
[Return to article page](#)

To print: Select File and then Print from your browser's menu.

This story was printed from FindArticles.com, located at <http://www.findarticles.com>.

EMedia Professional

June, 1998

Intranet searching a light at the end of the tunnel.(search systems for the Internet)

Author/s: Robert J. Boeri

Since the earliest days of the Internet, information technology professionals have struggled to devise a scaled-down, in-house version for their corporations. The fruit of their labor is the intranet, a smaller network for transporting information within and among corporations that promised the same hyperlinked cohesion that made the Internet such a compelling distribution tool. With the evolution of the intranet, of course, came a need to organize and access the information contained within it and to live up to the expectations placed upon it. And with those needs came the emergence of intranet search systems.

A quick glance at the market reveals a variety of choices. Some run on the platforms through which a corporation's Web is hosted, thus providing a virtual out-of-the-box indexing of the system's HTML files. In many cases, users can begin searching the entire collection in short order, perhaps as they do with a commercial Internet search engine. What more, then, could anyone want? Isn't this problem like buying a new car--more a matter of taste and affordability than a subject of lengthy analysis?

More often than not, a corporation's intranet is maintained not by a single individual, but by a team comprising varied interests, points of view, and responsibilities. Moreover, Web technologies continue to grow exponentially. Every month, in fact, brings nearly a year's worth of changes. And corporations often have already selected search systems or, increasingly, want to be able to search beyond their own intranet.

The sheer quantity of information to be searched and the range of users' needs compound the problem of selecting and configuring an intranet search system. For instance, will colleagues be happy when a routine search behaves as it does on a typical Internet search engine, giving them 20,000 possible items to review in response to a simple question? To understand fully the mechanisms that enable a useful, query-sensitive response, one must first understand the nature of a document and the process by which that document is

categorized and accessed in a search.

YOU HAVE TO START SOMEWHERE: THE EVOLUTION OF THE DOCUMENT

Everybody knows what a document is, yet surprisingly, few can define the term. What's worse, attempting to pinpoint a definition typically leads to more questions than answers. For instance, is a document a single Web page or a collection of them? Are the sound or animation objects part of the document, or separate documents? Despite the countless questions that emerge, it is reasonable to begin with the circular assumption that a document is a book-like collection of related information objects; an information object, in turn, is any meaningful set of data that can be tied to other sets of data in a comprehensive search.

A typical document search involves a scan of its text for designated words or concepts. The search systems that scan for these words include full-text indices with pointers to essentially every word in a collection of documents. Queries using this index can range from simple words or phrases to Boolean AND-OR operations to extended operators like proximity. As the number of indexed documents increases, the need for more sophisticated search techniques increases as well. These search aids often include thesauruses, language support, and even facilities for searching general concepts.

One company that has successfully addressed intranet document management issues is Sunnyvale, California-based Verity. Verity's comprehensive search engine--embedded in many of its competitors' products--offers "topic" queries that can be combined for increasingly rich concept searches. For instance, a basic topic query for "garden" quickly expands to subordinate queries like "vegetable," "flower," and "herb" gardens, which are themselves distinct queries. By building well-designed families of topics, extraordinarily detailed concepts can be searched and found in large document collections.

Although the most talked-about stage of document development is the search--and-use phase, effective intranets must also carefully consider how they will receive documents, what binary types of documents will be accessible, how long documents will remain available to users, and what becomes of them after their expiration date has passed.

NOT ALL DOCUMENTS ARE CREATED WHERE AND HOW TO SEARCH

In many companies, departmental or divisional HTML pages are constructed consideration for how the content be searched. The home page's usually provides access to content frequently subdivided by corporate organization or function. When the amount of content is relatively small, searching is not difficult. A point-and-click approach, in these cases, is all that's needed.

When the amount of content grows, however, searching by navigation alone becomes far less simplistic. New categories of

information are set up, and if these mirror the firm's organization, the information layout changes with the corporation. Add to that difficulty the reality that different groups may begin growing their own subnets, or have preferences for their own document management and search needs, and you have either a disaster brewing or a great opportunity, depending on the color of the lenses through which you view the world.

If paper documents are part of the content to be searched, they are typically digitized using an Optical Character Recognition (OCR) system. OCR renders text searchable and creates files that are always smaller than their image counterparts. Unfortunately, OCR systems do not preserve word processor structures, but throw away font and layout information, as well as pictures and graphics. Systems based on Adobe Acrobat Capture, however, will create Portable Document Format (PDF) renditions that are not only searchable, but preserve many of these elements, including graphics.

The information objects inherent to documents almost always have attributes that simplify any given search. PDF documents, for one, have built-in attributes that can assist in searching; new attributes can also be created, if appropriate. Even word processor files have attributes--saved as "summary" or "cover" page information--that generally include the author's name and the files' subject matter. With these document attributes, users can divide and conquer portions of their document base and then apply full-text queries to the remaining database of information.

Equally important, though often forgotten, is the host language in which the documents are written. While simple 8-bit ASCII is common to English-read pages, it may not be accessible in other parts of the world. And while HTML tags are written in English, what lies between them may not be. UNICODE, for instance, supports non-English languages ranging from the common European FIGS (French, Italian, German, and Spanish) to kanji. Likewise, Acrobat PDF files can express non-English text. Particularly useful to the development of a search system strategy is an upcoming standard called the eXtensible Markup Language (XML), which is designed to support foreign languages and add more structure to electronic documents.

AN INTRANET AND ITS SEARCH SYSTEM: MAKING IT WORK FOR YOU

Further complicating the quest for an effective search system is the purpose of the corporate intranet itself. Is it simply a communications vehicle, or is it a virtual workplace where employees work and share files? If it is a virtual environment, users will want to extend their boundaries and search content within and beyond their company's intranet. Likewise, they will demand a unified process of searching. For corporations that have already settled on a separate search system for accessing information outside their own intranets, the process of integrating these search systems and techniques creates yet another obstacle.

Given the diversity of factors that must be considered, it is not surprising that the search for an intranet search system can seem daunting. The investigation becomes more manageable, however, through needs analysis. Specifically, the user must consider a series of questions about the types of documents to be examined, the characteristics of the people who will be using the search system, and the required capabilities of the search system itself to identify those products that are most likely to take full advantage of an intranet's information archives. For example, the IT professional responsible for initiating a mechanism for intranet searches must ask what types of documents will be indexed and searched, whether paper legacy documents will be searched, and whether documents will be searched by predefined attributes.

Also critical to the search for an appropriate intranet searching mechanism is an understanding of the user and his or her needs. If the average user does not meet certain qualifications, or if funds are some searching mechanisms will better than others.

A final consideration centers on the that will be placed on the search being investigated. Questions of importance include the platform upon which the system will run the types of operations and level of customization desired. When reviewing the possibilities, it is wise to have a test suite of documents representing the kinds your organization uses readily available for indexing and searching.

MATCHING THE SYSTEM TO THE NEED

Every intranet is different, just as almost every living organism is different. However, some systems stand out for their ability to minimize clutter and streamline the search process. While Web search agents within programs are common, few are fully capable of delivering only what you want and minimizing the clutter. To assist agents, document collections must themselves be sifted into rational categories which can be automated.

One system that successfully improves the value of Web search agents and automates category building is Information Access Systems' Judgement Space, or J-spaces[R]. Originating from the U.S. Air Force's Artificial Intelligence Center, J-space has been commercially available through integrated products for more than ten years.

As companies grow, their document collections tend to evolve into islands of information searched and managed by incompatible systems. To solve this problem, Infodata Systems, of Fairfax, Virginia, has developed a product called Virtual File Cabinet (VFC). This customizable, Web-based system allows users to access, organize, and share documents. With VFC, users can search, retrieve, edit, and file information throughout the enterprise, regardless of where the documents originated or are stored, by navigating a hierarchy of collections that uses an intuitively obvious metaphor.

But all that power may go for naught if a system is too difficult to use. Several years ago, a Massachusetts-based commercial property insurance company implemented a powerful industry-leading search system. It was second-to-none in its power and customized to provide search features appropriate to the business' needs. Unfortunately, it was not easy to use, and the system never met the company's usage goals. Lesson learned: If a system flunks the useability test, or does not meet a user's needs, it will either be underused or replaced.

The best systems not only include useful search aids, but will chunk document collections into Yahoo-like categories, thus reducing the number of responses. Furthermore, systems should ably rank results by relevancy and provide summaries of results, clusters of results, and the option to "find me more like this." While doing all this, the systems should also provide an automated setup of categories for searching, and search constantly growing, heterogeneous groups and types of information, including nontextual media and structured database information.

XTENDING HTML: WHERE SEARCHING IS HEADED

Given the need for continuously improved search systems, there are several key areas to watch. For example, anyone considering implementing or using intranet searching systems must pay close attention to XML, which will likely facilitate more tailored searching.

Metaphorically, HTML can be compared to Henry Ford's Model-T: It made automobiles available to the masses, was simple and affordable, and you could get it in any color you wanted as long as that color was black. SGML--HTML's parent standard--is like a Mercedes-Benz, with a full range of mix-and-match options. Specifically, it is built for the long haul, but quite expensive. XML, like the majority of vehicles on the market, is fully customizable and affordable, but lacks some of SGML's capabilities. Although the XML specification is only about one-tenth the size of the SGML specification, it is still remarkably powerful; thus, the mad rush to make everything, including browsers, publishing systems, and search engines, support XML.

The biggest benefit of XML to search systems will be its ability to perform zoned searching, or full-text searching, within custom document elements. Likewise, XML supports ISO 10646 (the UNICODE standard), enabling support for international languages and the use of those languages in XML tags. Though no search engine or document management system currently supports XML, it is a sure bet that many will be pledging support within the next year.

Yet another aspect of XML that deserves consideration by potential users is membership in the World Wide Web Consortium. By accessing its Web site, at <http://www.w3.org/Consortium/Member/List>, one may determine whether a particular Search System Analysis Checklist vendor is committed to emerging Web standards, including XML. Microsoft, Netscape, and Digital Equipment, for instance, are all members of

the World Wide Web Consortium, yet only one company's product-- Microsoft's Internet Explorer 4.0--had sufficiently committed to XML in early 1998.

Of course, developing and marketing search systems has not proven to be the highly profitable endeavor originally anticipated. Vendors who seemed to hold solid positions in the top tier of vendor comparison lists have struggled recently. Verity, for one, is not the economic powerhouse it used to be. Fulcrum, once a top-tier vendor with references including Microsoft, was acquired by PC DOCS, developers of the document management system of the same name. And new search system vendors like Oracle are emerging from other disciplines, too.

Given these trends, and the evolution of the intranet into a viable repository for important data, the need for effective searching mechanisms seems more important than ever. The demand for intranet searches--which, at first, seemed to be a simple problem--is rapidly becoming a noticeably complex undertaking. Luckily, search technologies and vendor offerings are maturing with no end--but yet, a light--in sight.

RELATED ARTICLE: Document Analysis Checklist

Type	Response
Exactly what kinds of documents, and what versions of each kind, will you want to index and March?	HTML version 4.0 is being considered by the International Standards Organization; word processor formats abound. Leave yourself the option of future revisions by selecting a product likely to support future versions.
Will you want to search what are currently paper legacy documents on your intranet?	Be sure the OCR package your firm uses supports the binary formats you plan to search. One good choice: Acrobat Capture and PDF, which preserve the look-and-feel of documents, and place graphics in context.
Is it important to search documents by their attributes?	If so, choose a product that supports attribute searches and allows you to add your own to documents. TIP: Adding custom attributes will increase the expense of preparing documents for indexing.

RELATED ARTICLE: User Analysis Checklist

Characteristic	Response
Low willingness to learn details of search features	No need to implement complex features if they won't be used. Make technical or special librarian services available to coach or actually perform searches.

	Possibly develop templates or reusable queries to facilitate searches.
	Consider system with natural language queries or ability to "find me more like this."
Technically curious; willing to perform power searches	Select system with rich search features, including ability to model concepts. Preserve ability to save and re-use queries. Combine content with attributes searching. Consider developing customized attributes for each document to enhance search precision.
Pro-active or passive users?	If passive, consider facility to "push" selective information to the desktop. Give users the option to turn off or alter the flow of information if they want to. If pro-active, emphasize ability to get information on demand by formulating queries.
Requires complex, specialized, or technical vocabulary	Develop specialized dictionaries or thesauruses to aid in searching. Case-sensitive searches needed? (May increase index size and slow down searches.) Can you define stop words to reduce index size, speed indexing, and speed searches? Is it possible to predict which words users must be able to find?
Demands collaborative culture	Develop facility for sharing and combining search queries.
Requires single native language or multilingual?	Be certain system supports required languages. Understand lexical requirements of searches. That is, will special pie-characters (such as [R] and [TM]) be part of searches or actually interfere with searches? Defining certain characters to avoid in searches will make searching easier and more efficient.
Heterogeneous groups of users?	If many different vocabularies and user profiles are present, consider offering both simplified searching and power search facilities. May also reduce ability to customize thesauruses or automated search assists.

RELATED ARTICLE: Search System Analysis Checklist

Capability	Response
Match your document and user requirements with the search system features.	This is so obvious, it is often overlooked in the glow of glossy vendor literature and demos.

TIP: Do the vendors on your short list have plans that assure their systems will continue to evolve?

On which platform will the search system run?

Be sure IT provides operational support to the search system you have in mind.

Documents continuously changing?

Select system with continuous "Web crawling" and index updates.

Do you want to customize your search system?

See if the system allows you to customize thesauruses, stop lists, and lexical character sets.

Be sure your system can index and allow searching of document attributes (such as HTML information and word processor attributes). Will the system allow you to add new, custom attributes to documents you will index and search?

TIP: Technical librarians can help develop appropriate customizations, delivering more benefits to the effort.

Do you need document management functions besides searching (e.g., link management)?

Consider a Web-aware document management system that also provides searching (e.g., Documentum's RightSite)

RELATED ARTICLE: Companies Mentioned In This Article

Adobe Systems, Inc. 345 Park Avenue, San Jose, CA 95110-2704; 408/536-6000; Fax 408/536-6799; <http://www.adobe.com>; InfoLink #401

Digital Equipment Corporation 30 Porter Road, Littleton, MA 01460; 800/336-7890; Fax 508/486-2017; <http://www.altavista.software.digital.com>; InfoLink #414

Documentum, Inc. 5671 Gibraltar Drive, Pleasanton, CA 94588; 510/463-6800; Fax 510/463-6850; <http://www.documentum.com>; InfoLink #415

Fulcrum Technologies, Inc. 785 Carling Avenue, Ottawa, Ontario, Canada K1S 5H4; 613/238-6452; Fax 613/238-7695; <http://www.fulcrum.com>; InfoLink #419

Infodata Systems, Inc. 12150 Monument Drive, Suite 400, Fairfax, VA 22033-4058; 703/934-5205; Fax 703/934-7154; <http://www.infodata.com>; INFOLINK #420

Information Access Systems, Inc. 3085 Bluff Street, Boulder, CO 80301; 303/442-6224; Fax 303/442-4530; <http://www.J-Space.com>; InfoLink #421

Microsoft Corporation One Microsoft Way, Redmond, WA 98052-6399; 425/882-8080; Fax 425/883-8101;

<http://www.microsoft.com>; InfoLink #430

Netscape Communications Corporation 501 East Middlefield Road,
Mountain View CA 94043; 650/528-2555; Fax 650/937-21;2;
<http://www.netscape.com>; InfoLink #431

Oracle Corporation 500 Oracle Parkway, Box 94065, Redwood
Shores, CA 94065; 660/506-7000; Fax 650/506-7200;
<http://www.oracle.com>; InfoLink #433

PC DOCS, Inc. 25 Burlington Mall Road, Burlington, MA 01803;
781/273-3800; Fax 781/272-3693; <http://www.pcdocs.com>;
InfoLink #434

Verity, Inc. 894 Ross Drive, Sunnyvale, CA 94089; 800/935-6246,
408/541-1500; Fax 408/542-2010; <http://www.verity.com>; InfoLink
#444

Robert J. Boed (bboeri@worldstdcom), co-columnist for Information
Insider, is an Information Systems Publishing Consultant at Factory
Mutual Engineering in Norwood, Massachusetts.

Comments? Email us at letters@onlineinc.com, or check the
masthead for other ways to contact us.

COPYRIGHT 1998 Online, Inc.
in association with The Gale Group and LookSmart. COPYRIGHT 2000
Gale Group
