



Enterprise Search | Tools and Selection Techniques

Gilbane Conference, Washington, DC | June, 2007

Theresa Regli
Principal
tregli@cmswatch.com

Copyright © 2007 CMS Watch | All Rights Reserved

Agenda

- Search in the broader ECM Market
- The difference is in the details: key features of enterprise search
- Tiers and characteristics of enterprise search vendors
- The path to selecting the “right” enterprise search tool
- Q&A



Independent, buyer-centered evaluations of content technologies & practices



The Web CMS Report evaluates 30 Web CMS packages.



Enterprise Search Report evaluates 28 Search vendors.



Enterprise Portals Report evaluates 15 products of 13 major Enterprise Portals vendors.

New:

- The ECM Suites Report
- Web Analytics Report

Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

3

“Markets are Conversations” -- Cluetrain Manifesto

Industry-leading, scalable solution for the enterprise, fully customizable for your needs, with robust API, Web Services support, and lowest TCO in its class, blah, blah, blah...



How does it work? How much does it really cost? What's wrong with it? Who are their real competitors? Do their consultants know my industry? Will my co-workers actually use it?



Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

4

The Search Marketplace



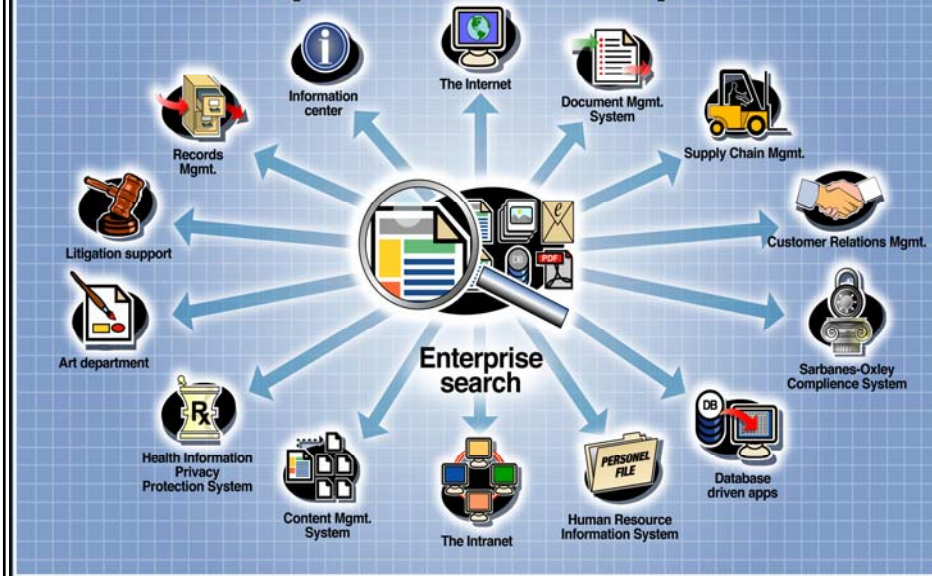
Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

5

Enterprise Search - Which Option



Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

6

What comes first?

Search Vendors Say:

We're more than « Search »

"Search is the human interface into all corporate information."

-- Dave Girouard, Google, AIIM 2006

"Search is the future of content management."

-- John Lervik, FAST Search, CMF 2005

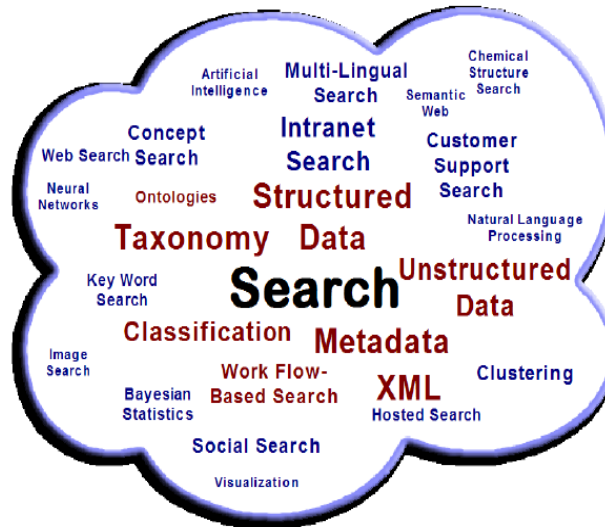
But the reality is...

- Vendors recognize importance of search
 - Beware of how they push enterprise search as the answer to an organization's need for a single, unified window into everything the organization knows at any point in time
- The ultimate knowledge management machine simply does not exist: the typical enterprise search system does not contain "all" the organization's content
- Limitations on available information include:
 - Security considerations
 - Inability to integrate specialized content
 - Difficulty reconciling structured and unstructured content
 - Cost, time, and difficulty required to incorporate diverse content repositories



Current trends in search

- As search sector changes, distinctions among different “flavors” of search technology, features, and functions become more difficult to make.



The Google Effect

- Increased staff expectations
 - Everyone has their “Google experience” in mind when selecting a search engine
- Google revolutionized search with its relevance ranking – which relies heavily on link popularity
- Within enterprises, corpus of information is much smaller and often “unlinked”
- In reality, Popularity ≠ Authority



Two primary types of search

“Web search” – systems intended to find web content (public or intranet)

“Enterprise Search” – systems intended for use internally within an organization in a variety of formats

Key differences:

1. What content it finds
2. How the system finds that content

Web vs. Enterprise Search

Category	Web Search	Enterprise Search
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 if there is a web application front-end	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 a wide range of techniques. Detailed reports required to comply with copyright or security mandates.

Web vs. Enterprise Search

- There is a right time, place and budget for both
 - For searching websites – public or Intranet – an enterprise search tool may well be overkill
 - To **extract** and **merge** content from multiple file systems, databases, and application repositories – especially those without a spiderable Web front end – a basic web search engine will likely not suffice

Agenda

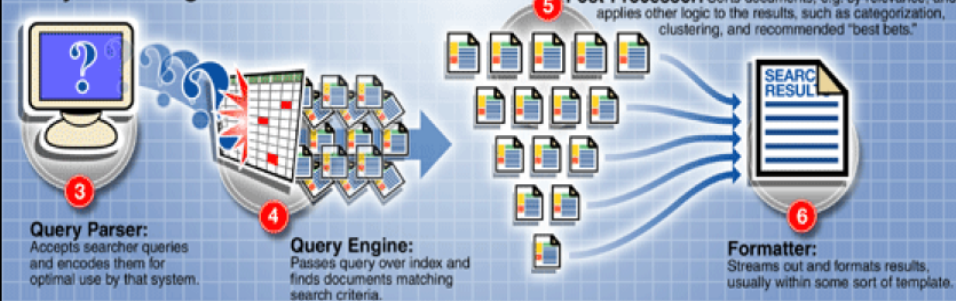
- Search in the broader ECM Market
- The difference is in the details: key features of enterprise search
- Tiers and characteristics of enterprise search vendors
- The path to selecting the “right” enterprise search tool
- Q&A

How Enterprise Search Subsystems Work Together

Content Indexing:

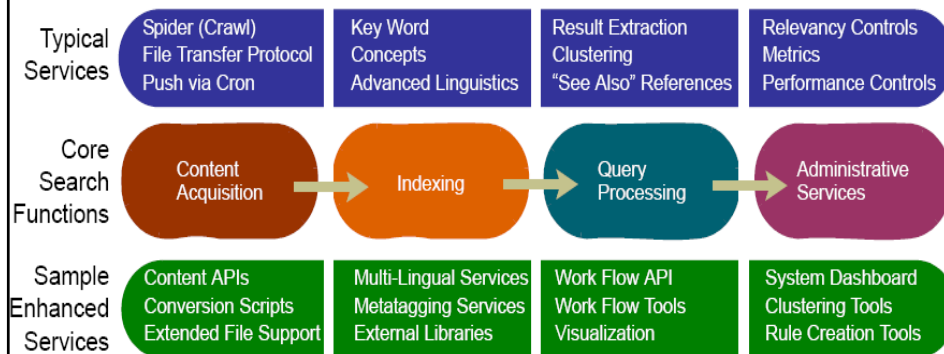


Query Processing:



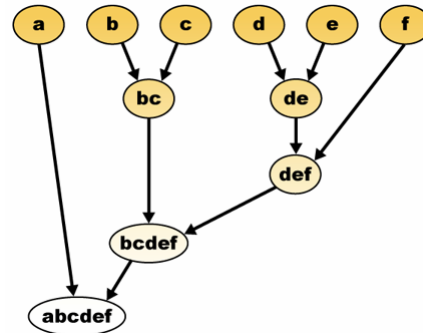
Differences are in the details | The HOW, not the WHAT

- The core functions of an enterprise search system can be sandwiched between tailored services and a wide range of enhanced services. Even "simple" enterprise search systems are complex.



Differentiator | Clustering Results and Related Terms

- Clustering search engines organize results into broad topics, allowing users to narrow their search results.
- Related terms support users who need more, rather than fewer, results by pointing them to terms they may not have used, but relate to the original query.



Differentiator | Metasearch or “Federated” Search

- Allows a user to search across multiple disparate databases, using one integrated search interface
- Useful in companies that have licensed multiple search products or employ various content management systems that embed their own search facilities
- Security is important
- De-duping, merging, and returning results are overhead-intensive

Differentiator | Content Types

- **Diverse Content Types:**
 - Unstructured
 - Microsoft Word documents, emails
 - Structured
 - Database tables
 - Semi-structured
 - XML, HTML
- Analysis of the information to be indexed as part of the requirements and pre-indexing phase of a search project is *essential* to controlling the costs associated with enterprise search

Three main options for delivering enterprise search

- Local installation
- Hosted Search
 - On premises
 - Off-site
- Appliances

Options for delivering enterprise search | Local install

- Search installed on your organization's premises by your staff or by people working under contract and acting on your organization's behalf



Advantages	Disadvantages	When to Use
More perceived control, usually more customization options	Customization, tuning, and other basic functions may be outside the IT department's skill set.	When management or operational issues warrant keeping the search system "in house"

Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

21

Options for delivering enterprise search | Hosted

- Search system is located at a data center operated, in part, by employees or contractors with clear divisions of labor

Advantages	Disadvantages	When to Use
Tightly defined functions with some customization options; no burden on the licensee's IT staff	Security via virtual private network or other means must be set up; customization options may add to monthly fee	When basic search is needed and IT and other resources are limited

Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

22

Options for delivering enterprise search | Appliance

Appliances

- Customer gets a “box” and the licensee accesses search on a dedicated server or servers
- Appliances are usually local installations



Advantages	Disadvantages	When to Use
Easy to install, maintain, and scale	Costs can be difficult to control when the number of documents and their changes rises rapidly	Departmental and small business should consider an appliance if a hosted service isn't appropriate.

Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

23

Agenda

- Search in the broader ECM Market
- The difference is in the details: key features of enterprise search
- Tiers and characteristics of enterprise search vendors
- The path to selecting the “right” enterprise search tool
- Q&A

Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

24

Categorizing search vendors

- Infrastructure vendors
- Microsoft-centric
- Traditional majors
- Specialized / Niche
- Text miners
- Turnkey appliances
- Public sector

Infrastructure vendors

Search tools provided by the major infrastructure vendors

- IBM

Products include STAIRS, DB2 SQL, OmniFind, iPhrase, X1 and others

- Microsoft

SharePoint search proliferates, "ready to run" solution for Windows-compliant environments

- Oracle

Oracle "Text" may be the most heavyweight engine you already own; SES 10g is enterprise product

- SAP

TREX works with different data types managed by SAP platform and integrates with BI functions

Microsoft-centric vendors

For searching in Microsoft-based environments

- Coveo

Tries to out-search SharePoint search. Specialized relevance and ranking, summarization, concept extraction etc.

- Mondosoft MondoSearch

Search, automatic categorization, advanced analytics. Tight integration with MS server products.

- dtSearch

Fast indexing and response, excellent MOSS integration, support for distributed searching

- GoDaddy

Low-cost and quick, for office files and HTML

Traditional majors

The big boys...or girls, if you like!....of the enterprise search space.

$$p(\theta|x) = \frac{p(x|\theta).p(\theta)}{\sum_{\theta' \in \Theta} p(x|\theta').p(\theta')}$$

- Autonomy

Built a mixed reputation on hard-core statistical data-crunching methodology and automated pattern recognition. Acquired prime competitor Verity

- FAST

Focus on speed and scalability against large, dynamic web repositories

- Endeca

Popularized “guided navigation,” works best with highly structured data repositories

Turnkey appliances

"Plug-and-play" boxes are increasingly popular, and while they remain best-suited for Web search, that's changing.

- Google
Impeccable brand compensates for less mature technology
- Thunderstone
Strong technology compensates for weaker brand

Agenda

- Search in the broader ECM Market
- The difference is in the details: key features of enterprise search
- Tiers and characteristics of enterprise search vendors
- The path to selecting the “right” enterprise search tool
- Q&A

Before you begin any IT project...

Make sure you have solidified business objectives

- Make a list of what you expect to achieve
- Keep them close by throughout the project

Develop your requirements

- Consider outside help
- User-centered design principles apply

Essential to identify content types

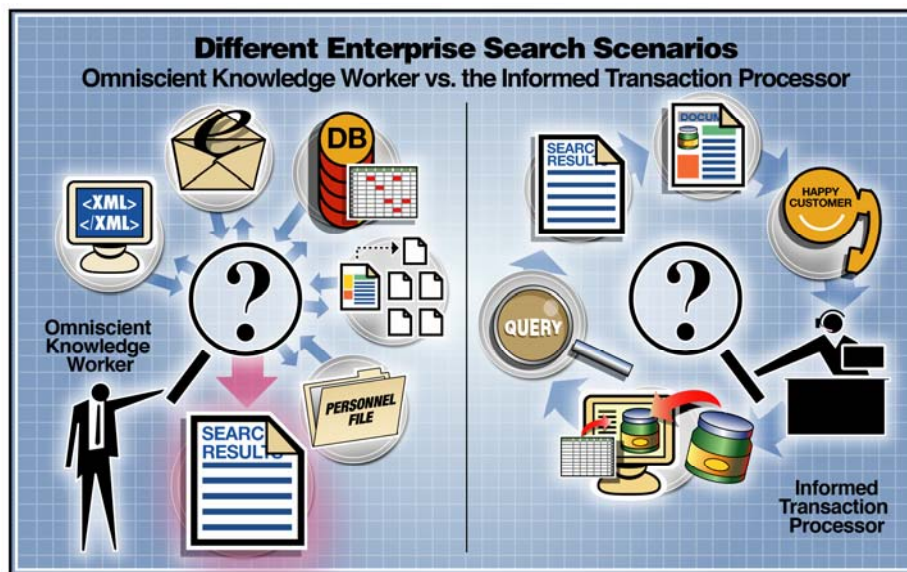
All successful enterprise information projects start with content analysis

- Unstructured data such as office files (Word, email)
- Structured data such as database tables
- Legacy mainframe systems
- Semi-structured data such as XML and HTML

Roles in an enterprise search project

- **Project Manager**
 - “Owns” the project
- **Information Professional**
 - Determines what documents should be indexed on what cycle and with what thesauri, vocabularies, or word lists
- **Information Technology Professional**
 - Solves problems with programming, system and network skills
- **Financial Professional**
 - Tracks costs and develops the analyses which support additional investments
- **Consultant**
 - Helps during critical product-selection and testing phases
- **Searcher – Most important!**
 - Provides feedback and ideas at every stage of the search system project

The Importance of Scenarios



Typical search scenarios

- Customer / Self-service
- Website search
- E-commerce search
- Legal search
- E-discovery
- Departmental search
- Hybrid internet search
- Executive dashboard
- Sci/Tech/Medical
- Multi-repository search
- Desktop search

Parting selection advice....



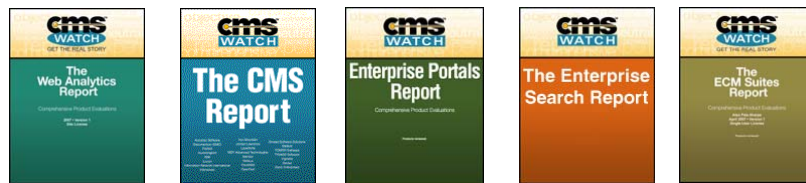
Enterprise Search

- Put together a team
- Assemble a sample / **test** corpus
- Devise initial scenarios
- Initial user **testing**
- Architect infrastructure changes
- **Test** some more
- Anticipate UI modifications
- **Test** still more

Thank you!

tregli@cmswatch.com

www.cmswatch.com



Copyright © 2007 CMS Watch | www.cmswatch.com



Differentiating Search Products

37