20 Questions (With Answers)

Enterprise search represents a new opportunity for information professionals. Long accustomed to searching traditional online services and Web databases for external information, we are now turning our attention to internal search as our organizations gear up to manage their own information resources. Under the general rubric of "content management," the issue of an enterprise search system inevitably rears its head. If the information/library unit has positioned itself as the "go to" place for search help, then those charged with implementing internal content management and search will show up requiring advice and counsel. On a more proactive note, we should be seeking out those within our organizations charged with this task and telling them how we can make their job easier.

Personally, I'm fond of an analogy IDC used in "A 360-Degree View of Enterprise Information," a white paper written for FAST [www.fastsearch.com/us/esp]. "Today's business workers resemble yesterday's riverboat pilots. They are constantly in danger of running aground while navigating the shoals and backwaters of enterprise information flows. Putting a next-generation search system front and center as the key enabler for a business worker holds out the very clear potential to transform the riverboat pilot into a top-gun fighter pilot."

An excellent place to ground yourself in content management topics, issues, and terminology is Tony Byrne's Content Watch Web site [www.contentwatch.com]. Stephen Arnold, a frequent contributor to ONLINE, has been working with Tony in a study of enterprise search. Some of his findings were presented at the Enterprise Search Summit, held in May 2004 in New York City. I've asked Steve to share some of his thoughts on how to choose an enterprise search system so that information professionals can quickly get up to "fighter pilot" speed. This is only a portion of a much fuller report to be published by Content Watch.

–Marydee Ojala, Editor

This article consists of some of the questions most frequently asked during the course of an enterprise search system implementation; the ensuing quick answers are the result of a 6-month study of enterprise search. For many questions, the answer is not complete. What many professionals believe is a "slam dunk" is more correctly a different ball game. An effort has been made to provide guidance on some complex issues. The questions and answers are intended to provide some broad compass headings for a procurement team working on an enterprise search adoption.

1 What search system should we buy?

Don't buy any search system, at least not before examining some other approaches. Use the search tools available within the present applications. Test them. Determine their limitations. Then download demonstration copies or obtain demonstration software from several vendors. Get hands-on experience with "free" systems before embarking on a formal acquisition process. If a procurement team does not have time for this step, do not embark on an enterprise search project. The risk and cost are too

About Enterprise Search

By Stephen E. Arnold

great for today's business environment. Knowledge first. Financial resources second. Formal procurement process third.

2 What search-related costs are the most difficult to control and why? What types of expenses are we looking at for hardware, programming, network services, third-party add-ons, etc.?

The complexity of enterprise search creates a situation in which cost control is often difficult. Among the most difficult to control costs are those associated with human labor—programming, manual fixes to an index or classification schema, and troubleshooting. Two costs that usually blindside even sophisticated organizations are the capital and human investments needed to handle unexpected increases in the number of documents processed and the number of simultaneous users of the system. A third type of cost spike occurs when business processes must be analyzed and the search system used to deliver information to support an employee who is providing customer support.

The process of analyzing a work flow, crafting a saved search that fires automatically when an event occurs in the employee's work process, and delivering that information to the employee at the precise instant the data are needed is expensive.

When a security breach occurs, the cost associated with finding and remediating the security issue can be impossible to control, and most organizations spend big bucks to keep their information secure from attack, theft, or corruption. The final cost that is difficult—and sometimes almost impossible—to control is an expense associated with a particular feature. When an organization wants to implement a feature such as on-the-fly translation of documents from a source language to the user's native language, the process can be human-intensive and resource-intensive. What seems like a simple request in today's hyperfast computer age is often easy to demonstrate but fiendishly difficult to implement so that the service works quickly, accurately, and intuitively.

3 In what parts of the overall search system is human expertise typically most essential?

There is a need for human expertise at all points in the search system procurement, deployment, and operation. In procurement, good judgment is needed to determine what the search system is supposed to do. Many search systems fail to meet expectations because the expectations are unrealistic at the outset of the project. The deployment of the search system requires managerial, financial, and technical expertise of the highest order. The most trivial decisions, such as the type of hardware to use to host the search system, can have enormous financial consequences throughout the lifetime of the project.

In operations, humans are needed to determine when and how to intervene in automated systems. Search systems are still unable to perform some of the interpretive tasks that humans take for granted. Language is a very difficult and challenging area for most computer systems. Although progress is being made, the idea that a search system can operate without human intervention in document selection, indexing, security verification, relevancy, and dozens of other search functions is not true. A failed search system is almost always a result of poor human work, not poor search software work.

4 How do we search our Web content without going to

➤

the time and expense of licensing search software and running it inside our own company?

There are three different approaches to offering enterprise search without having the search system located on site at an enterprise. The first is to use a third-party service such as Atomz or Blossom. Both companies can index enterprise content in standard office file types from their servers. Both provide the code snippet to place on the portal page for the intranet so employees can query the index. The results list points to documents that reside on company services.

A second approach is to pay a search system vendor such as FAST, Verity, Convera, or Autonomy to arrange for remote indexing. However, the cost of this service is generally higher than fees associated with such search providers as Blossom and Atomz.

A third option is to retain a managed services company such as AT&T, MCI WorldCom, Globix, or other similar company. The enterprise enters into a contract with a managed services company to license the search system, set up the infrastructure, and operate the search service. These approaches are similar in that the burden of work shifts from the enterprise to the contractor. The variables among these companies include security, technical support, and cost.

5 We have heard many seemingly similar descriptions of search software from vendors of search, customer relationship management systems, application server software, and content management system software. Are these companies selling the same basic capability? Do unique offerings exist?

Differentiating among search vendors on the basis of a written description is very difficult. Only head-to-head tests using an identical corpus of sample documents on identical or functionally equivalent hardware will provide the points of differentiation. A marketing person can write a sentence with the terms "natural language processing," "seamless integration with industry standard databases," and "full support for CRM system content." Critical factors such as logic of the administrative interface, document processing throughput (basically, the number of documents the system can index per hour), speed of displaying results in response to a user's query, system stability (the amount of time the system is up and running and not down for maintenance), the vendor's responsiveness to a glitch, and similar factors are more important than the buzzwords.

Remember that search system vendors explain their system to individuals who are generally unfamiliar with the internal processes required to take a Word document, index it, update the index, take a user query, match that query against the index, and display the results in a way that make sense to the user. As a result, the marketing literature and most vendor dog-and-pony shows never reveal the essential facts about the search system. Vendors such as Open Text, Hummingbird, Endeca, iPhrase, and InQuira, among others, avoid describing their search systems as search systems. The reason is that the language used to describe search and retrieval has been devalued to such a degree that a different vocabulary allows the companies to talk about a different value proposition. Examples include iPhrase's work flow and ROI positioning; Open Text's collaboration suite, and Endeca's guided navigation approach. Search is one of the great challenges in computing science, and it is also one of the challenges that is difficult to explain.

> A good rule of thumb is to increase storage capacity for a search system by 50 percent each year.

6 What are the essential differences among Autonomy, FAST, Convera, Verity, Hummingbird (Fulcrum), and Open Text search-and-retrieval products?

These are companies that have built above-average brand visibility. The differences among them are massive. Autonomy focuses on processes that are automated through statistical analyses of content. FAST is a system built for speed that has been engineered to deliver a highly flexible and configurable search system for enterprise and Web content. Convera, like Verity and Hummingbird, offers modular systems that can be configured to handle most search requirements. Convera, Verity, and Hummingbird, unlike FAST, represent architectures that have been refined for 6 to 10 years and deployed in hundreds of client environments. Open Text is a firm that offers several different search engines. These range from ones optimized for collaborative environments and electronic mail to traditional Boolean engines that deliver incredible power and precision but require the user to enter command line queries. From the point of view of a procurement team, any one of these companies is a "safe" buy. These companies represent more than 60 percent of the enterprise search licenses. The companies are publicly traded, have stellar customer lists, and offer a mind-boggling range of features and options.

7 How many search-and-retrieval solutions are commercially available at this time?

Including open source and demonstration systems from university research laboratories in the major countries, there were more than 200 search-

and-retrieval options available as of December 2003.

8 What is the ratio of license fee to professional service charges for an enterprise search system that serves 1,000 users, runs on hardware that has eight processors, and indexes 2 million documents?

Cost estimates vary widely. For mainstream software, the base license fee is in the $50,000 to $200,000 range. Services are likely to add anywhere from two to six times to this base cost. This means that in the first year of the license, the total cost for fees and services can range from about $250,000 to more than $1 million. The actual cost will be higher when internal staff time, hardware, and network infrastructure are included. The first year costs for indexing can be viewed as a cost per document, so 2 million documents works out to $5 to $10 per document, excluding internal staff time and additional purchases.

9 How much does it cost to index manually 100,000 documents? What about using an automated system?

If the cost per document touched is 50 cents for a 4K Word document in English with six terms applied, an author identified, and a data assigned, 100,000 documents cost $50,000. If the documents are patents or scientific papers, the cost can reach $20 per document or $2 million. These costs are the reason that commercial databases are so expensive in comparison with an advertising-subsidized service from the Wall Street Journal. The costs are why Autonomy's argument that the system can work with minimal human intervention has been such a powerful sales tool.

When documents are converted from paper to a digital form such as an Adobe PDF file, it is impractical to have a person index millions of legal documents for a large court case. The fuzziness setting in search systems is actually a way of compensating for poor indexing of these digital copies of paper documents.

If the indexing is narrowed to inserting keywords from a list of controlled vocabulary terms, there are two ways to view the costs. First, a human indexer can apply terms. The per-document indexing cost decreases as the number of documents goes up. The best prices for human indexing are typically obtained from companies sending the indexing work to lower-cost operation centers in the Philippines, China, and India.

If humans are spot-checking the indexes produced by automated indexing systems such as those available from nStein, ClearForest, Inxight, and Stratify, among others, the costs drop to as little as 15 cents per document.

With an automated system, the cost depends on three factors: [a] the cost of the indexing software, [b] the systems needed to operate the indexing software, and [c] the human set-up time. Automatic indexing software can cost as little as $20,000 up to $1 million or more. Set-up fees typically increase this base cost by a factor of two to eight times. Sophisticated systems such as those from Stratify or ClearForest benefit from upfront editorial work. Lower-cost systems focus on providing an "out of the box" solution that accommodates some manual adjustment, but the focus is on the software's ability to make an acceptable indexing decision.

At this time, for most large document sets, human indexing is essentially too costly. However, certain types of information—chemical structures, engineering drawings, documents that contain little text and image data—require manual indexing unless the metadata attached to the information object are available.

Therefore, automated indexing is an essential technology. However, additional quality control is desirable to [a] handle documents with complex or arcane terminology or [b] speed up the response time of a search and retrieval system. As indexes become large, response time of the system slows. Indexing documents with an automated system reduces the number of terms held in the index. Fewer terms generally yield faster system response.

> The search system must be set up to handle, track, and comply with third-party content requirements.

10 What is the best way to calculate the amount of storage we need for our enterprise search-and-retrieval system. We have about 2 million documents to index, and the number of documents is growing at about 50 percent a year.

Enterprise search engines will build an index that is anywhere from 1.5 to 2.5 the size of the indexed documents. If 2 million documents average 50,000 kilobytes each, the storage these documents require would be about 100 gigabytes of data. The index for this document collection—sometimes called a corpus—would run from 150 gigabytes to 250 gigabytes. With a growth rate of 50 percent each year, every 12 months an additional 75 to 125 gigabytes of storage would be required.

Many search engine vendors use compression techniques and proprietary index formats to reduce the size of their indexes. FAST Search & Retrieval is one of the leaders in minimizing index size. Nevertheless, storage is a key consideration. The number of documents and the average size of the documents increase each year. Electronic mail can pose special challenges because a mail message can include an attachment several megabytes in size. An example is a short message containing a PowerPoint deck or a PDF of a 16-page, four-color brochure.

➤

A good rule of thumb is to increase storage capacity for a search system by 50 percent each year. The search architecture should accommodate plug-and-play storage devices. Mirroring, RAID technology, and clusters are storage approaches that a search system requires. Loss of an index means that a new index must be built from ground zero if a current back-up or fail-over mirror is not available.

11 What is the most cost-effective, easy-to-operate way to integrate content produced by our employees with content from an outside Web source? Outside commercial sources such as the Associated Press? A mixture of Web and branded third-party data in text and row-and-column format?

Integration of content requires the same sequence of actions regardless of the search system or the sources of the content:

a. Identify the content to be indexed.
b. Analyze the format of the data.
c. Determine if a format that is directly supported by the document processing subsystem import module filter of the search system can be used to acquire the content. Determine if the search system can automatically identify the file type and acquire it without coding.
d. If no filter is available, a filter must be 1) licensed from a third party, 2) coded by the search system vendor, 3) coded by another third party
e. Test the filter.
f. Determine the flow of content and establish a schedule for acquiring, indexing, and scripting it.
g. Monitor the system so that when an exception or format change causes a record to be rejected, the specific scripts can be adjusted.
h. Repeat the process for every file type to be processed.

Acquiring data from structured tables works the same way. If there are 100 different table structures, the process must be repeated for each table. Some of the newer database table indexing systems integrate with the source tables and provide some automation tools to reduce the manual set-up. Even these modern tools such as those from Speed of Mind and utility vendor Pervasive Software [www.pervasive.com] cannot eliminate the data analysis and set-up processes.

Once the source data are in a format that can be ingested by the document processor, the licensee of the search system can determine if the content should be held in separate collections, which is desirable for third-party content with reuse or copyright restrictions, or placed in a single index with appropriate value-added tagging to permit content to be sliced and diced (sorted by attribute).

12 Our organization has never had an enterprise search. Do most organizations index "from this day forward," or do organizations do retrospective indexing of older material?

Most enterprise search systems look at existing files on servers in active service on the organization's network. Therefore, enterprise search builds an index at the point at which the source content is collected and processed. However, many organizations go through a search system learning process. After the electronic documents have been made available, other content sources are identified. Among these are:

- content residing on proprietary systems such as Lotus Notes or in databases running on AS/400 or HP/UX servers that have not been made "Web aware."
- information in hard copy, such as legal documents or copies of engineering drawings in different sizes and form factors.
- specialized electronic content in an electronic mail system or in a proprietary accounting system running on separate hardware in the company's data center.
- Web information in the formats supported by a standard browser such as PDF, images, or copyrighted and government information.

There is no simple answer to any of these content types in terms of "older material." Content analysis is needed to generate rules and guidelines for

Guidelines for Indexing Content by Time

	Yesterday	Today	Tomorrow
Proprietary or legacy system content	Ignore unless mandated by federal requirements or management directive.	Export in a file format that the search system supports.	Wrap proprietary system so that content can be moved automatically from proprietary or legacy system to enterprise search system.
Hardcopy	Ignore unless required for federal compliance.	Ignore because about 93 percent of information is in electronic form.	Develop a policy for handling hard-copy documents.
Electronic mail	Destroy.	Index only if required by law or a specific company policy.	Monitor policy and shift to "destroy after X days" approach if possible.
Web information	Ignore.	Identify specific URLs and rules for obtaining specific content.	Accept content via RSS feed.

specific types of documents. The table at left provides a snapshot of representative approaches to each of these document types in terms of yesterday, today, and tomorrow categories.

13 We have an enterprise search system from one of the leading vendors. We want to switch to another product. What are the steps? Where are the pitfalls? What are the risks?

Switching search systems is a relatively straightforward process technically. The old system is uninstalled. The new system is installed. The principal challenges are user training and familiarity. The steps are the same as for procuring enterprise software. A thumbnail summary: Identify a product, license it, install it, and deploy it.

Unfortunately, enterprise search systems don't work like a desktop application, so the actual process is more complex and trouble prone. The pitfalls are that the new system may not perform any better than the old system. When a search system fails to deliver, the problem is usually with the set-up and management of the system, not the search system itself. This is a key point when one well-known search system is replaced by another well-known search system. The risks remain the same. A rip-and-replace solution is a last resort, usually indicative of other problems unrelated to indexing and displaying results. In general, little of an existing search system can be repurposed, so rip and replace means starting from ground zero.

14 When can we have a Google-type search system for our internal digital content? How can we best implement Google's Search Appliance to protect content and prevent indexing of restricted material?

Google's search algorithms work best with hyperlinked documents, something internal documents rarely have. Without hyperlinks to assist with relevance, Google's engine performs on a par with other low-end enterprise search systems.

With Google Search Appliance, index entries are deleted using its administrative tool. The appliance can be prevented from indexing certain documents by excluding the folder, file name, or server via the administrative interface. Another appliance, from Thunderstone, is plug-and-play. Both Google and Thunderstone provide graphical user interfaces to common administrative functions.

15 What issues are associated with having our enterprise search operated by a third party under a managed service contract (outsourcing)? How do we protect our proprietary data? What are the costs for this type of enterprise search service?

There is one issue that dominates all others—security. Hosted search raises questions about access to the organization's most valuable possession—its data. The data can be protected using industry standard security guidelines and company specific policies. The federal government makes security guidelines available from NIST, the Department of Defense, and the Office of Management & Budget. Industry groups provide security guidelines. Vendors provide security guidelines. The issue is developing a policy and enforcing it.

Hosted search offers one major advantage in the form of lower overall costs. A hosted solution to search can be deployed for about one-third the cost of a fully burdened in-house enterprise search solution. The catch is fully burdened. Most organizations do not capture every cost associated with enterprise search, so unless such cost data are available, demonstrating a specific savings may be difficult.

16 Our information center has licensed content from The Thomson Corporation and United Press International. Is it really true that a single enterprise search system can address both internal and external content? If we do integrate some of this external data, what happens if we break our agreement with these third-party providers?

From a technical point of view, the entries to the content in the index can be removed. From a legal point of view, the existence of the content itself within the organization varies by contract and third-party vendor. When a new source of third-party content is acquired, that content must go through a data analysis step so that it can be acquired and processed by the document processing subsystem.

Many third-party content providers impose additional requirements on their customers. Among the most common are stipulations for how long the content may be retained and what type of usage tracking data are needed to ensure compliance with copyright and royalty requirements. The search system must be set up to handle, track, and comply with third-party content requirements. In general, third-party providers ignore certain types of laxity in enforcing the contract's provision. However, if a contract or standing order

A good rule of thumb is to increase storage capacity for a search system by 50 percent each year.

is cancelled, the third-party provider may use compliance with the signed agreement as leverage. Copyright transgressions can be embarrassing and expensive to resolve.

17 We have seen some very interesting visualization and graphic search demonstrations. One was from a company called Plumb Design, which offered a visual approach to our thesaurus. Kartoo, a company in Paris, demonstrated a three-dimensional display of search results. Should we consider adding a visual display of search results to our enterprise search to improve its usability and appeal to our users?

At this time, visualization is not a significant part of enterprise search, although it is becoming more evident in special purpose applications of search, such as chemical structure searching. Laundry lists of search results are difficult for the employee to process quickly.

Visualization of result sets is one way to address the laundry list of hits problem. Vivisimo places hits in folders and generates labels for these folders on the fly. Kartoo is a metasearch engine that passes a query against different collections and presents the results in a graphical map. Plumb Design uses a hyperbolic map similar to the one developed by Inxight, a spinoff of Xerox PARC.

For enterprise documents, visualization can aid special-purpose queries against certain domains. However, visualization may add to the computational load at indexing time (more value-added tags are often required). Newer systems perform the processing prior to displaying the result sets to the user. The computational load may be trivial in a demonstration environment. However, for a large-scale deployment, the visualization processes may require a separate subsystem so that performance remains acceptable to the user.

18 Why does Yahoo! offer a "search box" and a point-and-click list of headings? Do users prefer one way to search for information over another? We do not have a classification system for our documents. What is required to create a useful directory or classification system so our users can point and click their way to information without typing a query in a search box?

Yahoo! combines the search box, a Library of Congress type of classification scheme, and stored searches because users demand different modes of access. Yahoo! also offers domain-specific searching. The search engine for the Personals section has very limited features. The search engine for Shopping allows result sets to be sorted by price. About half of Yahoo!'s users use the search box; the other half use a point-and-click approach to content. Enterprise search systems, particularly for large user communities, should take heed. A search box is not a useful tool for about half of any population's users.

> Visualization of result sets is one way to address the laundry list of hits problem.

19 We have heard a great deal about "guided navigation." Is guided navigation search and retrieval, or is it a separate type of search functionality that can be added to our existing system?

The company most closely identified with "guided navigation" is Endeca. The firm's search software was chosen in 2003 to provide a search tool for the U.S. Library of Congress. The core of Endeca's approach is an interface that allows the system administrator or Endeca's professionals to create a sequence of easy-to-use interface screens that "walks" the user through the search process. Endeca has invested considerable effort in its proprietary tools that allow "guided search" screens to be assembled and mapped to indexed data in unstructured (text) format or in row-and-column format. Endeca's "guided search" allows text and row-and-column data to be integrated. Endeca's approach can be emulated with software from Easy Ask, Mercado, and Verity's K2 software.

20 We have a top-notch corporate information center. How can we best utilize the expertise of the professionals in this unit for enterprise search?

The manager of the information center, or an outside consultant with equivalent experience, should be added to the procurement team. Individuals with training in information science should be part of the testing team. Indexing and classification processes and tuning should include individuals with formal training in indexing, library science, and search and retrieval.

Enterprise search offers a wide range of choices. None of the search systems are flawless, and most will be complemented by specialized or nice systems. Careful planning and head-to-head tests are essential in making the correct choice today and for the foreseeable future.

Stephen E. Arnold *[sa@arnoldit.com] is president of Arnold Information Technologies (AIT), based in Harrod's Creek, Kentucky.*

Comments? E-mail letters to the editor to marydee@xmission.com.