



# EContent™

This article is reprinted with permission from EContent magazine, June 2006. © Online, a division of Information Today, Inc.

ron miller

## Complex Search Your @ Web Service

As enterprise users face growing repositories of valuable data, it becomes increasingly important to be able to search across these data storehouses (and the Web) to find the best available answers to a query, then to effectively apply that data. While enterprise search technology has been capable of searching multiple repositories for some time, it required a great deal of programming and computing overhead and didn't necessarily allow users to manipulate the results. XML and other Web Services have changed all that, making it possible to search multiple repositories and across various Web sources and then use the data in various ways, while using fewer resources.

This means that the enterprise user can not only produce more pinpointed search results but also use that information in interesting ways as programming logic is applied to those results. For owners of content, Web Services provide a way to

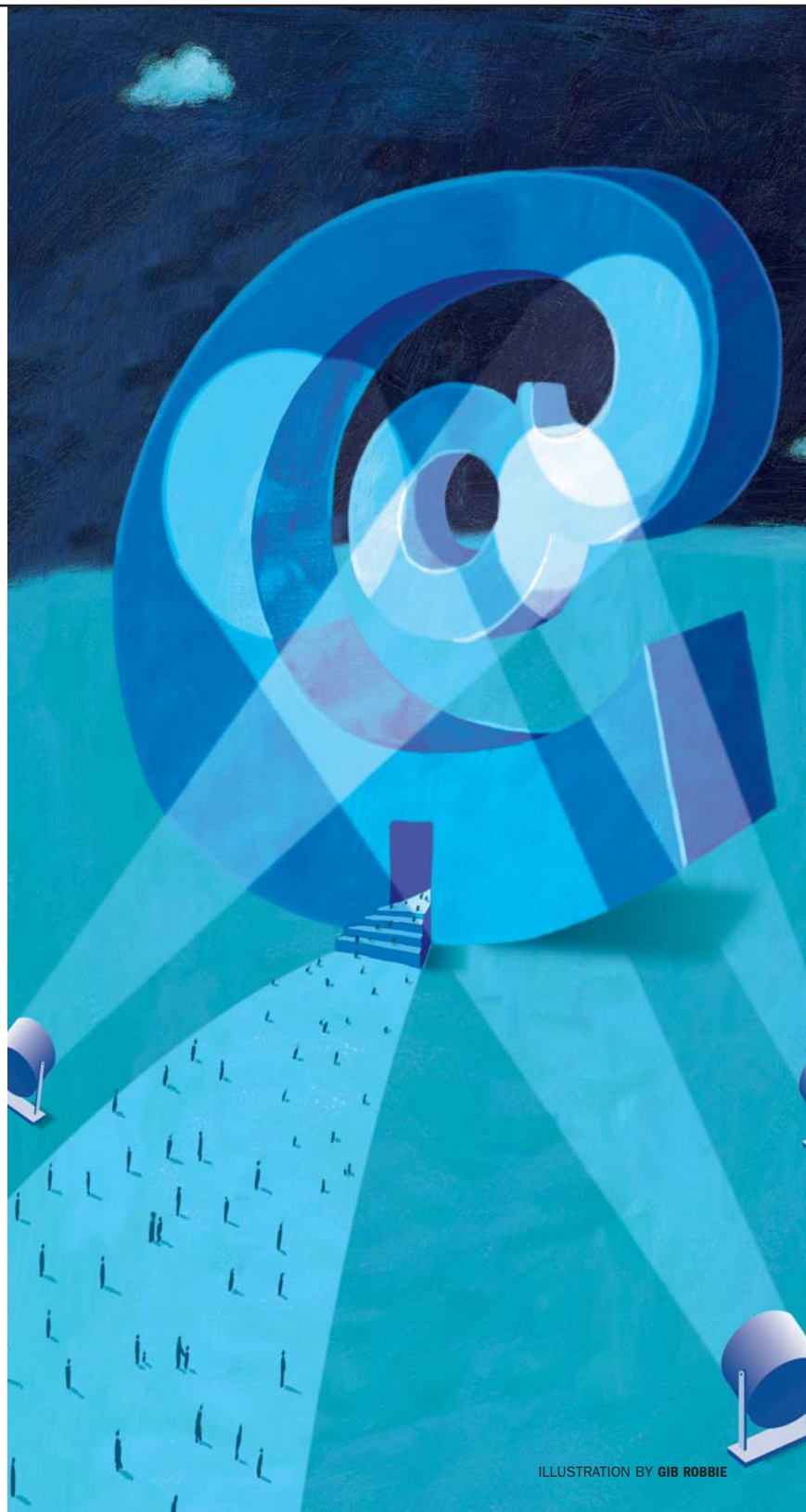


ILLUSTRATION BY GIB ROBBIE

use the metadata associated with content to slice and dice the content in new and meaningful ways for customers and internal users alike, to combine the data with other Web Services or internal applications, or to incorporate data from searches into the fabric of the enterprise where it is needed most. What's more, search engine companies have opened

up their search components (and other services) as Web Services and invited programmers to come up with creative ways of producing custom sets of results.

### SEARCHING FOR MEANING IN WEB SERVICES

Web Services have been around for a number of years and take advantage of the ubiquity of the World Wide Web, metadata tagging, and coding standards like XML to deliver a unique solution inside a Web browser. Tim Matthews, co-founder and VP of marketing at Ipedo, an enterprise information integration company, says that when it comes to search as a Web Service, it's all about what users can actually do with the results. "In general, I would say if you are looking to do something programmatic with the results, search, as a Web Service, gives you a great advantage because it normalizes the search results in a form you can manipulate." This means that a search solution can examine the tags (metadata) associated with a particular item and simply render the results in a page, or more importantly, use the tags to combine with other sources to produce more meaningful results.

together in a single list of results.

With Web Services, instead of using a program to retrieve the data, the data is tagged with metadata and compiled. Then, using this information, you can build programming logic to manipulate the results. Bradley Allen, founder and CTO at enterprise search vendor Siderean, says it doesn't affect his company's application whether the content comes from a structured database with relational tables, a content management system exposing XML feeds, or even information out on the Web. "We bring that information into a central metadata repository set up to extract information from those sources, categorize them, and put them into the context of a metadata-level description, and then index those in a way to provide low latency navigation services." He says this information can then be presented as a page of results or the user can slice and dice it or take queries and turn them into RSS feeds that they can plug into other applications such as an RSS reader. Schireson says this type of data flexibility is not possible in a traditional search model without significant coding and performance overhead.

### A DIFFERENT KIND OF SEARCH

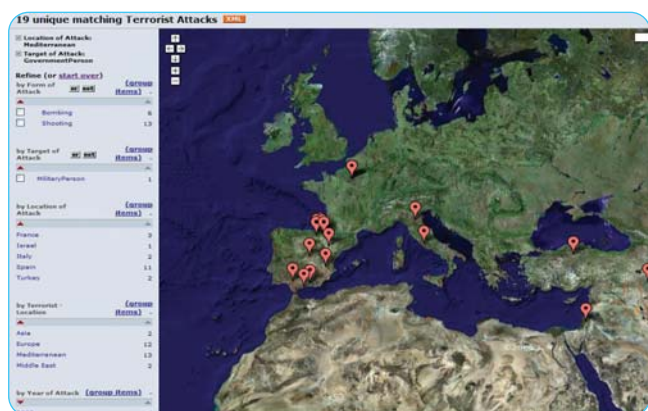
Traditional search produces a set of results, but the search tool only looks at text and does not have the ability to manipulate the results. Schireson says, "The first major limitation of traditional search is that it really looks at documents as being text, and most content has some structure associated with it, which is ignored in traditional search engines." This structure is the metadata associated with the result, and by taking advantage of this metadata, users can achieve finer control over the results.

"Traditional search engines," says Schireson, "just return a URL and a snippet of text." He says a Web Services query allows you to be more specific about the query because the markup can be more finely grained, and as a result, the search results can be all that much more precise. "Our server allows you to manipulate and render content to find the most relevant paragraph on a given subject, and return not just the URL of the book where it occurred, but return the finest-grained sub-section around that paragraph," he says. This approach allows the user "to understand the information in context and where it fits in their overall work."

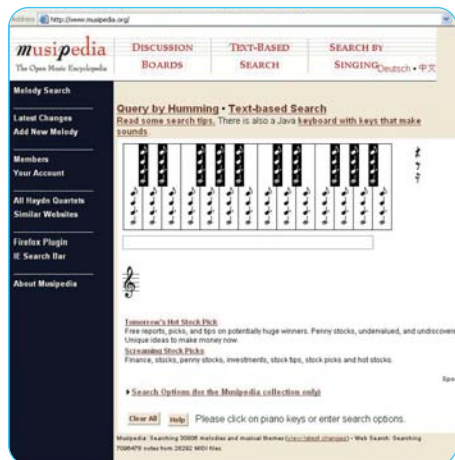
As an example, Schireson says one of his customers, science and medical publisher Elsevier, was looking for a way to better



Using the Siderean search tool, users can isolate videos based on tags, in this case: markets, food, and color.



Siderean took the results of a search for terrorist attacks, then used these to pinpoint the location of each attack on a map. Such manipulation of results is part of what makes search, as a Web Service, so useful.



Using Alexa's Web Services, this developer tapped into a database of midi files and built a Web search component that provides a way to search for music by melody.

Prior to the Web Services model, companies typically used a database to locate information, but this approach had inherent limitations, according to Max Schireson, VP of customer solutions at Mark Logic, which offers an XML content server product. "Traditionally, if you wanted to go beyond text search, you would have to take content and shred it in a relational database. It required that you know a lot about your content to be highly consistent, and you lost your flexibility going forward."

Another way this might have been done, according to Frank Gilbane, CEO of Bluebill Advisors, a content management consultant and analysis firm, is to build a federated search, a program that checks several different repositories and builds queries across these different systems, then pulls them

understand its vast repository of data and target the results to the needs of the individual user, to present them in context of need. "Elsevier is the world's largest scientific, technical, and medical publisher. Their big asset is a bunch of content." Elsevier executives wanted to know how "to maximize the value of this content to the company by maximizing the value to a customer in any individual interaction," says Schireson. "There is an imperative for them to develop new products that target users specifically. Well, it turns out that the most time consuming part of that is assembling content behind it. What they've done using our technology is built a repository in XML, which they can then use to develop new products."

Schireson says that Elsevier has built a Web Services layer on top of its content repository, and new products are essentially an HTML application that talks to the Web Services layer. This gives the company a lot of control over the search, what results will be returned, and how these results are going to be presented.

By building search as a Web Service in this fashion, search becomes a platform on top of which companies can build HTML applications that provide more concrete and specific ways to get at the data. According to Gilbane, "By using the search engine as a platform, users have a much more lightweight and feasible way to get their hands around large amounts of information. The advantage Mark Logic has," Gilbane continues, "is its XML structure and granularity. So if you are building an application on top that is going to use Mark Logic as a service to feed it, then you can build some sophisticated metadata into the database in advance."

### SEARCH ENGINES CAN PLAY TOO

It's not just vendors like Siderean, Mark Logic, and Ipedo that are using XML to enhance the search experience. With increasing frequency, Web search vendors such as Yahoo! and Alexa are building Web Services interfaces to content and inviting programmers to build corresponding applications. In fact, Allen of Siderean says that whether you are talking about his Seamark product or Alexa exposing its tools to developers, it's all about accessing needed information and finding ways to use that data beyond the initial results. "It's this aspect of treating results as a resource . . . and eliminating the overhead that people had to wrestle to get that search information into a useful form." He says that

# New England Journal of Medicine Uses Mark Logic Technology to Improve Search

Kent Anderson, executive director of international business and product development at the *New England Journal of Medicine*, wanted to make better use of the Massachusetts General Hospital (MGH) case records. This repository was used by students and instructors to provide sample cases for teaching purposes.

Users were not able to pinpoint the cases they wanted. What they had in their case repository, according to Anderson, "had been popular for decades, but there was no way to mark up and extract the really salient points, the presenting complaints, what somebody shows up with, and then the final diagnosis. Now we can do that."

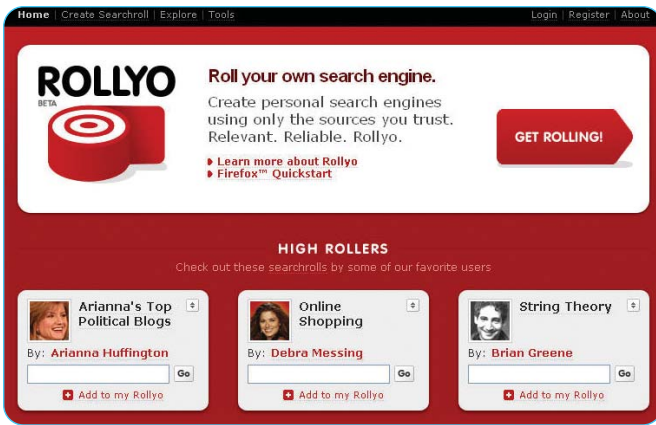
Anderson says he had a discussion with Mark Logic and was impressed with what they had to offer. "We had been looking around, but Mark Logic came up very early in our examination and seemed to have the right attitude and technology." When the *Journal* implemented the Mark Logic solution, they added a tab in their search tool to search cases, and this enabled students and instructors to isolate and retrieve cases in the MGH repository based on different data types such as the diagnosis, the tests given, and other specific information. Users can quickly identify the type of information using an icon system on the search results page.

Anderson says that before they implemented this new system, they would just get a list of cases without any context or understanding of which tests were given or how they arrived at the diagnosis. With the old system, users could enter a complaint or a diagnosis such as heart disease, but just got a list of results in which the term was mentioned. In the new system, he says, users can isolate results around specific issues such as the type of test, final diagnosis, or patient complaints.

He says end users love it because they can pinpoint the cases that match their needs with far greater ease and accuracy than was possible under the old system. "Physicians-in-training have found it remarkable."

The screenshot shows the search results interface for The New England Journal of Medicine. At the top, there is a navigation bar with links for HOME, SEARCH, CURRENT ISSUE, PAST ISSUES, COLLECTIONS, and HELP. Below this is a 'SEARCH RESULTS' section with tabs for TITLE/FULL TEXT, AUTHORS, IMAGES, CASES, CME, and MEDLINE. The 'CASES' tab is selected. Search criteria are displayed: Search Term(s): heart disease, Search Within: Full Text, From: Jan 2000 through: Mar 2006. There are 1-20 of 144 results. A legend below the search criteria categorizes medical tests into Hematology, Chemistry, Other/mixed, General, Radiology, Histology/pathology, and Other. The search results are presented in a table with columns for Case, Labs, Radiology, Histopath, and Other. The first result is a case titled 'A 71-Year-Old Woman with Urinary Incontinence and a Mass in the Bladder' from N Engl J Med 2006;354:850-856, February 23, 2006. The presenting complaint is: 'A 71-year-old woman sought a second opinion on the management of a malignant tumor of the urinary bladder.' The 'Radiology' column contains a 'CT' icon, and the 'Histopath' column contains a microscope icon.

Using Mark Logic Server, the New England Journal of Medicine can isolate given cases based on any number of criteria such as diagnosis or test type. Users can identify specific information types based on the icons.



Using Yahoo! Web Services, this company has designed a tool for creating your own mini search engine with a set of "trusted" results. On the Rollyo home page, shown here, there are several examples of celebrity searches.

"things like the Alexa engine that was put out, and the A9 initiative to extend RSS as a carrier for search information . . . make it simple for a developer to grab and augment a given application, where the query is a way to filter down to a broader set of things, something that a developer is trying to build into a workflow or a decision-making process." And, he concludes, "the closer we get to delivering results easily in workable forms, the closer we get to this notion of search as a Web Service."

Jeremy Zawondy, technical lead at Yahoo!, says although it made its APIs available for some time, the company did so only through specialized business relationships involving search syndication. About a year ago, Yahoo! began offering access to search features as a Web Service as a way to give developers an idea of what kinds of features were available from Yahoo!. He says, "We didn't have a way to let the general population of Web developers or smaller emerging companies plug into what we're doing and make it available to a broader set of end users and developers." The Yahoo! Developer Network, according to Zawondy, gives developers access to Yahoo! offerings such as its Web Search and Photo Search. They also

have a term-extraction Web Service that Zawondy says developers have been using as a way to tag content on the Web.

According to Zawondy, developers use these services in many different ways. For instance, a company called Rollyo helps users build a small customized vertical search engine. "The idea is, I can

go to Rollyo and build a list of trusted sources, then provide a customized search box and put it on my Web site, and anyone who visits my Web site can conduct a search across those sources. In effect," he says, "they are tapping into my knowledge and the sources that I trust in order to get the information for whatever topics they are looking for.

"There's a whole other class of applications that fall into technology demonstrations, show-off, or mash-up applications where people are taking our search interface and building fun things or new visualization tools, things that aren't standalone products but are still demonstrations of where things could be headed if we decide to go one direction or another with next-generation products," Zawondy continues. "One of those I've seen, someone took some RSS feeds from Yahoo! News and used our search interface to Yahoo! Image Search, and provided a new way to navigate news stories on Yahoo! by building a navigation scheme where you navigate using pictures."

Alexa, a service owned by Amazon that provides Web statistics data opened up some of this data as a Web Service about a year ago under the name Alexa Web Information Service. Among the services Alexa offers, according to Niall O'Driscoll, VP of engineer-

ing at Alexa Internet, is a service that provides a way to tag a collection of content and then, based on this, build a set of trusted search results (much like what Zawondy describes for Yahoo!). O'Driscoll says one developer in Germany is running a music site using the Alexa Web Service that enables visitors to search for music by melody. The Alexa search gave the developer access to midi files and he was able to extract this information and build a database of music files and present a unique melody-based search engine.

#### GOOD OR JUST GOOD ENOUGH?

Gilbane points out that no matter how good a tool may be, there are always going to be search repositories that, for whatever reason, lack good metadata. At that point, he says, you become dependent on full-text search to go that last mile. Gilbane adds that there may be some applications where this type of searching won't be good enough to get you what you need. "In healthcare applications, for example, in a medical situation or emergency where you are looking for a particular medication or drug you need quickly, you can't depend on full-text search, you need something more rigid. If you were doing research or writing a paper, it wouldn't be as critical, but if it is a life-threatening situation, you need to be absolutely sure."

No matter how good the conversion tool is, according to Gilbane, conversion doesn't always parse the content as well as it should. Thus, search tools won't be able to make it granular enough, regardless of XML conversion.

Certainly, companies can construct XML data stores where none exists, or build an XML-based cache of critical data. In many cases, though, creating data stores from whole cloth will not be as feasible as converting existing ones to XML. However, even without the most perfect conversion, Web Services-enabled searches give enterprise users the ability to take data results and better incorporate them into the workflow cycle and real business applications. This not only provides a methodology for producing a more useful result set, but a mechanism to better incorporate found information into producing meaningful results—and that is something all businesses are searching for. **CC**

**RON MILLER** (RONSMILLER@RONSMILLER.COM) IS A FREELANCE TECHNOLOGY WRITER BASED IN MASSACHUSETTS.  
**COMMENTS?** EMAIL LETTERS TO THE EDITOR TO ECLETTERS@INFOTODAY.COM.



Mark Logic Corporation  
2000 Alameda de las Pulgas  
Suite 100  
San Mateo, CA 94403  
650.655.2300  
[www.marklogic.com](http://www.marklogic.com)

Mark Logic Corporation is the provider of the industry's leading XML content server. Mark Logic works with providers of information products to accelerate new product creation, deliver products through multiple channels, integrate content from different sources, repurpose content into multiple products, build custom publishing systems, and mine content to find previously undiscovered information. MarkLogic Server does this by enabling companies to load, query, manipulate, and render XML content using the W3C-standard XQuery language. Designed for high performance and scalability, MarkLogic Server can deliver millisecond response times against multi-terabyte contentbases. Mark Logic is privately held and backed by Sequoia Capital and Lehman Brothers.

For more information, please visit [www.marklogic.com](http://www.marklogic.com).