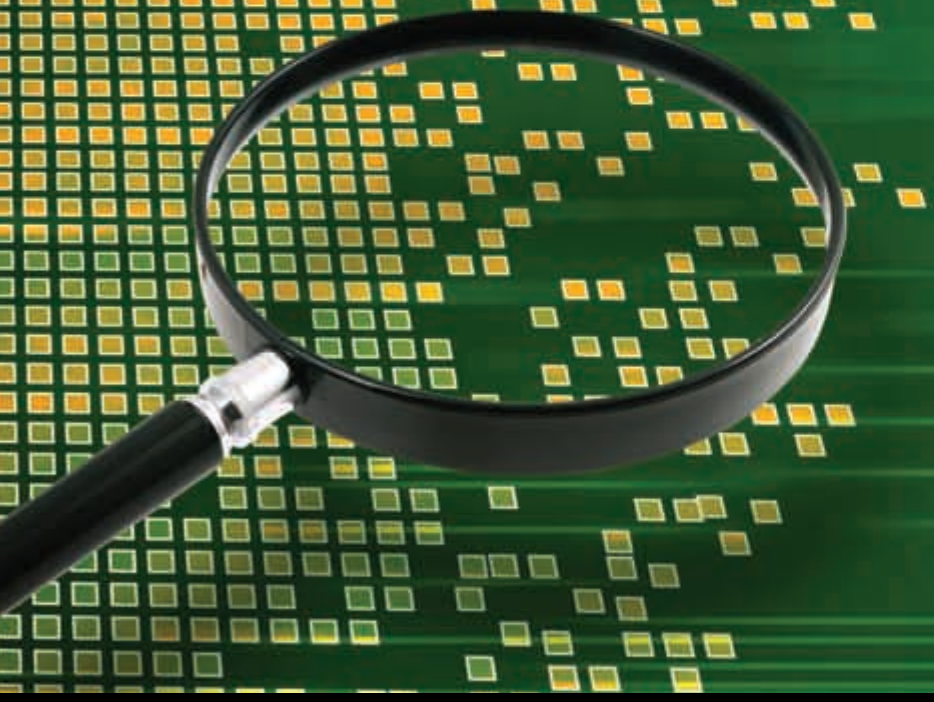
Federated Search

For Your Library and in Your Enterprise



Thanks to the sponsors who helped to bring you this issue of *Computers in Libraries*.

Swets
Building a Better Search Query
Through Content Mining

Gale
Discover: It's Not About Federated Search, it's About Discovery

Ex Libris
Primo Discovery and Delivery—Beyond the OPAC
A Unified Interface for Finding and Getting All Library Resources

Information Today, Inc.
Taming Multiple Search Engines In Your Organization
By Jean Bedord

Building a Better Search Query

Through Content Mining

SWETS
Swets Simplifies

Federated Search Alone Is Not the Complete Answer

With the amount of online information rapidly expanding and residing in increasingly disparate sources, organizations need a way to simplify how their users discover and access the information they need. Federated search is designed to help organizations meet this challenge, enabling users to simultaneously search multiple sources and quickly obtain relevant results using a single search query.

Yet federated search on its own can often lead to information overload, making it time-consuming and difficult for users to locate the most relevant information. We will examine how SwetsWise Searcher combines content mining capability with federated search to offer the most effective solution. More than just a categorization tool, content mining does not force categories on users, but instead "mines" content and guides users to build the most effective search query for the most relevant results.

Why the Content Mining Approach Was Developed

Developers of the content mining feature recognized several issues with standard federated search options. As they analyzed the offerings, they realized that introducing the ability to simultaneously search multiple sources can quickly lead to information overload. Users are confronted with sorting through lists of hundreds, even thousands of search results, with relevant information buried among randomly scattered topics. They spend an immense amount of time and effort browsing, analyzing, and interpreting the results.

Along with wasted productivity, information overload negatively impacts the quality of research. Time forces users to ignore many results, and they miss critical information that is buried in random lists. It also prevents discovery of new content and relationships that are relevant to users' research.

The quality of returned search results is also driven by the effectiveness of the user's search query—something federated search on its own does not address. A user's first search query frequently produces results sets that are not the best, or are completely unrelated to their research needs. This is compounded by user search behavior that is conditioned by web keyword searching. Users often begin with very broad search queries, relying on the speed and extensive indexes of internet search engines, revising their query based on the results they see returned.

Because the most relevant results are often not in the initial results set, categorization and grouping tools are not always useful. Presented with the grouping, the user drills into the result set, only to find it is not the information they needed. In addition, categories can rely on taxonomies or other pre-indexed material and often do not help users generate the most effective search query.

All of this leads to users searching over and over, refining each search query until they achieve better results—a time-consuming query/search/re-query process. Furthermore, this process is more difficult to reproduce with bibliographic, citation, and full-text databases.

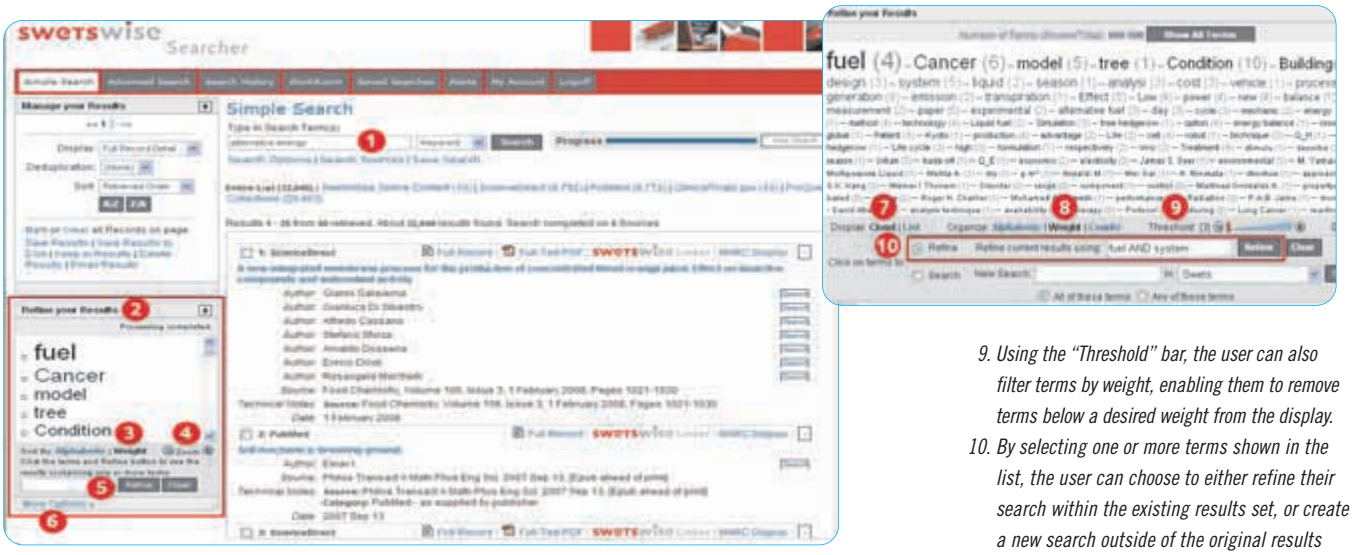
A Better Search Query for More Relevant Results

SwetsWise Searcher's unique content mining functionality "mines" terms and phrases from content retrieved "on the fly." It then analyzes the terms and produces a term weight, determining the relative importance of that term within that set of results. This content mining process is completely independent of other categorization products and does not rely on taxonomies or other pre-indexed material. In addition, it is applied to content from all sources searched with the SwetsWise Searcher application, including databases not indexed by web search engines.

Terms and phrases are mined and weighted according to where they occur in the document. For example, terms from important fields, such as title, are given a higher weight than those found only in the description of a record. Term weight is then represented visually by font size—the more important the term, the bigger and bolder its font—making it easy for users to identify the best terms for refining or expanding their search query.

The number of results where a term is found is also displayed next to the term, and a value is applied to each occurrence of the term within a document, with higher values assigned to more important fields. This method produces a weight reflecting the actual content of the result, rather than a weight simply reflecting the title or tag of the result. By showing both term weight and result count, users can easily distinguish results where the target terms are central from those where the terms simply occur frequently or are included as minor points. As a result, users can focus on major or minor ideas within a results set.

Users can choose to display the mined and weighted terms in a list or cloud view —a methodology familiar to many users, as it is often utilized in blog sites and interactive forums that incorporate Web 2.0 methodologies. They can then easily select one or more terms to refine their initial search within the existing results set, or create an entirely new search, enabling them to quickly and effectively "drill down" or "expand out" without starting over. As a result, content mining guides users to intuitively build the most effective search query to obtain the most relevant results.



Content Mining Demonstration

1. *A search query for "alternative energy" is entered.*
2. *In the initial content mining view, mined and weighted terms are listed in the "Refine your Results" window.*
3. *The default list is sorted by weight, with terms of greater weight appearing in a larger font size. By selecting the "Alphabetic" link, the user can also sort the list alphabetically.*
4. *The user can also adjust the font sizes in the "Refine your Results" window by using the zoom control buttons.*
5. *By clicking on one or more terms shown in the list and selecting the "refine" button, the user can refine their search within the existing results set. A new window will appear containing the results that include the selected term or combination of terms.*
6. *The user can also click "More Options" to display additional tools that help refine the search.*

MORE OPTIONS

7. *Terms are displayed in a cloud view by default. The user can also choose to display the terms as a list, as shown in the earlier initial view.*
8. *By default, terms are once again sorted by weight, with terms of greater weight appearing in a larger font size. The user can also sort the terms alphabetically or by the number of counts. A term's count refers to the number of results where that term is found. The greatest number of term counts does not necessarily result in the greatest term weight, as explained earlier in the article.*
9. *Using the "Threshold" bar, the user can also filter terms by weight, enabling them to remove terms below a desired weight from the display.*
10. *By selecting one or more terms shown in the list, the user can choose to either refine their search within the existing results set, or create a new search outside of the original results set. The user can also conduct the new search using all of the terms in the query, or any single term in the query. In this example, the user chose to refine—or "drill down"—within the existing results set.*
11. *A new window now appears containing the new results that include the selected term or combination of terms. Both the original search query and the terms used to refine the query are clearly displayed. Also note that the list of mined and weighted terms has changed, reflecting the "on the fly" analysis of content within this new results set.*

Greater Productivity and Improved Research

To be most effective, federated search requires a tool that helps users more easily survey the information landscape, improves relevancy of results, and leads to content discovery.

SwetsWise Searcher addresses this need by applying content mining to the federated search environment. This combination of technologies results in an extremely powerful method for conducting more effective research, guiding users to build the most effective search query for the most relevant results. Users become more productive and are able to make better decisions during their research process, as they are less likely to overlook relevant information that is otherwise buried in long lists of randomly scattered topics.

To learn more, visit **WWW.SWETS.COM** or contact **SWETS** at 800-645-6595.

This article has been brought to you by the publishers of *EContent: Creating, Distributing, and Managing Digital Content.*

Discover:

It's Not About Federated Search, it's About Discovery.

GALE CENGAGE Learning

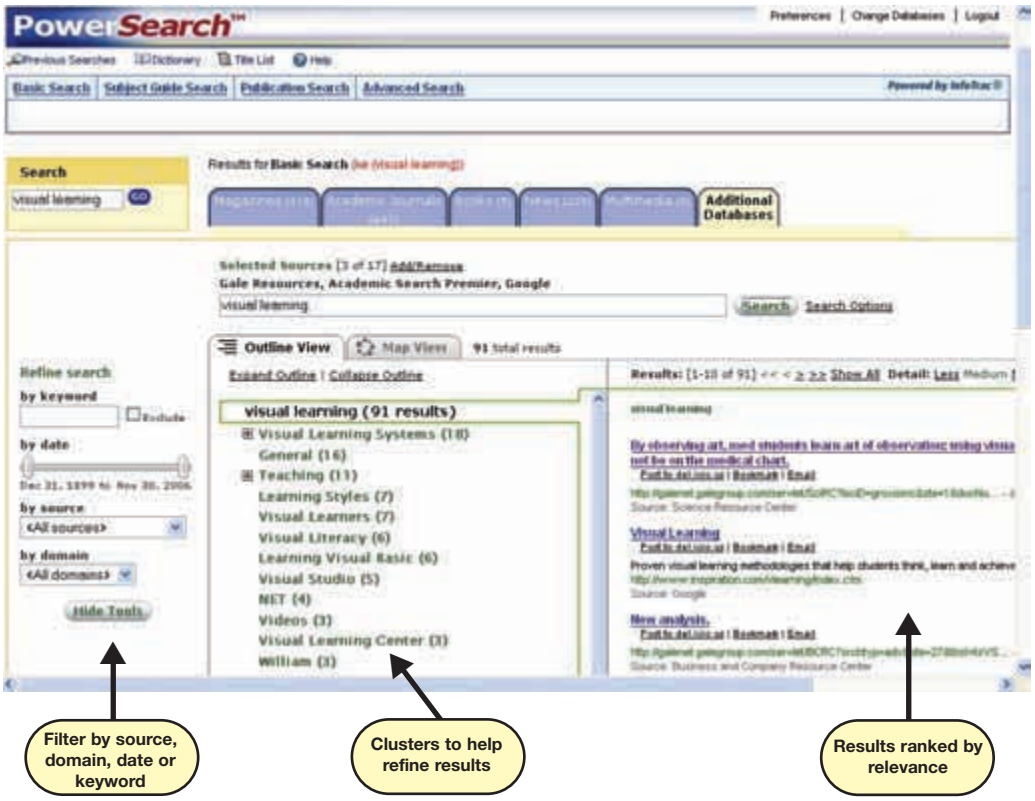
Federated Search does not have to be expensive, hard to install, slow to load and impossible to use.

In collaboration with Groxis, Gale's PowerSearch Plus changes the game

Launched in January of 2008, Gale collaborated with Groxis, an established San Francisco technology company to launch one of the most easy to use, easy to install, and affordable search tools available. As an enhancement to Gale's *PowerSearch* platform, which is installed in 60,000 libraries across the world, *PowerSearch Plus* allows libraries to access all Gale databases, library catalogues, databases from other vendors, and websites in one environment. While some call this federated search, *PowerSearch Plus* is really a great deal more.

In developing this tool, Gale and Groxis decided to take a different approach. *PowerSearch Plus* is not intended to be a final destination. It's more of a portal that allows patrons to easily find and access all the great resources found in a library -- whether print or electronic. While it offers strong search technology, accurate clustering and visual representation of results, usage reporting and more, *PowerSearch Plus* makes it easier for patrons to find your catalogue, databases and vetted websites. The hope is, through this discovery, they will come back time and time again to these resources when they have specific needs. In contrast, they will use *PowerSearch Plus* when they don't know where to start.

Currently, federated search installations can be long, rigid processes that cost significant budget dollars. Because Gale views this as a utility, and Groxis has created a tool that is significantly easy to use and install, Gale is able to provide a subscription to this federated access at a cost that is nominal compared to other federated search installations. It's completely flexible, with full administrative capabilities that allow libraries to make changes to defaults daily. This makes it available and reasonable for libraries of all sizes and all budgets. It's about equal access, after all.



This unique offering has already been installed in scores of libraries. By listening to users and customers, and developing these tools to addresses their specific needs, accolades follow. *PowerSearch Plus* the only library search tool to be named a 2008 CODIE finalist. The *PowerSearch* platform was recently awarded its second "Most Improved Product" award in a row by the *Charleston Advisor*. We expect to win that award every year.

How easy is it to set up PowerSearch Plus?

1) Choose the databases and library catalogues you want to connect to. Gale does not charge you for searching websites or Gale databases.

2) Specify your defaults -- the number of databases you want to search every time, the length of time to wait for each search, the number of results returned, etc.

3) Add a search box to your web page.

4) In hours, you're done.

Call Gale 1-800-347-4253 or visit www.gale.com/powersearch for more information.

Primo Discovery and Delivery—Beyond the OPAC

ExLibris
The bridge to knowledge

A unified interface for finding and getting all library resources

Heather strides quickly into her usual haunt, looking around anxiously to ensure that her favorite seat is still vacant before pulling out her laptop and getting down to writing her report. There are only two days left to complete this challenging assignment. All around her, fellow students hunker over their laptops, feverishly gathering research material. Heather draws a deep breath and savors the reassuring aromas: coffee, cinnamon rolls, croissants. What happened to the smell of print and leather book bindings? The university library is still there, but Heather and her peers are accessing its collections from their favorite coffee shop via their library's Primo® discovery and delivery system, which can be accessed from the browser search box, iGoogle home page, and college portal.

Primo offers Heather a single, intuitive interface for finding all the information she needs from diverse resources such as books, e-books, print and electronic articles, Web pages, and digital media from both local and remote collections. The findings are enriched by a mashup of additional data, such as abstracts, tables of contents, and book jacket images, so Heather is spared a search in multiple systems. Primo groups search results by facets, enabling Heather to tag and comment on them for her own benefit and for that of other users.

Local library data, such as the library catalog, digital collections, and course management systems, is harvested and indexed through the Primo publishing platform. Remote resources, seamlessly discovered via the Ex Libris™ MetaLib® premier metasearch solution, complement local data. Integrated with Primo, MetaLib provides a coherent, one-stop information environment. MetaLib also runs as a stand-alone product with its own user interface.

Just as she finishes the first section of her paper, Heather's older brother, Justin, stops by. He graduated from the same college five years ago and is in awe of the changes that have taken place in library services in such a short time. Justin remembers his freshman year in college. He and his classmates used on-campus workstations to search for research material from the library catalog and other information sources.

This search method was quite time consuming, as searching had to be done one resource at a time. To use resources effectively, students had to familiarize themselves with each and every one. Once a reference to an article of interest was found, it was often difficult to locate the article, since it typically resided in a different resource from the one with the reference—and no direct link to the article was provided.

The development of metasearch engines, enabling users to search multiple local and remote resources simultaneously, and OpenURL link resolvers, providing context-sensitive links to articles' full text and other library-defined services, such as print holdings in the library catalog, document-delivery forms, reference managers, and copyright clearance, made it much easier for Justin to access most of the information he needed during his college years.

Using MetaLib, Justin was able to search multiple heterogeneous resources at one time. The button providing access to the Ex Libris SFX® link resolver enabled him to link to the full text of articles. The SFX button could also be found in many information resources, such as PubMed® or ScienceDirect™.

By the mid-2000s, MetaLib's capabilities were expanded through the development and adoption of new library and industry standards such as SRU/SRW and NISO MXG.

Heather and Justin order another coffee and decide to share a dessert, imagining how easy it will be for their children to find scholarly information when they go to college. In actuality they realize that the revolution is really not that far away. If the world has gone from searching for print journals in the library's collection to accessing all types of information using Primo in just a few short years, Heather and Justin can be sure that more changes will soon follow.

The OpenURL standard (Now NISO Z39.88-2004) was originally developed by Herbert Van de Sompel, then of Ghent University in Belgium (currently at the Los Alamos National Laboratory); Oren Beit-Arie of Ex Libris; and Patrick Hochstenbach. To find out more about the OpenURL standard, see http://www.niso.org/standards/resources/OpenURL_FAQ.html.

Taming Multiple Search Engines In Your Organization

By Jean Bedord

One-size-fits-all enterprise search is dead—if it ever existed. At the same time, search that makes an organization's assets more readily accessible has become a critical mission. Customers, partners, and employees all expect to be able to "find answers" via search technologies.

Thus, organizations increasingly deploy search engines to help satisfy these expectations. These engines are typically called site search, department search, intranet search, or application search, rather than enterprise search.

However, the fact is that organizations that have one search application usually have several. A 2007 enterprise search survey I did for the publisher of the *Enterprise Search Sourcebook*, Information Today, Inc., and my employer, Shore Communications, found that 62% of respondents had more than one search solution in place. Noted industry analyst Steve Arnold reports that the typical Fortune 500 company uses solutions by at least five search vendors.

This proliferation (and multiplicity) may not be obvious, even to those within the companies deploying the solutions, since responsibility for search strategy is frequently undefined within the organization. One reason for this is because search doesn't exist independent of the content that is being accessed. It creeps in with add-ons to content management systems (CMSs), business intelligence (BI) applications, records management, document management, or knowledge management systems. In addition, ecommerce systems with baked-in search capabilities can add to the growing number of search vendors used by an organization. As an independent application, search can be tuned to specific business-process needs, which means it comes in many guises.

Thus, many companies face more than the already challenging issue of managing a single search engine: As the number of solutions deployed within an organization increases, distinct issues emerge in reconciling these disparate solutions. The major problem is that each search product builds indexes to the existing content using a "secret sauce" that's incompatible with other vendor indexes. Thus, finding answers across different application repositories, each with its own index, is not a straightforward proposition.

Strategic Planning for Search

One enterprise search myth is that all information in the organization should be accessible via search. Organizational content, however, is more difficult than the relatively simple world of consumer search on the open web, which is primarily HTML webpages and unstructured content. Information created and controlled by the organization is complex, from the content perspective and the technology perspective.

There are some facts to consider about content and how it relates to search functions:

The value of content within the organization varies. Indexing a server that contains years of cafeteria menus is a waste of resources, creating yet more information overload. Indexing technical reports representing the intellectual capital of years of R&D, thus enabling idea discovery, has potentially high ROI. Improved findability for ecommerce companies, particularly those with large product catalogs, can result in increased sales.

Job functions require different content. A financial analyst may need current sales by product line, even down to the individual part number. Customers and partners want data sheets with technical specifications. Defining subsets of content for job functions is the domain of management, not technology.

Security complicates search. Privacy of employee records is essential, though employee benefit information should be made available to everyone. Access to customer records may prove useful for certain job functions, but it may be regulated by federal and state laws. Access to trade secrets is typically restricted to those with a need to know. Real-time identification of fraudulent transactions may be mission critical. Thus, access and security policies need to be in place before unleashing the power of search technology.

Technology Planning for Multiple Vendors

Another myth of enterprise search is the feasibility of standardizing on a single search vendor. Search engine software creates proprietary indexes and relevancy ranking, with each having different strengths and weaknesses. Products from the same company, say IBM or Autonomy, do not necessarily provide compatible upgrade paths as search applications grow. There are some simple facts every company must face when attempting to reconcile search solutions:

Legacy systems are part of the search landscape. Many business intelligence and content management vendors integrate search into their product suites with OEM relationships with enterprise search companies. This ensures best-of-breed functionality within the suite, but not necessarily between vendor suites.

Mergers and acquisitions play havoc with standardization. Acquiring a company is based on business fundamentals, not compatibility of software systems. Provided the existing systems work well at the application level, there is little ROI in switching search

software. However, the consolidation of software as search companies are acquired by major players can result in product "orphans," leaving your organization without support.

Federated search needs to be part of the technology plan. Federated search refers to the capability to search multiple indexes without creating yet another index. Also called meta-, blended, or universal search, this capability can be implemented in various ways: real-time federated searching using on-the-fly queries, portals, webpage mashups, or structured top-50 queries. Implementation is highly dependent on access and security requirements, which can vary with each of the underlying content repositories.

Care and Feeding of Indexes

Search engine indexes and the information retrieved from enterprise repositories rely on the accuracy of the underlying data. Maintaining integrity of the applications that create content is crucial for business decision making, yet organizations typically underestimate the time and effort it takes to maintain clean and relevant content repositories.

Here are some issues that must be considered:

Records management is an underappreciated aspect of search. Regulatory compliance and ordinary retention schedules should be part of managing business risk. Indexes exist separately from the content, so they have their own update schedule. Overnight or weekly—rather than real-time—updates can make search results appear out-of-date. Search technology makes misspelled, noncompliant, and dirty data glaringly obvious.

Indexes grow as the numbers of records applications grow. Index files vary in size but can be an additional 50% to 200% larger than the size of the primary data files, depending on the expansion factors. Scalability of the search solution should be a major consideration in selection and implementation.

Relevancy of search results depends on context, not popularity. The purchasing department may work with a dozen "business card" vendors worldwide. A "business card" query for most individuals in an organization, however, means a short how-to procedure for ordering new business cards. Search technology alone doesn't solve the job of interpreting the query to provide relevant answers.

Managing expectations is a major challenge for search implementations within the enterprise. Search is a tool to provide answers to problems. Finding those answers across enterprise repositories requires a higher level of relevance than open web consumer search, yet the underlying content structure is more complex and access is more constrained. Just as there are no one-size-fits-all answers to organizational questions, enterprise search is not one-size-fits-all, and most organizations will need to manage multiple solutions to supply the different answers individuals seek.

JEAN BEDORD (jean@bedord.com) is a findability and search consultant with EContent Strategies, senior analyst at Shore Communications, Inc., and faculty lecturer at San Jose State University. Her report on "Enterprise Search Deployment, Usage, and Trends" is available at www.enterprisesearchcenter.com/Reports/Details.aspx?ResearchReportsID=38.]

This article has been brought to you by the publishers of ***Enterprise Search: Deployment, Usage, and Trends***.

SPONSORS

Swets
160 Ninth Avenue
P.O. Box 1459
Runnemede, NJ 08078

Phone: 1-856-312-2690
Toll Free: 1-800-645-6595
Fax: 856-312-2000
Email: info@us.swets.com
www.swets.com

Gale Cengage Learning
27500 Drake Road
Farmington Hills, MI 48331

Phone: 1-800-877-4253
Fax: 1-877-363-4253
Email: gale.customerservice@cengage.com
www.gale.com

Ex Libris
1350 E. Touhy Avenue
Suite 200E
Des Plaines, IL 60018

Phone: 847-296-2200
Toll Free: 1-800-762-6300
Fax: 847-296-5636
Email: infousa@exlibrisgroup.com
www.exlibrisgroup.com

Information Today, Inc.
143 Old Marlton Pike
Medford, NJ 08055

Phone: 609-654-6266
Fax: 609-654-4309
Email: custserv@infotoday.com
www.infotoday.com

Michael V. Zarrello
Advertising Director
609-654-6266 x132
mzarrello@infotoday.com

www.infotoday.com