on digital library aggregation services is a good survey of what aggregation services libraries are building.[18]

In addition, the CDL's recent activities in acquiring metasearch software provided further observations about products currently available. Together, these reviews and observations can give context to any new evaluation of products. The following are some general expectations for metasearch product capabilities, at this point in time.

Most vendors support the following:

- *Communication protocols*: Z39.50, HTTP (screen scraping), OpenURL, XML, Dublin Core, MARC, SQL.
- *Platforms*: vendor-hosted service, Windows, Unix. Some vendors support: Linux, Sun Solaris.
- *Modes of authentication*: LDAP IP address, domain name.
- *Administrative features*: Web-based admin interface, configurable statistics reporting through admin interface.
- *Search features*: merging and de-duping, post-search filtering and sorting, no limit to number of databases searched simultaneously, selection and manipulation of citations, alerts.
- *Interface customization options*: configurable patron interface, customization of interface to include library name, link, logo.

The following capabilities are still not widely adopted:

[16] The initial findings of this research were reported by Christy Hightower and Catherine Soehner, UC Santa Cruz.

[17] Dorner, Daniel & AnneMarie Curtis. *A comparative review of common user interface software products for libraries*, 2003, <http://www.natlib.govt.nz/files/CUI_Report_Final.pdf> .

[18] Brogan, Martha. *A Survey of Digital Library Aggregation Services,* Digital Library Federation and Council on Library and Information Resources, 2003, <http://www.diglib.org/pubs/brogan/>.

- *Platforms*: Linux, Sun Solaris.
- *Administrative features*: statistics at the level of members of a consortium.
- *Modes of authentication*: Shibboleth

For more details on specific products, it may be useful to consult the Library of Congress Portals Applications Issues Group's list of vendors and products.[19]

Competing Services

Recognizing the need for integrated search, a number of major database vendors have upgraded their services to allow search across multiple databases hosted on the same platform. Although this type of integrated search is limited to the specific databases offered by each vendor, it is well received by users, especially in cases where the databases are in a common subject area (e.g. ProQuest's news resources, CSA's social sciences resources).

Emerging services such as Google Scholar and Yahoo Search Subscriptions are also calling into question the need for libraries to offer a metasearching capability to their users. If a site like Google Scholar is providing search access across a wide array of resources, why do libraries need to build these kinds of costly and time-consuming services?

Indeed, these services may one day replace the need for libraries to metasearch, but if that day ever dawns it will not be soon. As relatively powerful as they may be in terms of certain technical aspects, they remain relatively primitive information finding tools compared to the kinds of services that libraries are beginning to deploy.

A simple example may prove illustrative. Imagine an undergraduate student is seeking some basic information on tsunamis – what they are, how they are generated, and the type and scale of damage they are capable of producing. Searching "tsunami" in Google typically returns a mix of a few good hits on the first page of ten results, along with links to sites providing tsunami relief and the ubiquitous paid ads. The same search performed in Google Scholar returns a set of very technical and narrowly focused scholarly articles. The National Science Digital Library, in contrast, returns a set of highly appropriate results sans advertisement.

Google, Google Scholar, or any general-purpose search tool cannot possibly serve all needs well enough to prevent the need for other search tools focused more specifically and appropriately on a particular set of user needs. A recent CDL survey supports this conclusion, showing that librarians view Google Scholar as one amongst many possible choices.[20] Librarians, in other words, still need a robust set of tools with which they can craft a set of search services tailored to a specific audience and/or need. In-depth evaluation

[19] Library of Congress Portals Applications Issues Group, *Federated Search Portal Products & Vendors*, <http://www.loc.gov/catdir/lcpaig/portalproducts.html>.

[20] *UC Libraries Use of Google Scholar*, August 2005, <http://www.cdlib.org/inside/assess/evaluation_activities/docs/2005/googleScholar_summary_0805.pdf>.

of the particular shortcomings of Google, Google Scholar, Yahoo Search Subscriptions, and other such services is out of the scope of this report, but the reader may wish to perform their own comparison using the principles outlined in this report. The only thing that is completely clear, however, is that we are in a period of great change and the situation tomorrow will be very different from today.

We must prepare to adapt as needed to the changing information environment around us, and the to do that is to have a robust set of tools at our disposal to allow us to search what we wish to search and (perhaps more importantly) not what we don't wish to search. Relying on any single search service is unlikely to be the best path forward no matter what tomorrow brings.

Emerging Standards and Best Practices

There are a number of notable emerging standards to be aware of, which will likely enhance metasearch capabilities in the near future:

The OpenURL standard (NISO 39.88 – 2004)[21] is a key part of getting the user to the appropriate copy of a journal article or to many other resources or services (for example, interlibrary loan). Thus it is an important part of any metasearch service, and is a standard offering of vendors offering metasearch products.

The NISO MetaSearch Initiative[22] is fully launched into the process of developing standards, best practices, and tools to make the metasearch environment more efficient for the system provider, the content provider, and the end-user.

The Digital Library Federation and NSDL OAI and Shareable Metadata Best Practices Working Group[23] is developing best practices for Open Archives Initiative[24] data and service providers.

The SRW/SRU[25] "next generation Z39.50" emerging standard is a vital standard for metasearch services, since it defines a robust set of XML-based services that can enable rich metasearching of search targets.

[21] NISO OpenURL Standard (Z39.88-2004), <http://www.niso.org/standards/standard_detail.cfm?std_id=783>.

[22] NISO Metasearch Initiative, <http://www.niso.org/committees/MetaSearch-info.html>.

[23] Digital Library Federation and NSDL OAI and Shareable Metadata Best Practices Working Group, <http://oai-best.comm.nsdl.org/>.

[24] Open Archives Initiative, <http://www.openarchives.org/>.

[25] SRW/SRU, <http://www.loc.gov/z3950/agency/zing/srw/>.

A9's OpenSearch offers a very low-overhead protocol for databases to send search results in XML. Although this standard is initially meant to allow databases to integrate with the A9 search service, the specification may be used in any context in which developers find it to be useful.

Conclusion

As of this writing, the metasearch software market is still in early days. Every application is painful to implement in greater or lesser degrees, and in often very different ways. Purchasing such software is as much an exercise in determining where one wants to feel pain as it is in deciding with which company you wish to form a long-term relationship.

Having acquired an application, key decisions must be made about how to deploy the service. We believe that although one-stop shopping may be appropriate for some libraries (such as small ones), many libraries will need to consider tailoring their services to specific audiences and purposes.

The marketplace and technological landscape will change, but the approaches, principles and practices outlined in this report should remain applicable for anyone who is evaluating the options for providing integrated search options to their organization.