Limited Access Document

Gilbane Seminar

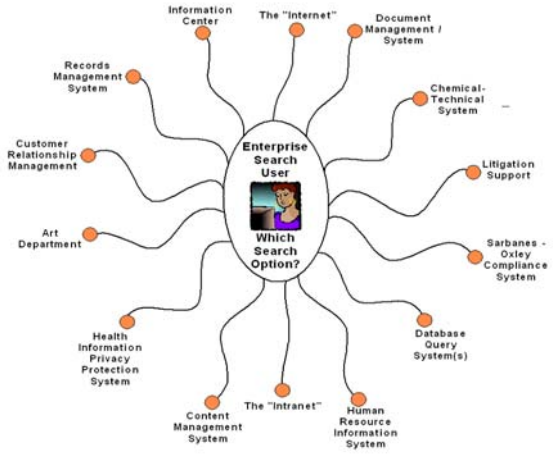
December 2, 2004

ARNOLD
Information
Technology

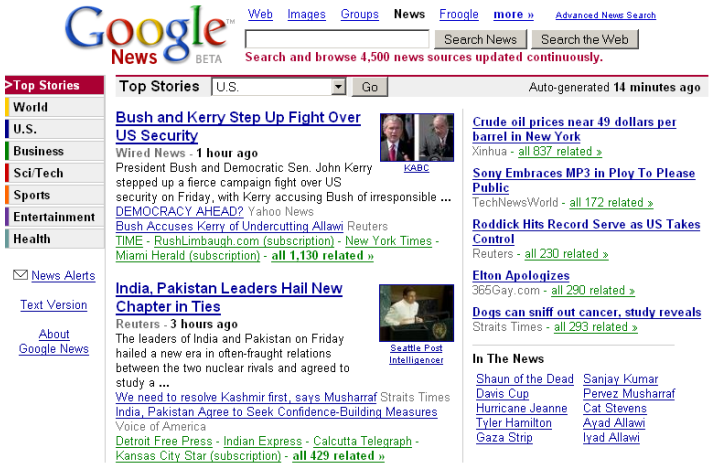
Introduction

- Background
- Requirements
- Selected Company Profile
- Cost Analysis

Search “Reach”



Google Defines “Search”



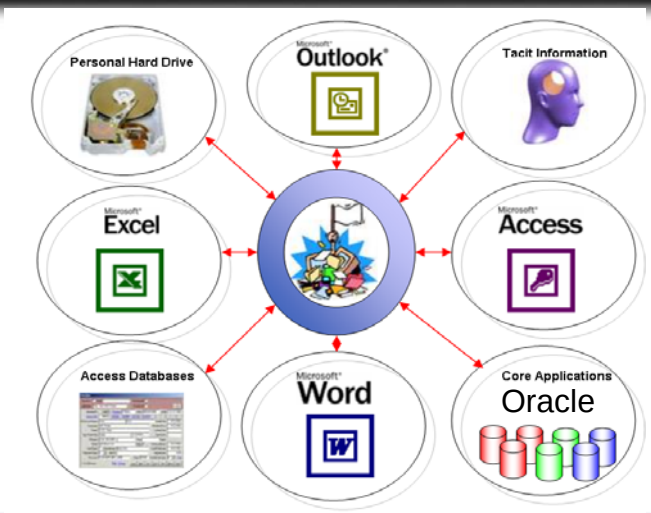
Some Terms

- **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)**

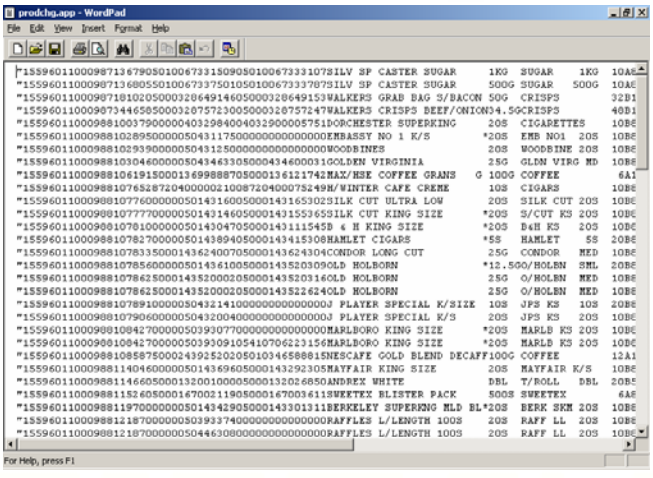
More Key Terms

- **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.**

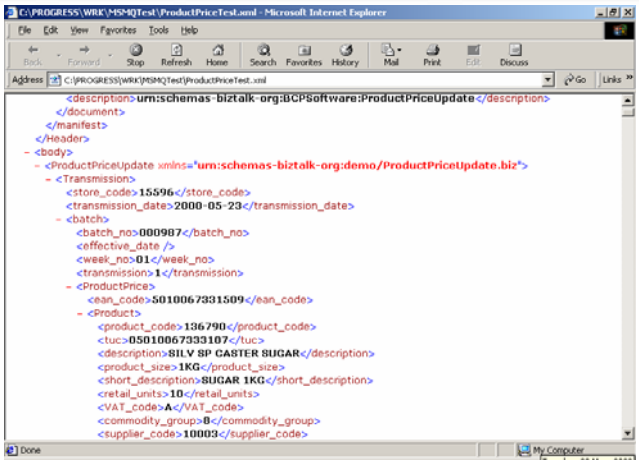
Content Types in Organizations



From This ...



To This...



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

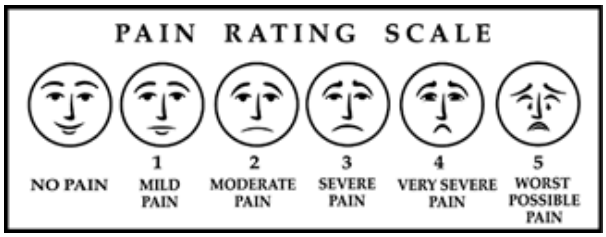
Enterprise Search Discussion

- 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?

Know Your "Market"

- **Customers = Your Employees**
  - **Goal is to provide employees the means to find information quickly that will enable them to do their job**
- **What do your "customers" want?**
  - **What information are they searching for?**
  - **What shortcuts are they taking because they can't find information?**
  - **How many interfaces do your "customers" use to find information?**

Rate Your Solution



Web Site Search

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

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**

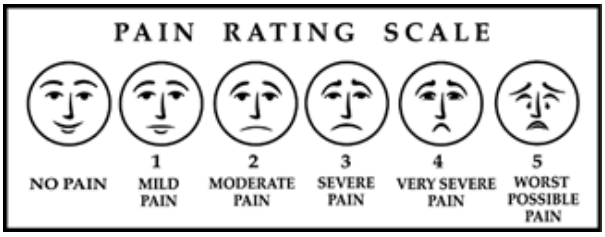
Discussion of Web Site Search

- **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?**

Know Your Market

- **Customers = Your clients**
  - **Goals vary based on your business needs**
  - **Sales goals different than information goals**
- **What do your "customers" want?**
  - **What information are they searching for?**
  - **Does the search engine match your business need?**
  - **How many interfaces do you use?**

Rate Your Solution



Web Search has low overall customer satisfaction

All Search is Not Created Equal

- **Enterprise Search**
  - **Helping Employees find info**
- **Site Search**
  - **Helping drive sales**
- **Intranet / Extranet search**
  - **Serving different needs?**
- **Web Search**
  - **"How do I make my site appear at the top of Yahoogle?"**

One Size Fits All?

Discussion

- **When a vendor say, “Our search system can do it all”, what does the vendor mean?**
- **What must be done to index information held in real-time systems running on mainframes?**
- **How do search systems deal with the jargon and specialized vocabulary in your organization?**
- **When an employee misspells a name, how can the search engine deliver the right results?**

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"

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

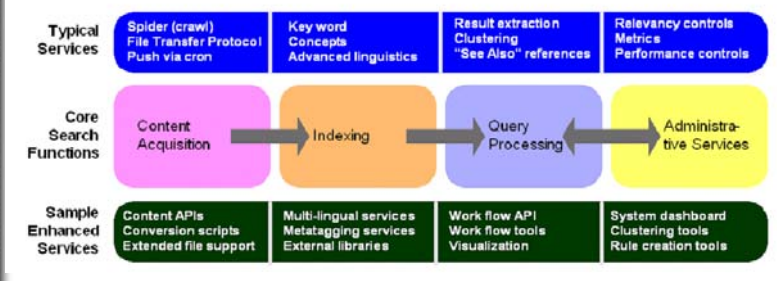
Web Search : Enterprise Search

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

Site Search : Enterprise Search

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

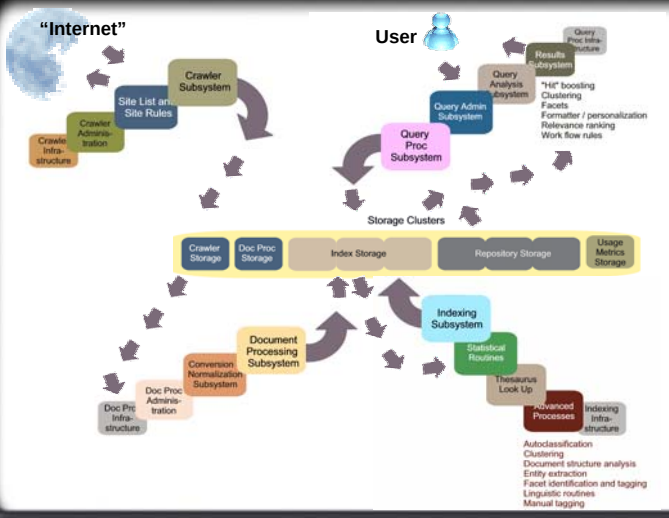
How It Works: Search "Sandwich"



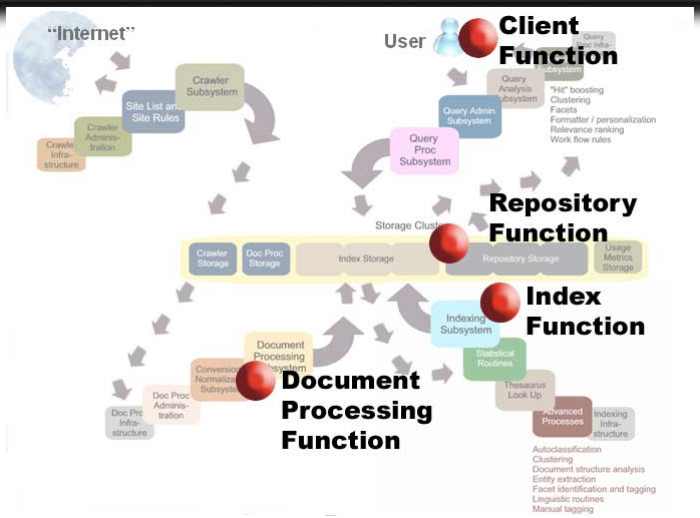
Search… Multiple Subsystems

- **Spider:**
  - **Locate content and changed content**
  - **Copy it to the search system**
- **Index:**
  - **Identify words**
  - **Extract / assign metadata**
- **Query processor:**
  - **Parse user's query**
  - **Process link for stored search**
- **Administration**

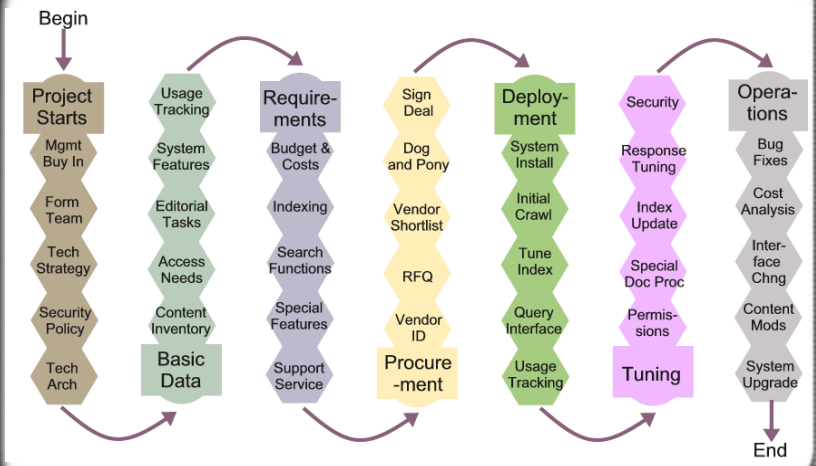
Search: One Machine or 10,000



Enhanced Search Option Location



Search Itself: Not Well Understood



How to Avoid Common Pitfalls

- **Get smart—Assume nothing**
- **Identify key stakeholders—Enterprise search is not a consumer audience**
- **Get the support of management—Lack of resources (people and money) means trouble**
- **Prepare a business case and cost analysis—Search is not perfect even with unlimited resources**

Summary: Four Search Myths

- **The myth: Search is trivial.**
  The myth: No, search—even with the Google Appliance--is hard due to access control and updating to meet the needs of colleagues.
- **The myth: Search is like Google.**
  The reality: No, search is not Google even when you have a Google Appliance (no secret "popularity" sauce…yet). Content types and user needs are different from a free Web search service.

Search Myths

- **The myth: Performance is not a problem.**
  The reality: Yes, performance is always a problem. Updating indexes requires network bandwidth, storage, and CPU slices.
- **The myth: Our IT people are able to do search.**
  The reality: No, search requires specialized support. One example: document retention for compliance with Federal regulations.

Next… Requirements

- **A short break**
- **Any questions?**