Building Blocks of Information Access:

Information Architecture, Content Management, and Search

Agenda

- The Findability Problem
- Enterprise Content Management as a Solution
- Enterprise Search as a Solution
- Information Architecture as a Necessity!
- Content Technology Trends

The Growing Problem

- Digital content is expanding at almost unmanageable rates
  - New information worldwide has been increasing on average 30% a year (doubling every three years)*
  - Getting **access** to the ***right*** information is an increasingly acute challenge for enterprise employees and customers alike

*http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Findability

Findability is *the quality of being locatable or navigable*

- At the core of information access is the findability of information.
- ***Information should be easy to discover or locate***
- Information access is about helping users find documents that satisfy their information needs
- Not necessarily something that you know that you're looking for
    - Remember, someone may be looking for something they've never seen or touched before

Building Blocks

Information Access

- Browse
  - Traversing an organized repository
- Search
  - Querying information sets and obtaining documents

Information Organization

- Content Architecture
  - Structure and composition of a repository, information collection, or individual document
- Content Intelligence
  - Enriching content with additional information

Agenda

- The Findability Problem
- **Enterprise Content Management as a Solution**
- Enterprise Search as a Solution
- Information Architecture as a Necessity!
- Content Technology Trends

Building Blocks

Information Access

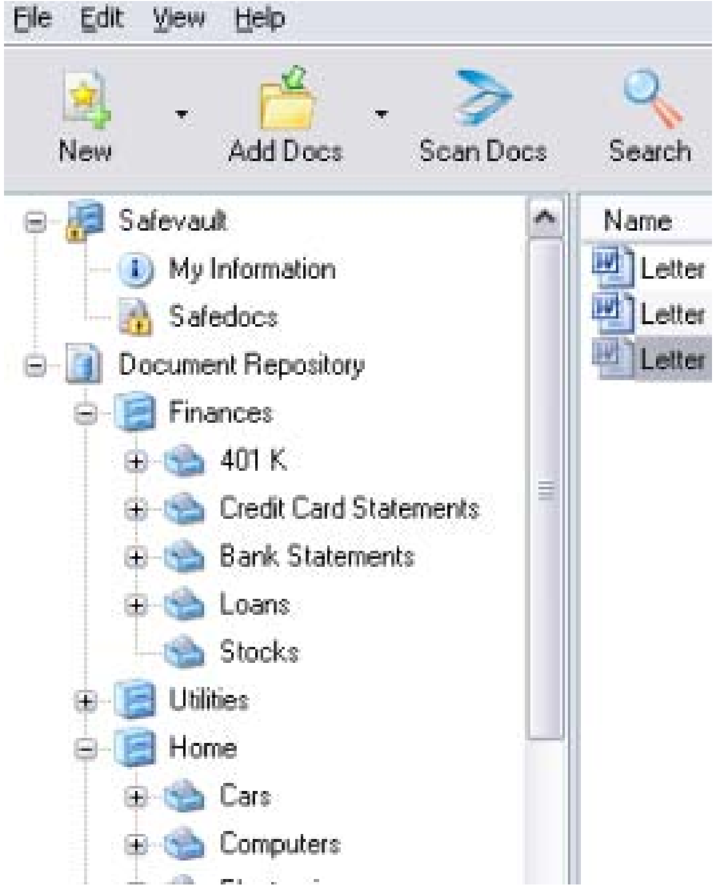
- Browse
  - Traversing an organized repository
- Search
  - Querying information sets and obtaining documents

Information Organization

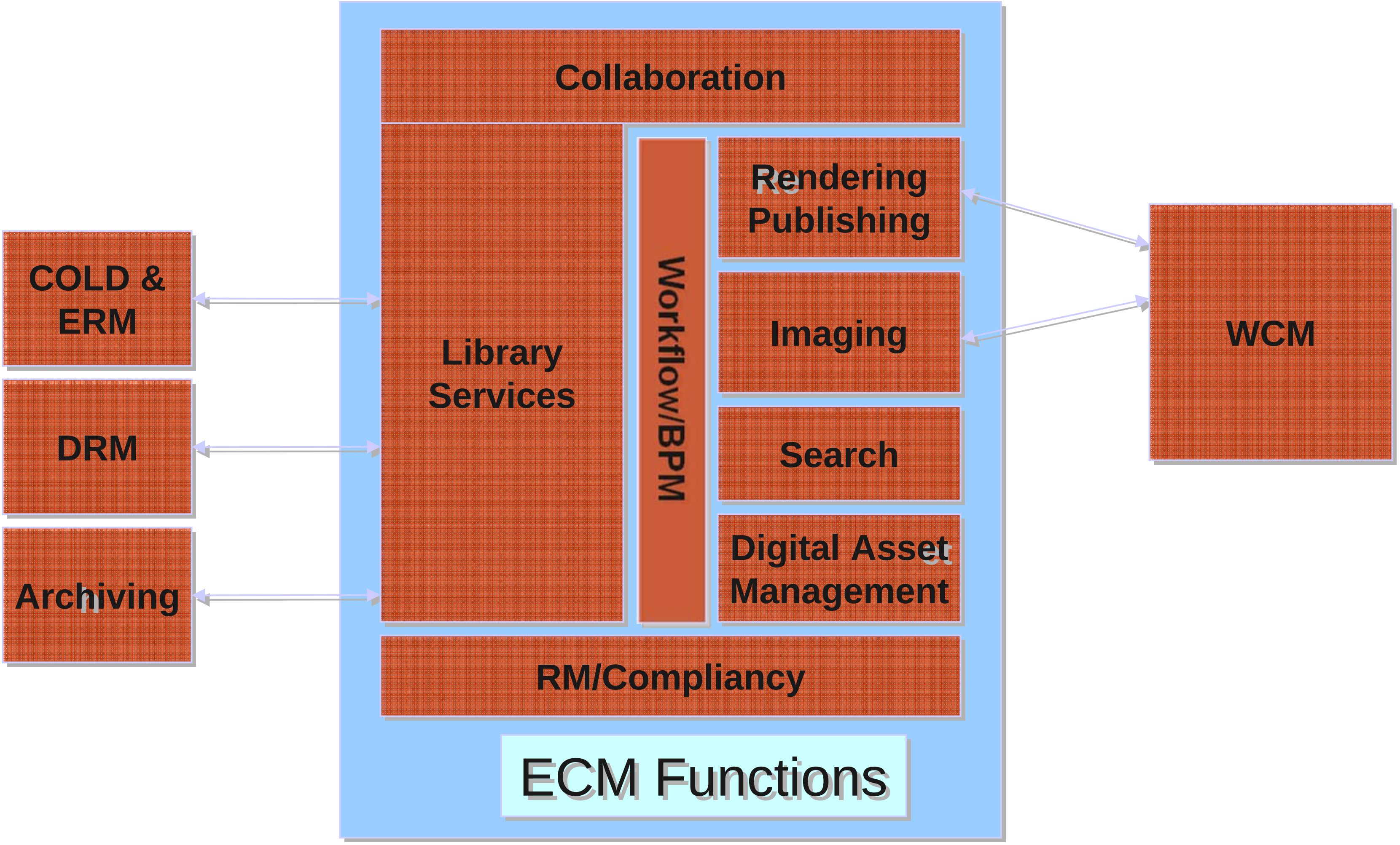
- Content Architecture
  - Structure and composition of a repository, information collection, or individual document
- Content Intelligence
  - Enriching content with additional information

Access via Browse

- Browsing is usually the first option for users seeking information or documents
  - Desktop and enterprise file systems
  - Content management system repositories
  - Intranets and Websites
- If users can't find via browse, then they resort to search
- Some users will go straight to search
  - This is partly generational
  - Depends on type of organization



What is ECM?


Collaboration
Library Services
Workflow/BPM
Rendering Publishing
Imaging
Search
Digital Asset Management
RM/Compliancy
ECM Functions
COLD & ERM
DRM
Archiving
WCM

Effective Browsing

- Browsing effectiveness is highly dependent on
  - navigational structure
  - folder labeling
  - the location of the content
  - In short: depends on how *organized* the content is…

- Content technologies typically use "virtual folders" to represent different classifications
  - These allow for multiple paths to the same content
  - In contrast: physical file system forces documents to a single "place"
  - Ideally content should be *cross-referenced*, but not *duplicated*

Scenarios vs ECM

- ECM as Business App
- ECM as Infrastructure
  - High Volume Imaging
  - Engineering Docs
  - Forms Processing
  - Regulatory Compliance
  - Case Management
  - Workgroup Collaboration
  - Marketing Information
  - Technical Docs
  - Enterprise Web Publishing

Good things about ECM Systems…

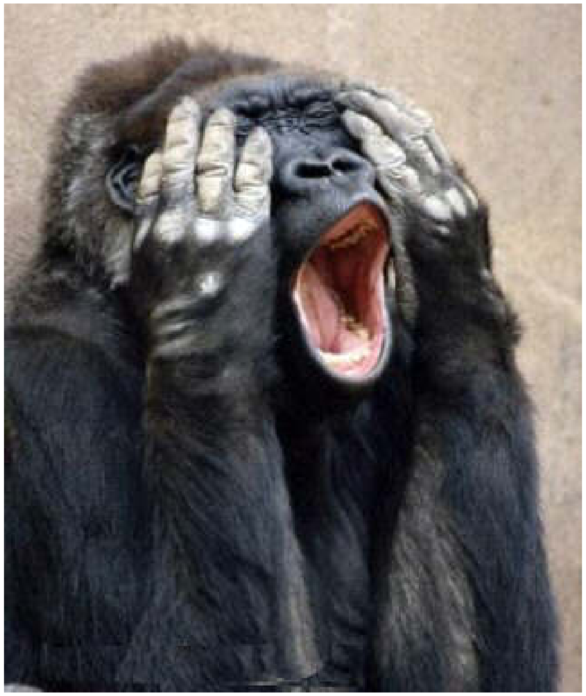
- Everything in one place
- Create once, re-use
- Process-oriented
- Easy-to-use – familiar models

- Enable knowledge management, compliancy, etc.

There is a compelling argument for using ECM only.

Bad things about ECM software

- Expensive
- Very expensive
  - Don’t forget services costs
- Major time-consuming updates
- Difficult to keep up with massive amounts of content



Agenda

- The Findability Problem
- Enterprise Content Management as a Solution
- **Enterprise Search as a Solution**
- Information Architecture as a Necessity!
- Content Technology Trends

Building Blocks

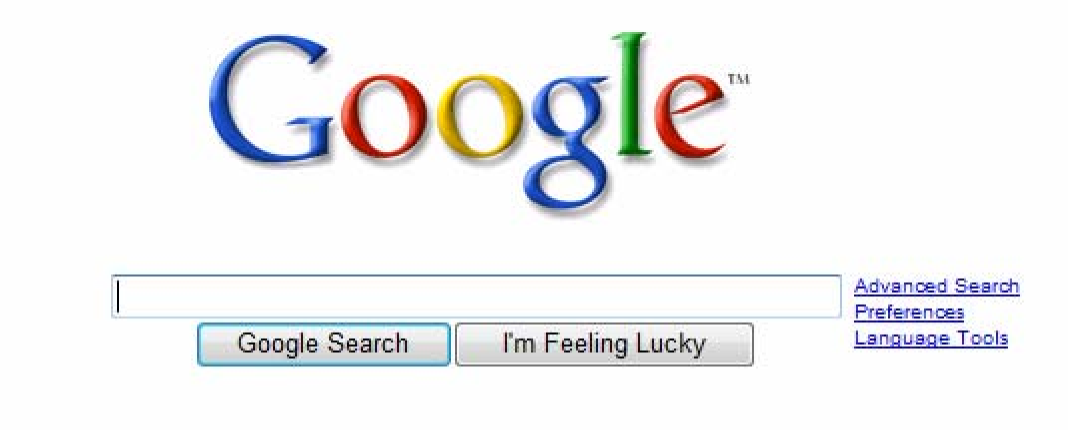
Information Access

- Browse
  - Traversing an organized repository
- Search
  - Querying information sets and obtaining documents

Information Organization

- Content Architecture
  - Structure and composition of a repository, information collection, or individual document
- Content Intelligence
  - Enriching content with additional information

Search is the answer! 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

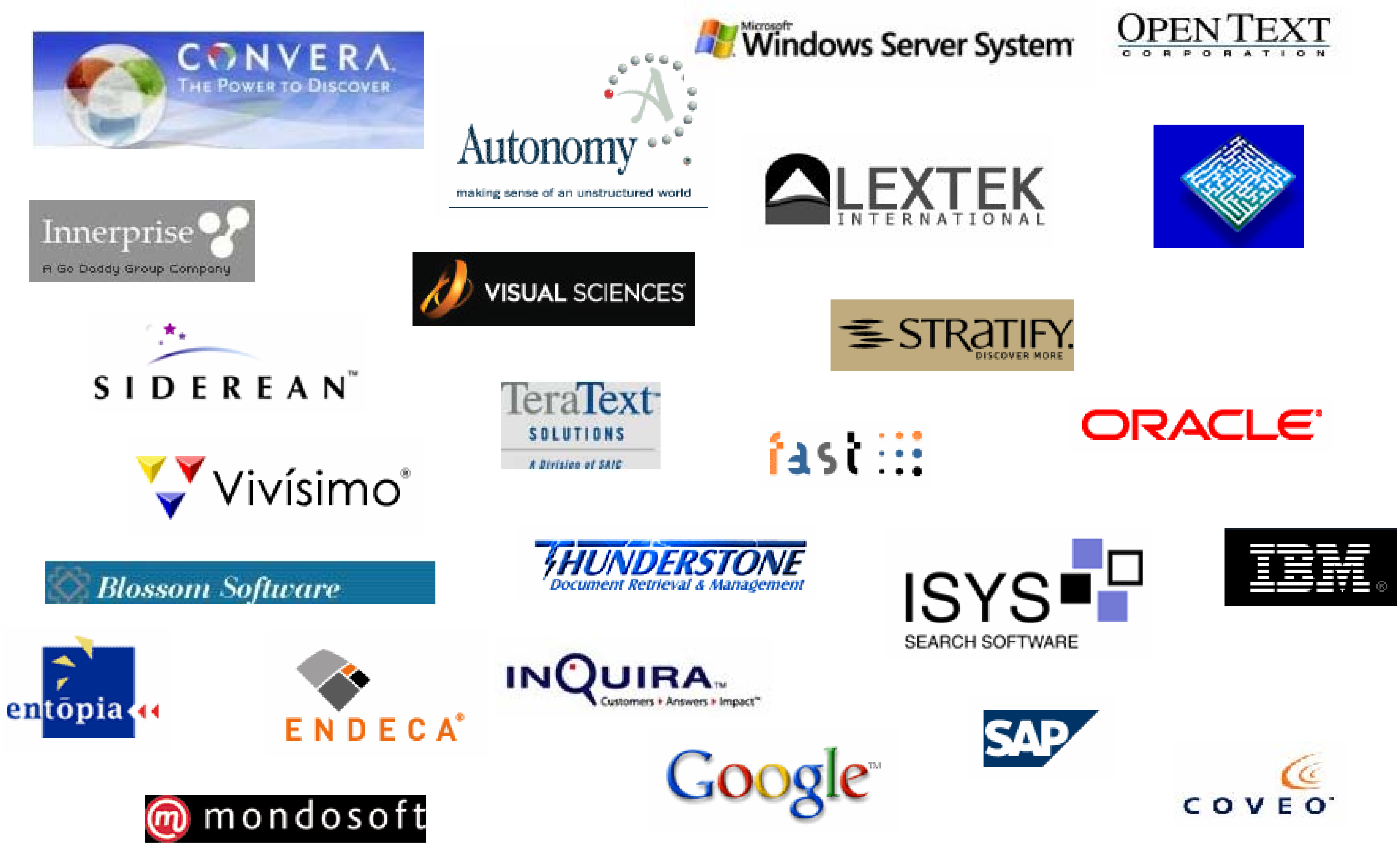
Google not necessarily the answer

- Within enterprises, corpus of information is much smaller and often “unlinked”
- In reality, Popularity ≠ Authority
- Example: Latest version of a document

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

The Search Marketplace



What is Search?

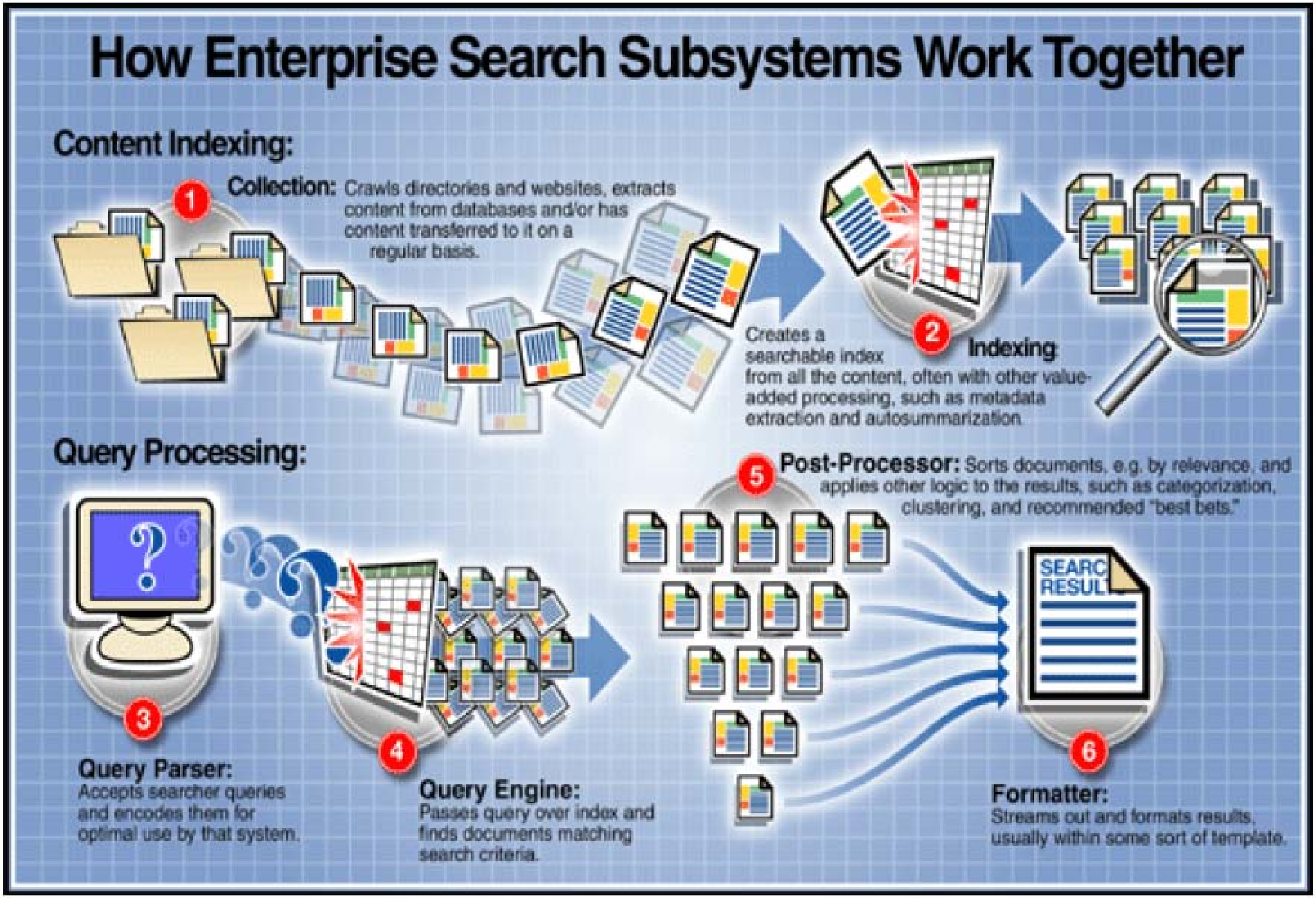
search box

- Search is an application or tool for finding information via search term
- Search is omnipresent, and essential
  - But: there is much ignorance about how search engines work
  - Most end-users shouldn’t need to know; they just assume “magic”
- Advanced display techniques can blur the line between search and browse

Components of a Search System

- Most search systems have four major interdependent components of varying complexity:

  1. Content acquisition
  2. Indexing - The technology to take a document, index the words in that document, and configure that index such that a user can search it.
  3. Query processing
     - Parsing
     - Matching
     - Post-processing
  4. Formatting results



How Enterprise Search Subsystems Work Together

Content Indexing:

1 **Collection:** Crawls directories and websites, extracts content from databases and/or has content transferred to it on a regular basis.

2 **Indexing** Creates a searchable index from all the content, often with other value-added processing, such as metadata extraction and autosummarization.

Query Processing:

3 **Query Parser:** Accepts searcher queries and encodes them for optimal use by that system.

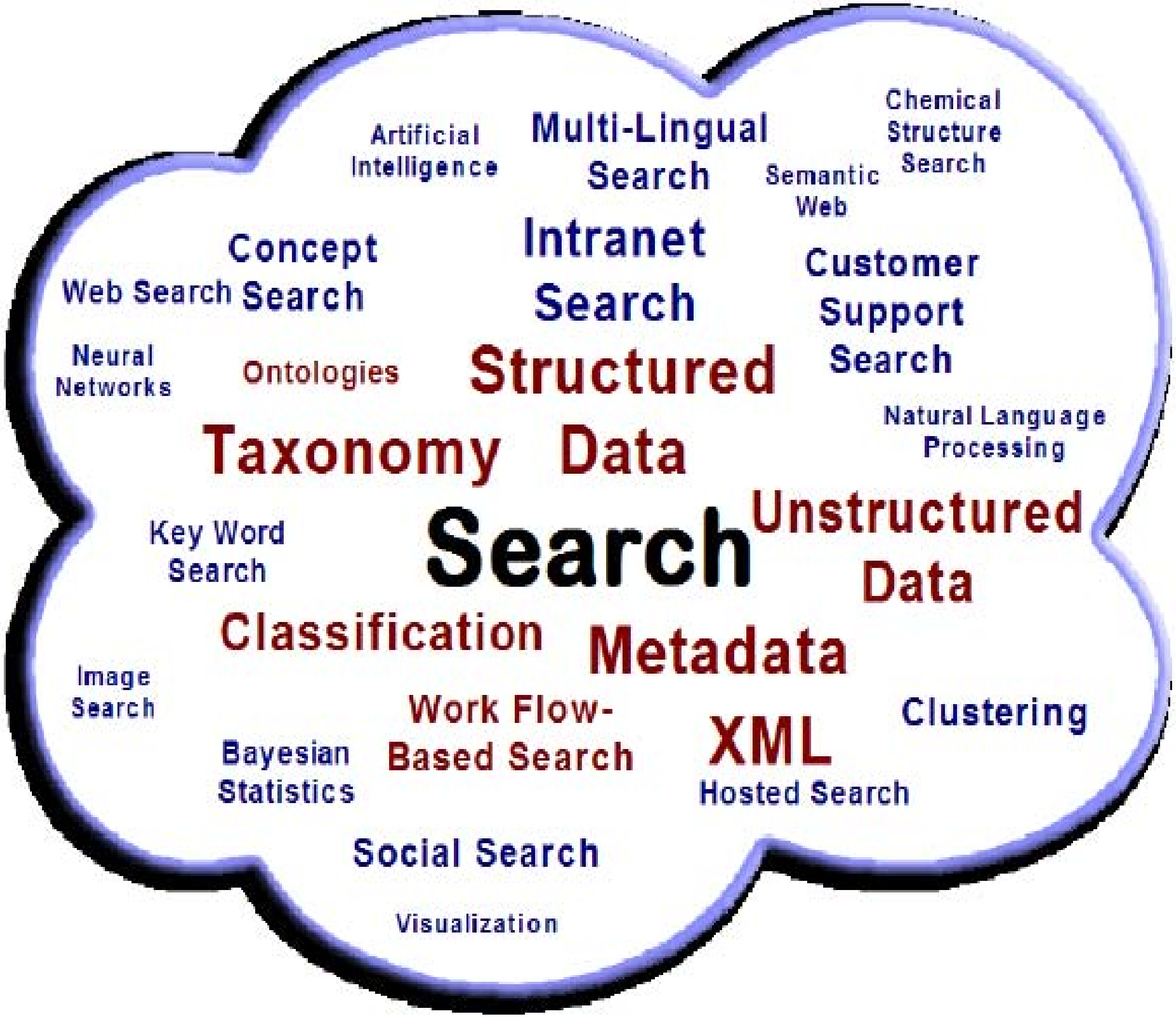
4 **Query Engine:** Passes query over index and finds documents matching search criteria.

5 **Post-Processor:** Sorts documents, e.g. by relevance, and applies other logic to the results, such as categorization, clustering, and recommended "best bets."

6 **Formatter:** Streams out and formats results, usually within some sort of template.

Current trends in search

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



Differentiator: Website vs. Application vs. Enterprise

- **Website Search**

  Systems intended for use by individuals seeking Web content, both within and beyond the enterprise

- **Application Search**

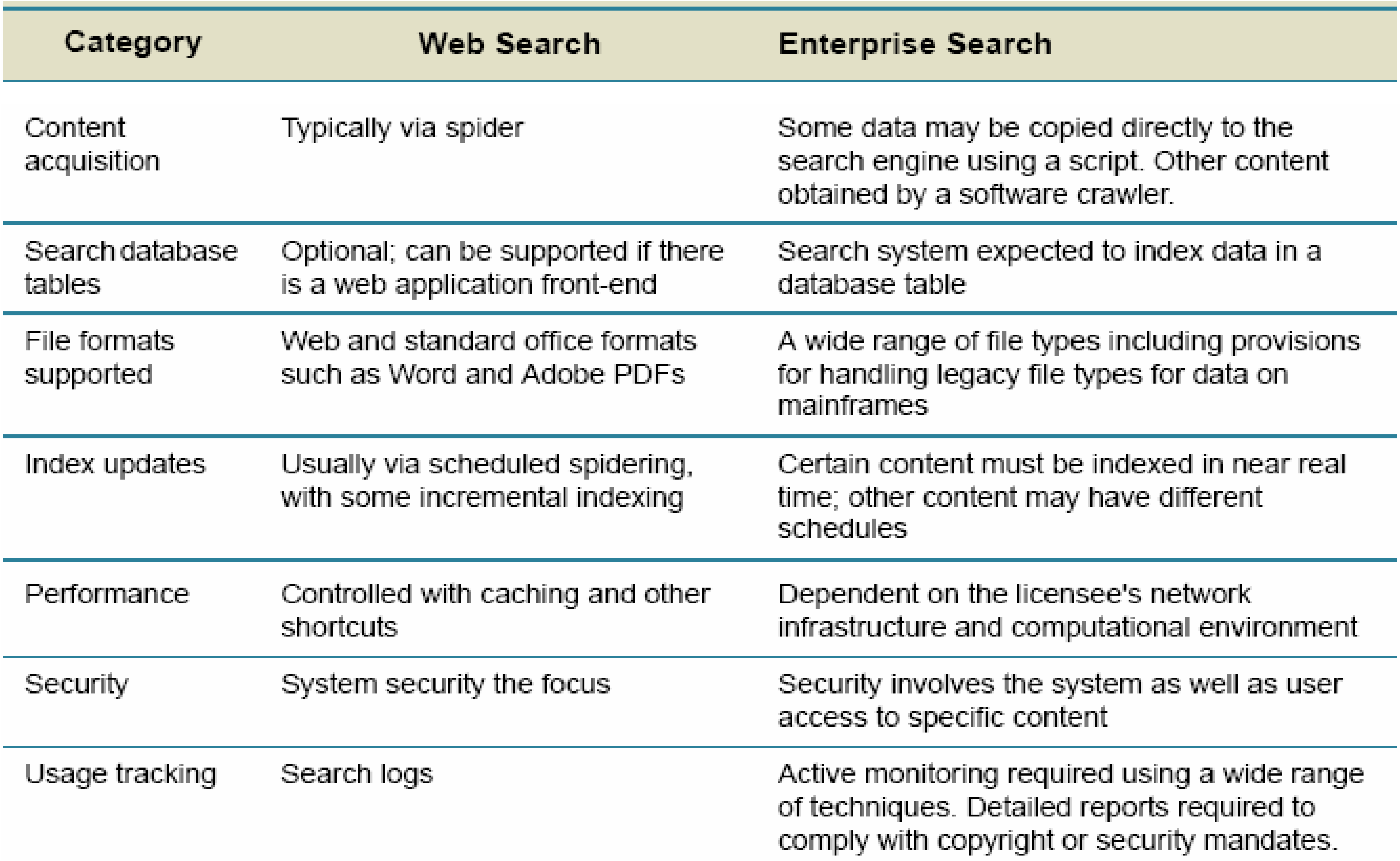
  Search systems within a single software application, designed to provide localized search services to application users. ***Desktop search*** works like application search – localized to your desktop.

- **Enterprise Search**

  Systems intended for use within an organization by employees seeking information held internally by the organization in a variety of formats and locations, including databases, document management systems, and other repositories

- **Key Differences:**
  - How content is retrieved and indexed
  - Breadth of content and file types
  - Cost and complexity

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.

Content Structure

Content can be structured, semi-structured, or unstructured

Structured	Semi-structured	Mostly Unstructured
Databased content	XML content Some Web content* Forms-based word processing files *HTML tends to be semi-structured	Scanned images Video Audio Email body Photographs Presentations Most word processing files Most Web content Chat / IM sessions Written correspondence

On average, 80% of all content in an organization is unstructured

The Importance of Scenarios

Domains:

- Desktop Search
- Departmental Search
- Website Search
- Hybrid Internet Search
- Multi-repository Enterprise

Functional / Industry:

- E-discovery
- Customer- / Self-service
- Executive Dashboard
- E-commerce
- Scientific / Technical / Medical (STM)
- Legal / Consulting

What your search engine probably *can't* do

1. Author content
2. Access control
3. Versioning and version control
4. Workflow
5. Localize
6. Transform into multiple formats
7. Declare a record (and all that goes with that)
8. And so on…

Recommendations and Best Practices

- Don’t think a search tool will be an enterprise panacea for information access
- Search is not one-size-fits-all!
- Plan carefully, index content incrementally
  - Each repository you add magnifies complexity with respect to security, performance, and precision
- Look for opportunities to consolidate search technologies across related applications
- But ultimately, findability is the goal, not technical consolidation
- Even within applications, search is not an excuse for poor content hygiene, disorganization, and limited classification

Agenda

- The Findability Problem
- Enterprise Content Management as a Solution
- Enterprise Search as a Solution
- **Information Architecture as a Necessity!**
- Content Technology Trends

Building Blocks

Information Access

- Browse
  - Traversing an organized repository
- Search
  - Querying information sets and obtaining documents

Information Organization

- Content Architecture
  - Structure and composition of a repository, information collection, or individual document
- Content Intelligence
  - Enriching content with additional information

Information Architecture and ECM

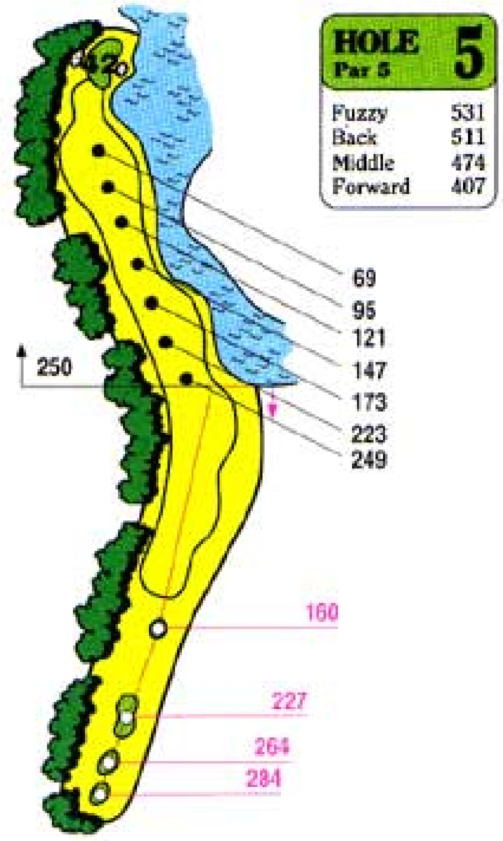
- Information organization structures often act as a “great unifier” in the area of content technologies and enable them to work together
- Many content management systems depend on solid library and categorization services order to add significant value
  - Essential for organizing any large content corpus
  - Required for meaningful records management
  - Critical to effective findability
- How you choose to design the repository, and how the system you choose can use certain repositories and content structures, greatly influence the business value you can realize

First: Know What You Have

- In order to improve findability, you need to know
  - How much content you have
  - What types of content you have, and its relative value
  - What content needs to be archived, retained, or deleted
- In order to undertake a successful ECM/WCM/RM/Search implementation or improvement effort you need to know:
  - What documents you possess before migration
  - Who “owns” the content in order to determine proper security, roles and permissions
  - Who or what creates content in order to properly tag/index and otherwise contextualize and enrich content
- Ultimately, you need to create an overall Content Model

What is a Content Model?

- Components or “elements” that make up a body of content
  - The folder or “meta”-structure of a repository or enterprise information set
  - The document types
  - Associated metadata
  - Elements within a (structured) document
- A framework applied to *content* to create relevant *information*
  - Making those related pieces useful to the people who need it



This is how you need to see and think about content

Basic Unit of Analysis

Content Type

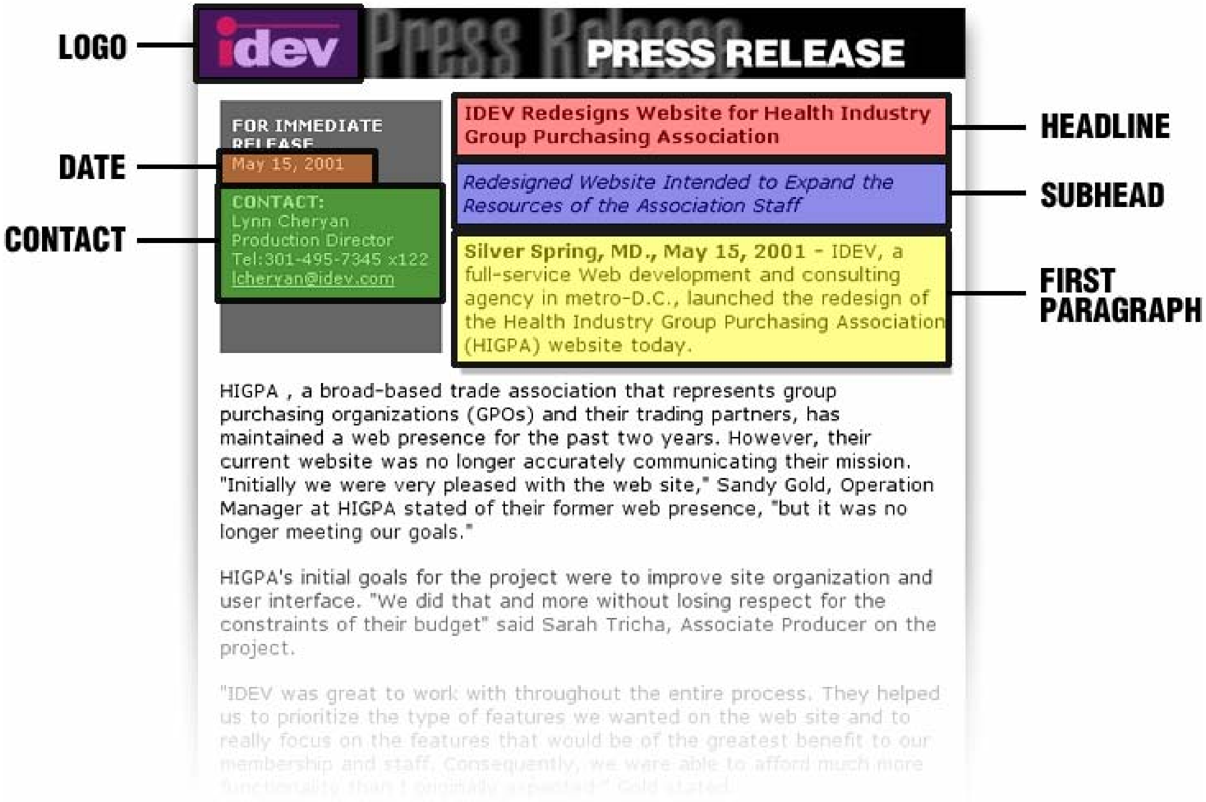
- Also called “Document Type”
- A content type defines the nature of a given piece of content
  - Press release
  - Medical record
  - Invoice
  - Product data sheet
- It can be in any format – thus it is not synonymous with *file type* or *mime type.*

Press Release as Structured Content Type

Structured content types have *elements*

Elements are:

- Named parts of a content type
- Individually stored and accessible units within a content type
- A basic unit of content



FOR IMMEDIATE RELEASE

May 15, 2001

CONTACT:
Lynn Cheryan
Production Director
Tel:301-495-7345 x122
lcheryan@idev.com

IDEV Redesigns Website for Health Industry Group Purchasing Association

Redesigned Website Intended to Expand the Resources of the Association Staff

Silver Spring, MD., May 15, 2001 - IDEV, a full-service Web development and consulting agency in metro-D.C., launched the redesign of the Health Industry Group Purchasing Association (HIGPA) website today.

HIGPA , a broad-based trade association that represents group purchasing organizations (GPOs) and their trading partners, has maintained a web presence for the past two years. However, their current website was no longer accurately communicating their mission. "Initially we were very pleased with the web site," Sandy Gold, Operation Manager at HIGPA stated of their former web presence, "but it was no longer meeting our goals."

HIGPA's initial goals for the project were to improve site organization and user interface. "We did that and more without losing respect for the constraints of their budget" said Sarah Tricha, Associate Producer on the project.

"IDEV was great to work with throughout the entire process. They helped us to prioritize the type of features we wanted on the web site and to really focus on the features that would be of the greatest benefit to our membership and staff. Consequently, we were able to afford much more functionality than I originally expected." Gold stated.

Building Blocks

Information Access

- Browse
  - Traversing an organized repository
- Search
  - Querying information sets and obtaining documents

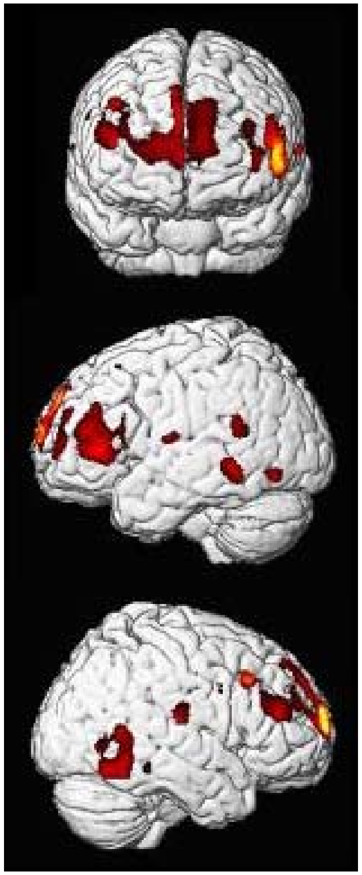
Information Organization

- Content Architecture
  - Structure and composition of a repository, information collection, or individual document
- Content Intelligence
  - Enriching content with additional information

What is Content Intelligence?

- Adding “meaning” to information by structuring, classifying, and/or labeling the content so it is more findable by both people and technology

- In short, ***enriching*** the content
  - Metadata
    - “Data about the data”
    - Usually a discreet component
  - Classification of content
  - Taxonomy
    - ***Law*** for ***categorizing*** information



Why Apply Content Intelligence?

- Taxonomy enables the broad categorization of objects – typically a tree structure of classifications for a given set of objects – in order to make them easier to retrieve and possibly sort
- The categories, sub-categories, and terms that make up a taxonomy are often employed as metadata
- Metadata can be leveraged by a system to find and display content easily and consistently
- Content Intelligence enables more precise browsing, search results, and personalization

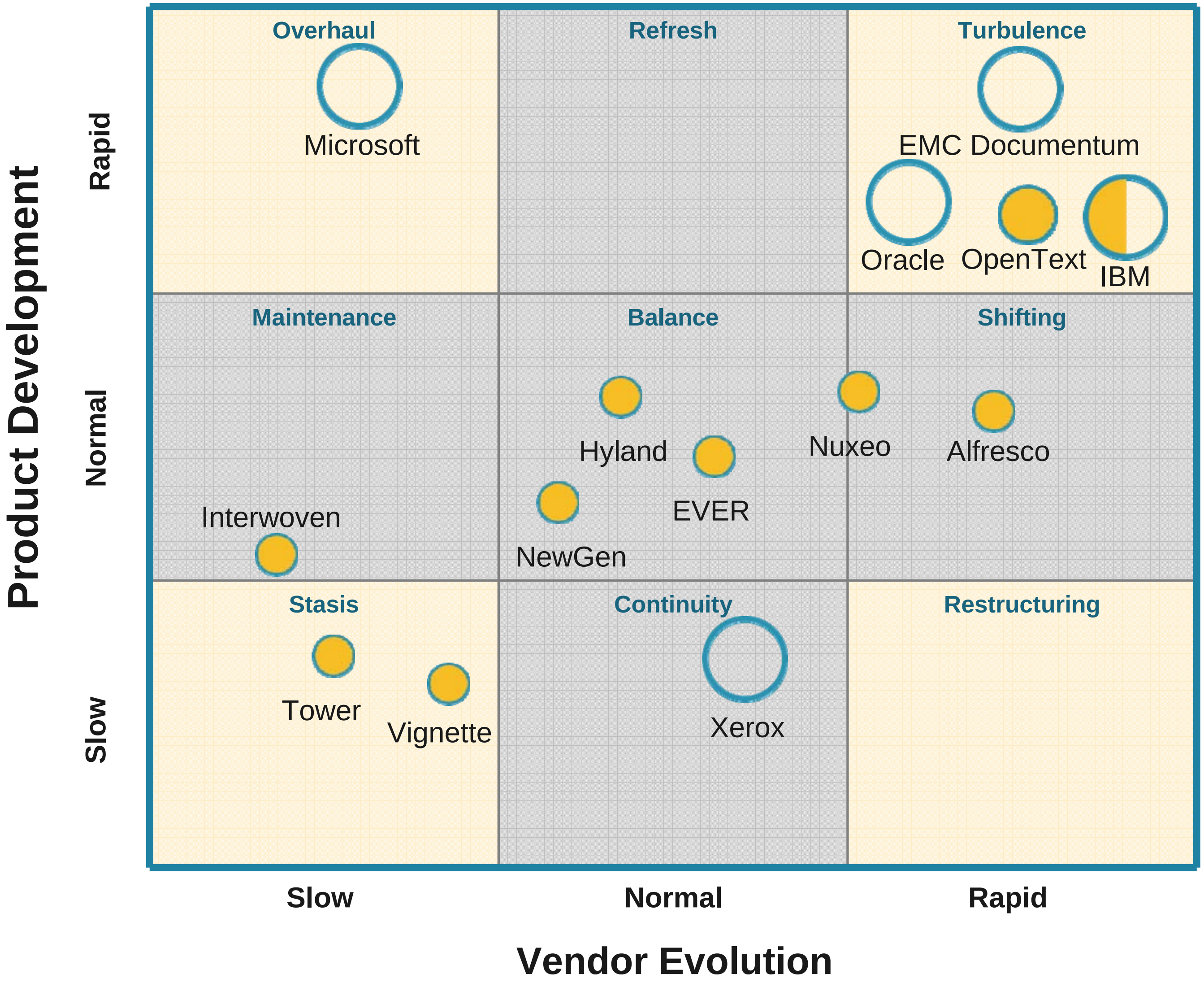
Agenda

- The Findability Problem
- Enterprise Content Management as a Solution
- Enterprise Search as a Solution
- Information Architecture as a Necessity!
- **Content Technology Trends**

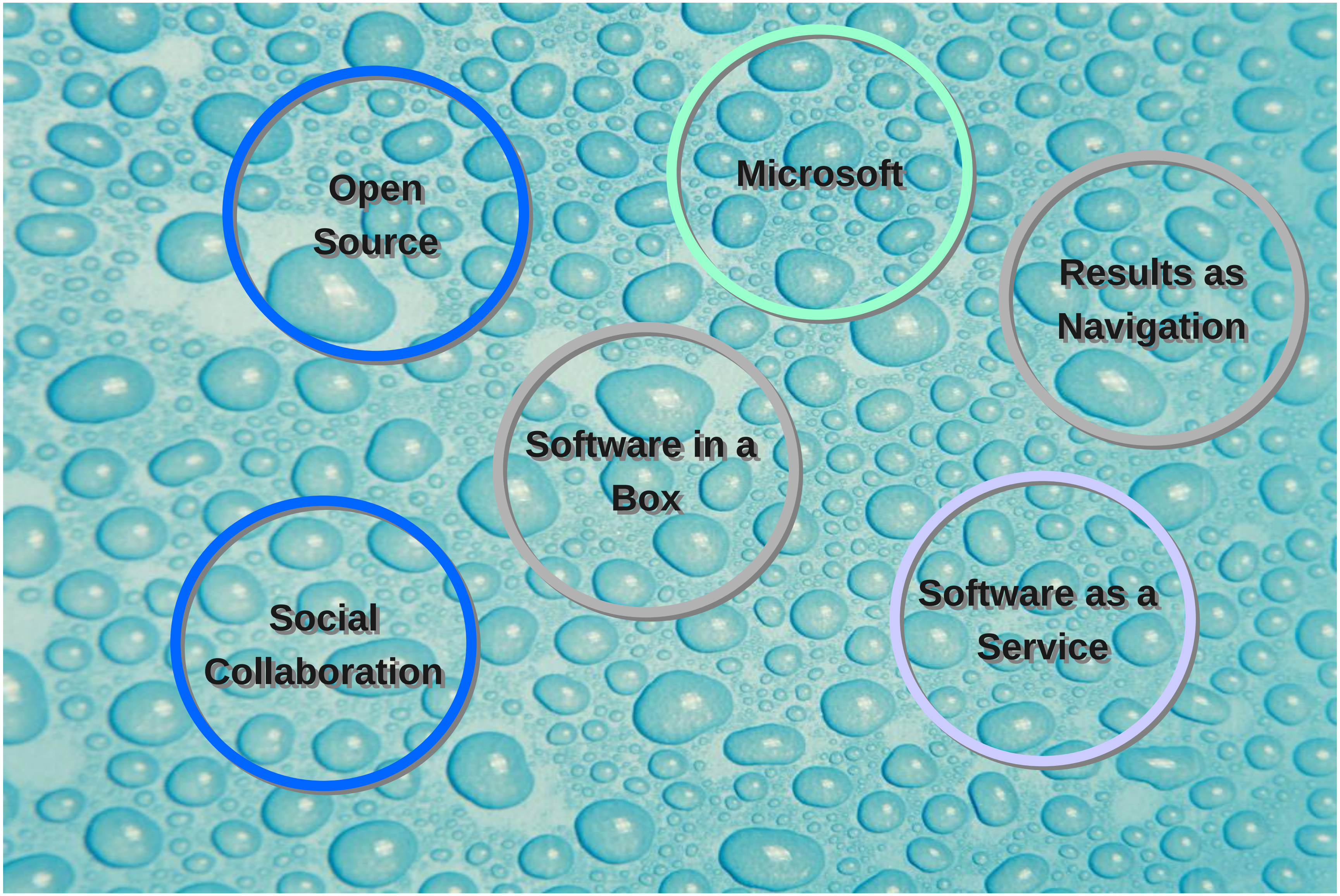
Be a smart buyer

- Many spend more spend more time buying a $20,000 car than they do selecting a $400,000 ECM or Search system
- Marketing hype obscures real capabilities
- Don’t be obsessed with technology
- Most ECM and Search systems are large, expensive and complex to implement
- Beware of a false sense of security
  - Vendor and <u>product</u> survival is not guaranteed; not even among the large vendors and products

ECM Suites: Vendor Risk Profile – H2/07



Bubbling under….



Open Source

Microsoft

Results as Navigation

Software in a Box

Social Collaboration

Software as a Service

Thank you!

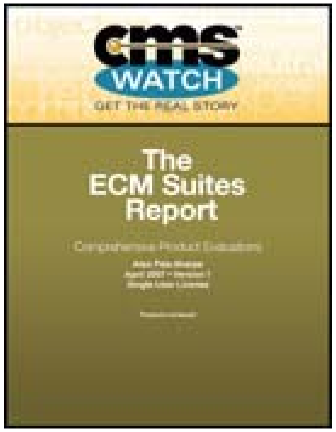
Jarrod Gingras
jgingras@cmswatch.com

www.cmswatch.com

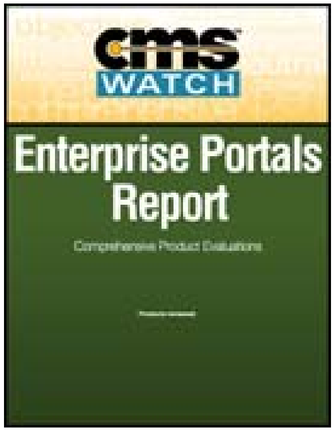
Independent Evaluation Reports:

The Web CMS Report evaluates 30 Web CMS packages

The Enterprise Search Report evaluates 18 Search vendors



The ECM Suites Report evaluates 30 products

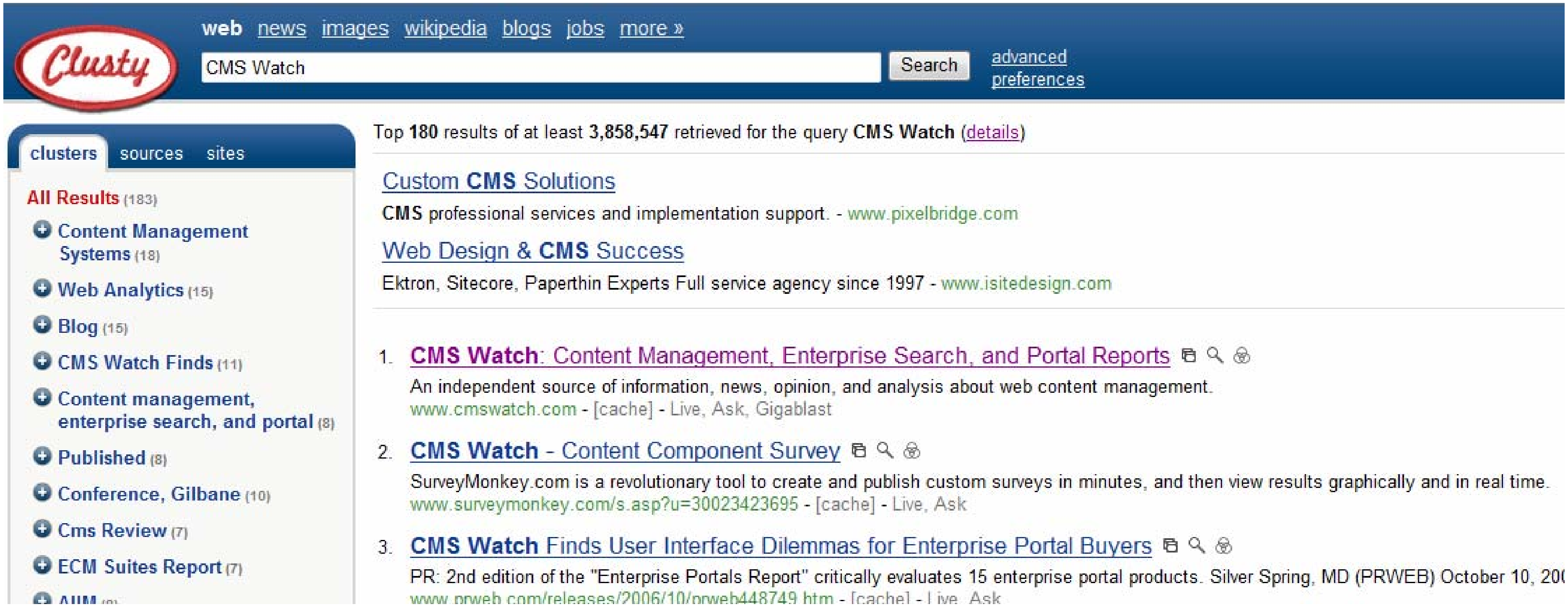


The Enterprise Portals Report evaluates 15 products

The Web Analytics Report evaluates 13 tools

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.



All Results (183)

- Content Management Systems (18)
- Web Analytics (15)
- Blog (15)
- CMS Watch Finds (11)
- Content management, enterprise search, and portal (8)
- Published (8)
- Conference, Gilbane (10)
- Cms Review (7)
- ECM Suites Report (7)
- AIIM

Top 180 results of at least 3,858,547 retrieved for the query CMS Watch (details)

Custom CMS Solutions
CMS professional services and implementation support. - www.pixelbridge.com

Web Design & CMS Success
Ektron, Sitecore, Paperthin Experts Full service agency since 1997 - www.isitedesign.com

1. CMS Watch: Content Management, Enterprise Search, and Portal Reports
An independent source of information, news, opinion, and analysis about web content management.
www.cmswatch.com - [cache] - Live, Ask, Gigablast

2. CMS Watch - Content Component Survey
SurveyMonkey.com is a revolutionary tool to create and publish custom surveys in minutes, and then view results graphically and in real time.
www.surveymonkey.com/s.asp?u=30023423695 - [cache] - Live, Ask

3. CMS Watch Finds User Interface Dilemmas for Enterprise Portal Buyers
PR: 2nd edition of the "Enterprise Portals Report" critically evaluates 15 enterprise portal products. Silver Spring, MD (PRWEB) October 10, 20
www.prweb.com/releases/2006/10/prweb448749.htm - [cache] - Live, Ask

Differentiator | Options for delivering enterprise search

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

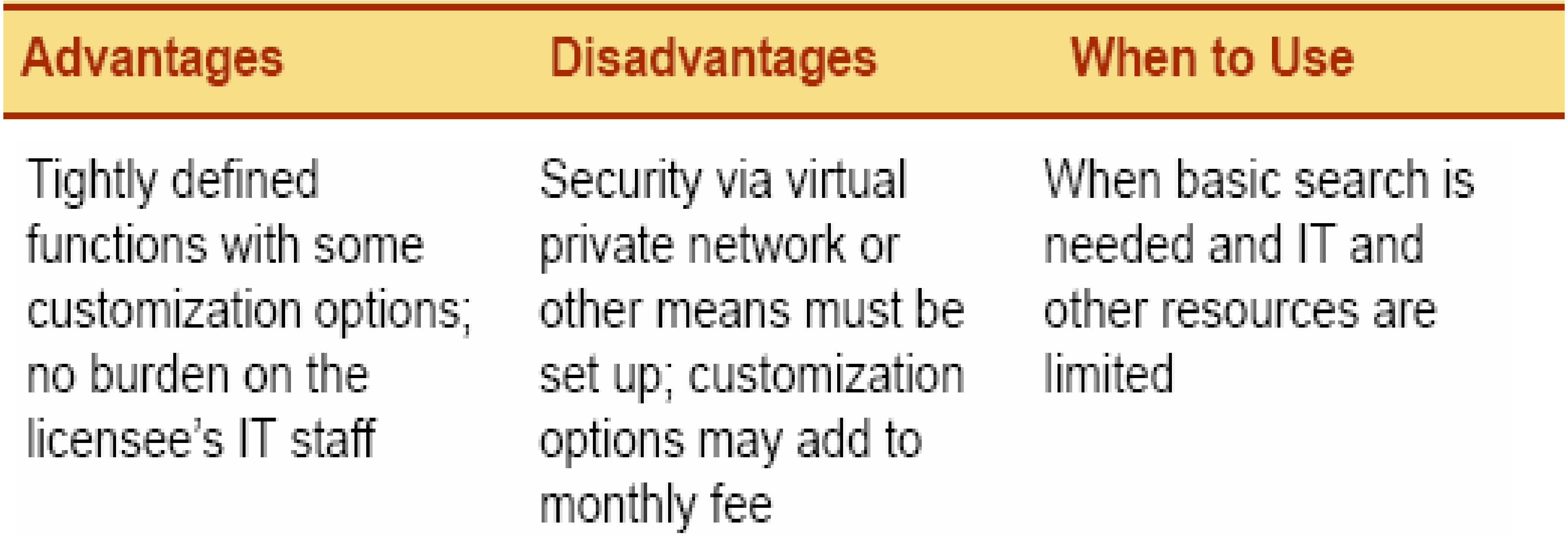
Options for delivering enterprise search | Local install

- 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”

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

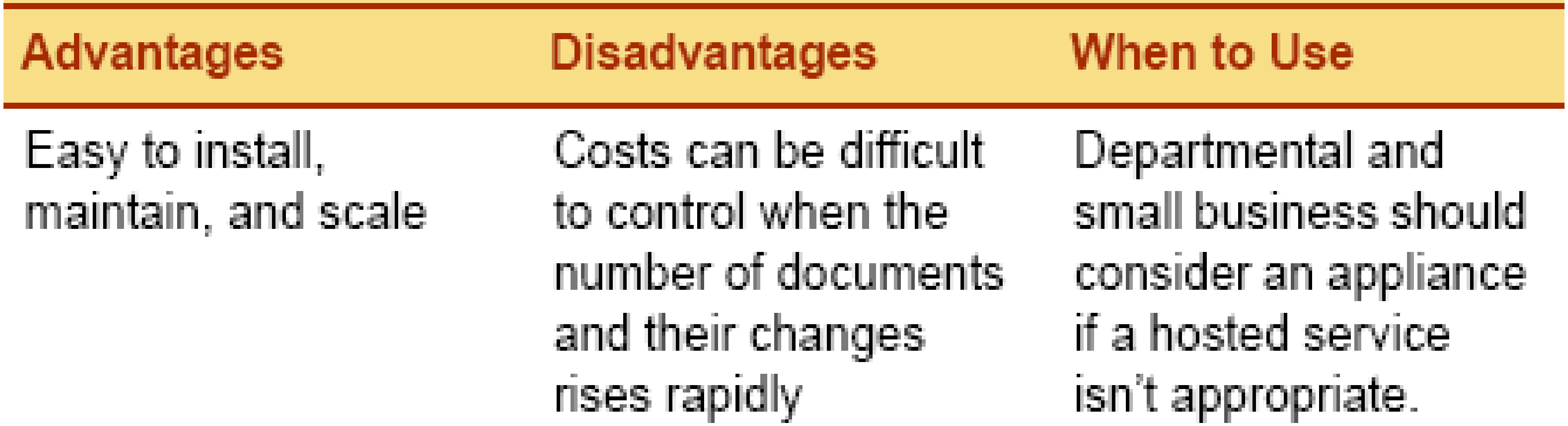


Advantages	Disadvantages	When to Use
Tightly defined functions with some customization options; no burden on the licensee's IT staff	Security via virtual private network or other means must be set up; customization options may add to monthly fee	When basic search is needed and IT and other resources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

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.