

# information insider

## Next-Generation Searching: Looking for the Right Stuff

Robert J. Boeri  
Martin Hensel

EMedia Professional, July 1999  
[Copyright © Online Inc.](#)

You can't be too rich, too slim, or have too much storage. And the more storage you have, the more information you save, and the harder it is to find it. While the Web has made storage virtually infinite, and both magnetic and optical storage costs continue their dizzy downward spiral, the resulting document deluge has complicated searching inordinately. Meanwhile, content-based retrieval has done little to change with the times, and vendors have been slow to adopt new technologies and standards, making it difficult to find a comprehensive solution.

Remember the 1960s Suzanne Pleshette movie, *If It's Tuesday, This Must Be Belgium?* The same odd logic must be applied to distinguishing search tools and how they apply the Boolean logic that's fundamental to searching. For example, if it's Verity, I use "AND" between the two search terms, but in AltaVista I use a "+" prefix before each term.

Pity the search system suppliers. The root causes of their difficulties (and ours) are the mysteries of language and the continuing exponential growth in documents--both in number and types. But if the search vendors can't agree on how to express "AND," should we expect relief from any more sophisticated search problems? The sheer volume of information to search, and the inadequacy of natural language and Boolean search techniques, are forcing some improvements. As elsewhere, the trendsetters are often on the Web.

### engines that can and how they do it

Yahoo!'s divide-and-conquer strategy, using large numbers of trained librarians to categorize documents, does make searching its content easier and more predictable. However, that predictability means serious restrictions on what it will accept for searching, and 80s-style labor costs to catalog documents. Some vendors are developing products with an 80-20 approach to automate categorization of documents. Infoseek runs its portal with its own Ultraseek server, which organizes content into browsable topics for easy searching.

Other vendors are also seeking the ultimate in catalog automation: Verity, Plumtree, and Autonomy come to mind. Verity's Knowledge Organizer develops enterprise-wide classification pages from which users can access wide ranges of database, document, spreadsheet, Web page, and other information types. Other vendors hope corporations will replace their House-that-Jack-built intranets with corporate portals--something akin to in-house Yahoo!'s, making it easier for employees to find what they need and surf only in-house. Plumtree's system creates document taxonomies by analyzing document titles, content, and other available attributes, such as summaries or metatags. Plumtree integrates access to document repositories, Web information, and data warehouses--all with what can be described as a dynamic card catalog. Autonomy claims its product enables companies to reduce their staffing for cross-referencing, hyperlinking, and categorizing information.

### speaking in tongues

And what about so-called "natural language?" So far this seems like just another perennial promise with only marginal and unpredictable support. Even on sites claiming to provide some natural language support, the results of "Bob's simple test" are unacceptable. This test consists of two queries, *to be or not to be*, both with and without quotations. The use of quotations avoids the interpretation of "or" and "not" as Boolean qualifiers. Without quotations, this simple, but tricky, test could yield a count of every single document on the server since every document would either contain "to be" or it would not. Results? Infoseek yielded a hit list of 57.4 million documents (both with and without quotes). Likewise, AltaVista returned a surprisingly small 9,876 hits. Excite gave 2,462 hits for the query sans quotes, and none for the quoted string, even though we know it indexes Shakespearean content.

Can XML help with the search problem? Since it adds structure to text, XML is a natural for more focused searching. XML-aware search systems could enable searching for terms or concepts within elements such as figure captions. However, because it lets you define whatever elements you wish, it forces systems to have multiple XML document models, collections of elements, and attributes. So XML search systems are probably best for sites that can restrict which models to search.

Infoseek's Ultraseek lists XML support as one of its features, yet forestalls field searches on XML tags for a future release. Autonomy's Agentware Knowledge Server claims an automatic XML tagging module. Oracle's ConText (now called "InterMedia" in Oracle 8i) has added XML support to database searching. Given Oracle's impressive support for natural language, it's no surprise that Oracle now offers structured document and database searching together. Enigma Corporation, an XML electronic book solution provider and Adobe partner, has developed an XML layer for Acrobat PDF searches even though Acrobat doesn't ship with this capability.

## getting meta all the time

Want to gain a strategic edge so the information you publish will be found before that of your competition? Prepare for ways to add metadata to whatever you publish, as you develop the information itself. Capturing this information today could give you a strategic edge tomorrow.

**Robert J. Boeri** ([bboeri@world.std.com](mailto:bboeri@world.std.com)) and **Martin Hensel** ([mhensel@hensel.com](mailto:mhensel@hensel.com)) are columnists for *INFORMATION INSIDER*. Boeri is Information Systems Publishing Consultant at Factory Mutual Engineering of Norwood, Massachusetts. Hensel is president of Texterity Inc., a Newton, Massachusetts-based consulting firm that builds SGML-based editorial and production systems for publishers, corporations, interactive services, and composers.

Comments? Email us at [letters@onlineinc.com](mailto:letters@onlineinc.com).

[\[EMedia Home\]](#) [\[Subscriptions\]](#)

Copyright © 1999, [Online Inc.](#) All rights reserved.  
[info@onlineinc.com](mailto:info@onlineinc.com)