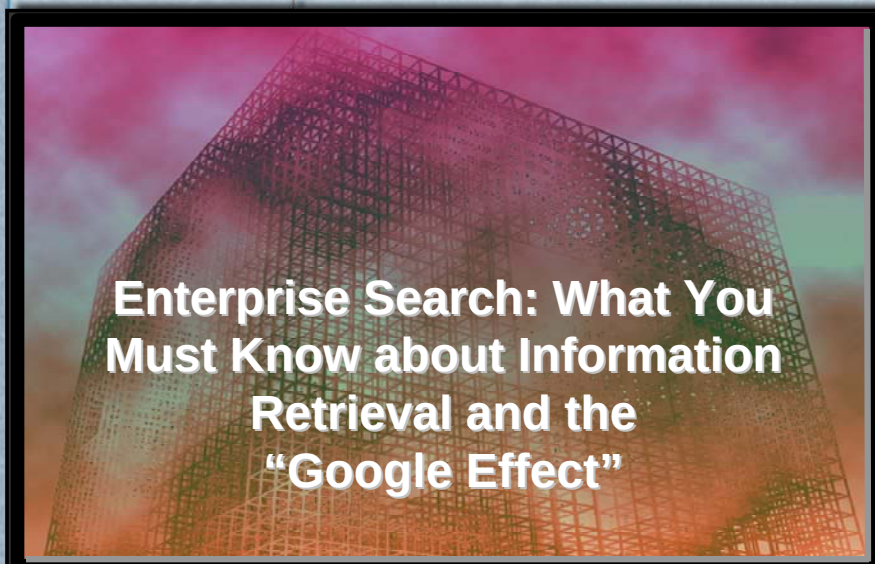


Enterprise Search: What You Must Know about Information Retrieval and the "Google Effect"

© Stephen E. Arnold, Postal Box 320, Harrod's Creek, KY 40027
Email: sa@arnoldit.com – Voice: 502 228 1966

Management consulting and strategy

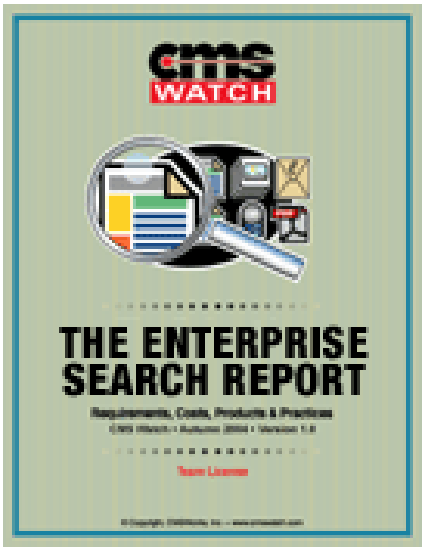
"The Google Legacy" became Available earlier this month Order at www.infonortics.com

"The Enterprise Search Report" 2nd edition now out. Order at www.cmswatch.com

Additional information at www.arnoldit.com/sitemap.html

Contact: sa@arnoldit.com

Limited Access Document



2
Terminology

Some Terms

- Spider—a script that copies content from a source to the search system. Sometimes called a *crawler*
- Indexing—the process of opening a document and identifying key words, phrases, and creating other information about the document
- Metadata—a buzzword that means information about the document that has been indexed; e.g., date (simple) to pointers to key paragraph (complex)

More Key Terms

- Query processor—function that takes the user's query and converts it to a form to allow documents matching the query to be displayed in a hit list (list of results)
- Saved search—A stored query so a user can click on a heading and retrieve hits that match that stored query; e.g., News heading on Yahoo is a stored query
- File or document types—Specific file formats such as Word, Excel, Adobe PDF, etc.

Other Terms?

- Taxonomy
- Ontology
- Controlled vocabulary
- Text mining
- Entity extraction
- Bayesian
- Linguistics / natural language processing

Enterprise Search

Identifying, indexing, and exposing via a "search box" or "point and click" interface information for employees or authorized users to access in order to do their work.

Enterprise Search Examples

- A company wants to index contracts and marketing literature for certain employees
- A trade association wants to index information related to annual conferences and the data about its membership
- A government agency wants to index proposals, statements of work, and reports for everyone in the agency

Enterprise Search

Category	Enterprise Search
Platform	Varies by search system vendor; usually run on licensee's premises. Solaris support more common.
Content acquisition	Some data may be copied directly to the search engine using a script. Other content obtained by a software crawler.
Search database tables	Search system expected to index data in a database table
File formats supported	A wide range of file types including provisions for handling legacy file types for data on mainframes
Index updates	Certain content must be indexed in near real time; other content may have different schedules
Performance	Dependent on the licensee's network infrastructure and computational environment
Security	Security involves the system as well as user access to specific content
Usage tracking	Active monitoring required using

Web Search + Enterprise Search

Category	Web Search	Enterprise Search
Platform	Often linux, occasionally Microsoft; may be hosted by a third party	Varies by search system vendor; usually run on licensee's premises. Solaris support more common.
Content acquisition	Typically via spider	Some data may be copied directly to the search engine using a script. Other content obtained by a software crawler.
Search database tables	Optional; can be supported	Search system expected to index data in a database table
File formats supported	Web and standard office formats such as Word and Adobe PDFs	A wide range of file types including provisions for handling legacy file types for data on mainframes
Index updates	Usually via scheduled spidering, with some incremental indexing	Certain content must be indexed in near real time; other content may have different schedules
Performance	Controlled with caching and other shortcuts	Dependent on the licensee's network infrastructure and computational environment
Security	System security the focus	Security involves the system as well as user access to specific content
Usage tracking	Search logs	Active monitoring required using

Enterprise Search Discussion

- What problems do you anticipate with determining what content should be indexed and what content should not be index?
- Which is easier? Indexing text or information in a database?
- What mechanism is needed to index new and changed documents for the government agency?

Web Site Search

Indexing information on an organization's Web site to allow visitors to locate and access that information.

Search Your Web Site

Category	Site Search
Platform	Unix variant or Windows
Content acquisition	Spiders a Web site or sites
Search database tables	More common
File formats supported	Web and standard office forma such as Word and Adobe PDF
Index updates	Web master controls
Performance	Controlled with caching and ot shortcuts
Security	SSL or other Web techniques
Usage tracking	Search logs

Search Your Web Site+Intranet

Category	Site Search	Intranet Search
Platform	Unix variant or Windows	Unix variant or Windows
Content acquisition	Spiders a Web site or sites	Some data may be copied directly to the search engine using a script. Other content obtained by a software crawler.
Search database tables	More common	Search system expected to index data in a database table
File formats supported	Web and standard office formats such as Word and Adobe PDFs	A wide range of file types including provisions for handling legacy file types for data on mainframes
Index updates	Web master controls	Certain content must be indexed in near real time; other content may have different schedules
Performance	Controlled with caching and other shortcuts	Dependent on the licensee's network infrastructure and computational environment
Security	SSL or other Web techniques	Security involves the system as well as user access to specific content
Usage tracking	Search logs	Active monitoring required using a wide range of techniques. Detailed reports required to comply with copyright or security

Web Site Search Examples

- A company wants to index the information on a Web server, located at a hosting company, for those visiting the Web site
- A trade association wants to index two Web sites: one site would be available only to employees; the other site would be available to anyone. The server is located at the association.
- A government agency wants to index three government Web sites and make the information available to anyone

Discussion of Web Site Search

- Which is easier? Indexing the content on the hosted site using software loaded on the hosting company's server or using a third party managed service to index the content?
- What's needed to make sure the right content is available to the authorized viewer?
- What must be done to ensure that the information on each server is not classified and up-to-date?

Enterprise Search and Site Search

- Enterprise search can be set up to include content on:
  - The organization's Web sites
  - Other Web sites
  - Third-party content (Factiva, for example)
- Enterprise search tends to be more complex than site search for two reasons:
  - Access to certain "sensitive" information
  - Need to make certain information available in "near real time"

Differentiate carefully...

Enterprise search

Web site search

Internet search

Points to Consider

- Vendors will explain that their search system can do enterprise search AND Web site search
- Depending on circumstances, the two can be:
  - Separated
  - Operated on a single system
- Mixing enterprise search which supports work tasks and Web site search which may have a marketing angle leads to potential misunderstandings