ENTERPRISE SEARCH CONSULTING

The Knowledge Worker
Waste up to 30% of their time seeking information
Only successful 50% of time finding right information

Recreating "The Wheel" Costs
For a company with 1,000 knowledge workers
that translates into $6 million spent each year
(IDC, Susan Feldman, "The High Cost of Not Finding Information")

Google Success Sets Expectations
"I want the same search experience..." (40% market share)
Average search string is 2.4 keywords
Less than 10% of searches use boolean operators

Registering Docs

SUBMIT DOCUMENT
- author information
- document information
- tags

Tagging system can be layered into any document management system (documentum, livelink, sharepoint)

concepts found directly in the article (listed by frequency)

system looks for other high frequency concepts in similar documents, but not one submitted

comma-delimited tags defined by the submitter

IN DOCUMENT:	SUGGESTED:	ADD YOUR OWN:
☒ operating dayrate	☐ deepwater fleet	
☐ existing contract	☐ contract option	
☐ drilling offshore	☐ deepwater project	
☐ appraisal drilling	☐ ultra deepwater	
☒ semisubmersible rig	☒ platform rigs	

CONCEPT LIST BUILDER
Phrases in Document · High Frequency Related Phrases · Manually Entered Tags

PRE-TAGGED
Registered Metadata · Tagged Metadata · Content To Be Indexed

Searching By Collection

Base on employee data or profiling

Mary is in investor relations (probably looking for board of directors)

Welcome, Mary — board — GO

Where SAM may be Human Resources "board" like room and board (meal)

Welcome, Sam — board — GO

options to browse other collections

Enter Search String — board — GO — Choose a collection

system allows/forces Mary to choose a collection for each search
optional: user can save default collection

Enter Search String — board — GO
automatically "suggests" the right collection based on query string content
other collections....

Requirements Phase

Reason to get it right? Switching costs...
Consumers can switch in an instant
Corporations cannot

IT INTERVIEWS

user interviews focus on scoping search needs and expectations (also to sell IT agenda/strategy)

IT sets scope, strategy and objectives for enterprise search project

User Interviews

products are "plug-n-play" solutions need a blueprint

Deliverables

ENTERPRISE SEARCH METHODOLOGY

- **"Collections" Map** — ontologies
  - can be used to build detailed schematics for data mapping
  - note: collections do not normally match organization structure (closer alignment is go-to-market)
  - study of the categories of things that exist or may exist in some domain
- **Business Rule Recommendations**
  - can be used to program business rules that drive the "what and how" results are presented to user
  - can be used to set spiders and frequency of updates
- **Document Retention Policy**
  - manage the duration of time content is kept
  - manage the final disposition of content
- **Technical Infrastructure**
  - analyze shared infrastructure
  - analyze application layer
- **Enterprise Search Blueprint**
  - can be used for roadshow (presentations to employee groups)
  - can be used to drive search and vendor selection process
  - can be used to build detailed programming specs/blueprints
- **Presentation Deck (detailed findings)**
  - can be used to secure project funding and sponsor endorsement
- **INFORMATION ARCHITECTURE**

people
technology is an enabler, not a solution

Indexing
93% of all data is born digitally

Scalability · Breadth & Depth

QUERY SPEED · INDEXING RATE · INDEX CAPACITY

not as important in controlled corporate setting time-to-[first]-index not as sensitive in corporations

opt out certain content from index
only return content for certain users
different search views; full doc restricted
all users can search; full doc restricted
docs accessible to everyone

delete document API to (re)move documents

Lifecycle of Content: **Purge** (Permanent Deletion) · **Archive** (SAN, Tape WORM)

only recommend indexing email when creating an expert finder application

EMPLOYEE EMAIL — **Desktop?**

recommend for later phase employees will have enough "cultural" change accepting enterprise search (at this time)

Federated Search? — yes, but recommend in phase II of the roll-out

Worldwide Web? — yes, but index must be kept separate from internal company data

CONNECTORS

SECURITY CONTROLS — inherit the security controls of existing systems

very important versus consumer solutions

access control list in the index [or]
search engine checks before results are presented [or]
parameters are added to search string during query

(Synthesized) Metadata Added — need one set of attributes to describe and tag all documents

Fixed Content or Active Content Management — execute policy for how frequently content needs to be indexed

Document Management

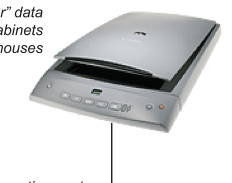
Rich Text Formats · Adobe Acrobat · Microsoft Office · Paper (Scanned Documents) · Document Management · Web Pages (HTML, Javascript, Flash) · Content Management (Interwoven, Hummingbird, etc.) · File systems · Databases (SQL Server, Oracle, etc.) · Microfilm & Microfiche · XML · Sharepoint (and other portal platforms) · Email (IMAP, POP3) · CAD Programs (PTC, AutoCAD, SolidWorks)

fixed content

one solution from one manufacturer not realistic
must meet the diverse (and changing) needs of the corporation

ability to automatically create structure from unstructured data sources is key to the success of enterprise search

unstructured data makes up approximately 85% of a corporation's data



harvest "paper" data archived in file cabinets and warehouses

information capture component

Pre-tagging Algorithms (Read and interpret content to build metadata) — annotated content — **Data Store** (PDF, JPEG, DjVu)

For large document stores:
recommend DjVu file format
10 to 20 times smaller than JPEG
3 to 8 times smaller than PDF (B&W)
note: requires UDV browser plugin to view

CLASSIFICATION / TAXONOMIES

challenge...
no hyperlinks
no anchors
no metadata
no consistency

word sense disambiguation
do not trust metadata
concept (phrase) versus single word

business rules
relevance weighting (date)
source weighting
classes of content
controls on indexing frequency

INDEX TRANSFORM & LOAD

BIASING CONTROL — date, document type, department, etc.

COLLECTION RULE MANAGER — GUI for building and maintaining defined taxonomies

Ability for admins to write and test business rules

BUSINESS RULE MANAGER

documents can belong to more than one collection

rules can "sunset" or "retire" dated content (still in index?)

technology
storage. data. organization. retrieval. decisions

Who Am I?
"shelving" information versus "filtering" information

COLLECTION · COLLECTION · COLLECTION · COLLECTION · COLLECTION · COLLECTION (dated information) · COLLECTION · COLLECTION

Collections can be referred to as "networks" (like the ABC Company Learning Network)

1 2 3 4 5 6 7 8 9 ... "relevance" rank each doc to overall collection, which can be used to improve search query results

Collections *reduce* the recall and *increase* the precision

"did you find" gives admins critical feedback

collections can be quickly tuned or split into additional collections

enterprise search system can have hundreds of collections

Recall — number of relevant records retrieved to the total number of relevant records in the database.

Precision — number of relevant records retrieved to the total number of irrelevant and relevant records retrieved.

RELEVANT ITEMS · sweet spot · ITEMS RETRIEVED
collections reduce irrelevant items returned
items I wish were in the results
items I wish were not in the results

Using Phrases

Who can use this collection?
Company filing new patent
Company suing a competitor
Lawyer defending a client
Lawyer wanting to do discovery
Lawyer wanting to sell services
etc...

Example:
What does "intellectual" mean?
What does "property" mean?
What does "property intellectual" mean?
What does "intellectual property" mean?

Should "patent" be in our concepts?
patent attorney
patent claim
patent suit
patent pending

I want to build a collection around patents:
inventing process
protect investment
market position
trademark office
legal right
patent application

Where do we get these concepts?
the company's data
u.s. patent and trademark office (ontology)

Instead, how about a rule around intellectual property:
copyright ownership
property rights
fair use
use work
exemptions provided
represented works

Search Engine
Passive Search · Web Service · Interface Design · Results Clustering

REBUILD SEARCH EXPERIENCE

discovery without user entering query (also: saved searches)
users can "subscribe" or follow new documents added to collections
can be email enabled
no complex APIs
maximum flexibility
prefer XML interface
do not recommend especially "auto" clustering
can use "collections" to drive clustered results

note: no "creative" visualization is needed or considered an advantage at this time

recommend placing search bars throughout the organization on intranet and multiple web sites present results "google-like"

consider spell checking feature in search engine (used to it in google)

<search string> GO

log search activity
log "null" results
log queries with 1+ click

Reporting:
reports on how users interact with system

can users find the search box
indexing the right information
protected information not searchable
enhancements to improve query
what search terms are not returning results
search terms returning too many results
categorize, describe, rank results
next steps if initial results not good enough

past user interactions with the search engine automatically detect "best bet" top results preferred by majority of users.

count/analyze documents most clicked on by peers and emphasize frequently clicked

can someday create "people" clusters, displaying similar results to those who share a certain profile

Best Bets

best bets do not always appear limit to just one or two best bets

almost 30% of queries are "navigational" and can be satisfied by a best bet

best bets · search results

Universal Document Viewer — in case user does not have appropriate software

example criteria Microsoft uses to determine "editor's choice" in Live.com

Presentation Features	
Result Position of the URL in Current ranking	
QueryTitleOverlap	Fraction of query terms appearing in result Title
QueryURLOverlap	Fraction of query terms appearing in the URL
QuerySummaryOverlap	Fraction of query terms in result summary
Clickthrough and Browsing features	
TimeToClick	Seconds between query and click on the URL
TimeToFirstClick	Seconds between query and click on any result
ClickFrequency	Fraction of clicks for this query on this URL
ClickDeviation	Deviation from expected click frequency
IsNextClicked	1 if there is a click below, 0 otherwise
IsPreviousClicked	1 if there is a click above, 0 otherwise
IsClickAbove	1 if there is a click above, 0 otherwise
IsClickBelow	1 if there is click below, 0 otherwise
TimeOnPage	Result page dwell time
CumulativeTimeOnPage	Cumulative time for all subsequent pages visited

example

note: this is poor interface design user has to "ask" for information rather than explore for an answer

unfortunately (thanks to google) keyword search is the current standard

monopoly GO

process
more than what did you say (dictation) -- what did you mean (understanding)

What did I mean?
When 2.5 search words are used, only 23 to 30 percent of the first 20 documents returned are actually relevant to the query.
[U.S. National Institute of Standards and Technology Study]

works much better if search engine can extract "dynamic summarizations" from checked documents and add results to search string

add terms to "monopoly" search — economics sole provider etc... — "refine results"

search for monopoly returns both the game (7 results) and antitrust (3 results)

user can check results and "refine" search string for better results

all 10 results on antitrust!

collects concepts and adds to search string

- ☐ Monopoly.com - The Official Site of Monopoly by Parker Brothers
- ☐ Play Free & Fun Online Games – Card, Strategy, Word & More
- ☒ Monopoly - Wikipedia, the free encyclopedia
  - ☒ Monopoly (game) - Wikipedia, the free encyclopedia
- ☐ Monopoly
- ☐ Probabilities in the Game of Monopoly®
- ☐ History of the Monopoly Game
  - ☐ Monopoly.com - Treasure Chest
- ☐ Amazon.com: Monopoly: Toys & Games
- ☒ U.S. v. Microsoft: Court's Findings of Fact

REFINE SEARCH

- Economics Lecture 1
- What You Need to Know About Monopolies and Monopoly Power
- Monopoly - Potential Benefits from Monopoly
  - GCSE Economics - Firms and Monopoly
- Economics Interactive Tutorial: Monopoly Price and Output
- Monopoly - Wikipedia, the free encyclopedia
  - Category:Monopoly (economics) - Wikipedia, the free encyclopedia
- Natural Monopoly of Economics
- Monopoly, by George J. Stigler: The Concise Encyclopedia of ...
- Monopoly (economics) - MSN Encarta

note: clicking on Google "similar pages" only reveals two related pages to the wiki entry for search string, monopoly

Data Forensics

Company data must be searchable and reproducible as legal evidence

Courts and regulatory bodies can impose stiff penalties for not producing "electronic-based" evidence

(Example) Securities and Exchange Commission fined Morgan Stanley $15 million when it was unable to produce email evidence in court

90% of US firms are involved in some form of legal action
$1B companies handle an average of 140 items at one time

Establish data retention & deletion policy
- Records management connection to legal department
- Smooth legal discovery process
- Where is the data in question?
- What actions were taken to preserve it?

Establish audit trails of usage
- Document "metadata" now expected as a matter of course
- Who created? When created? Who accessed? Who changed it?
- Verifiable chain of custody

data backup ≠ data archiving

backup is designed for points of failure and to reconstruct a system from a specific point in time

archiving determines what needs to be saved and why it needs to be saved

How will courts define good faith operation of an electronic information system?

Revisions to the U.S. Federal Rules of Civil Procedure (12/1/2006)
lawyers must produce electronically stored information in the event of litigation
must produce them as part of the pretrial process

rules were expanded to include more than "documents" (Rule 34)
litigators can request "electronic" versions of the data vs. paper records (Rule 33)

Litigators must discuss discovery of electronic data no later than the planning conference (Rule 16)

SIMILAR CONCEPTS
Please choose one or more search concepts below to improve your search results:

☒ operating dayrate	☐ deepwater fleet
☐ existing contract	☐ contract option
☐ drilling offshore	☐ deepwater project
☐ appraisal drilling	☐ ultra deepwater
☒ semisubmersible rig	☒ platform rigs

No thanks. Just take me to my results. · Improved Company Search

Browse By Collection

http://intranet.abccompany.com/search.aspx — Saved Searches

ABC SEARCH — COMPANY-WIDE SEARCH
All Content · Channel A · Channel B · Channel C · Channel D · More >> · Advanced Search
Company Search · Guide Me!
Search: ◉ Company-wide ◯ US Firm Only
Click here for more information about our DEF merger

while desirable to have search on all pages in the company suggest a "search site" also

users can search a channel from the search page most popular listed first with others in the more>> area

user can save searches on any page saves it to a cookie on their desktop user access "saved" searches from home page

search by geographic location defaults to all locations

very important! area for management to feature most important current event at company

left side of navigator always populated

click on channel and center populates

CHANNEL NAVIGATOR

My Personal View		
Operations	Roman Sea	Documents
Exploration & Production	Ocean Spirit	Regulations
Semisubmersible	Deep Sea 2	In-Country Info
Jack-up Rigs	Trans Admiral	Purchasing
Africa Deepwater	Raven 1	Reports
AsiaPacific	Ocean Nautilus	Headlines
Dry Dock	Wilmington	Platts Research
Purchasing	Millennium 2	CAD Drawings

based on apple finder concept will use scroll bars as content grows

Search Results

http://intranet.abccompany.com/searchresults.aspx — Save This Search

brings up channel navigator in center of page

suggested channels · best bets · search results listing

user can save any search

match search term to most clicked results for best bets

results can be customized to each client specification

based on search results can recommend channels

Channel Viewer

http://intranet.abccompany.com/channel/headlines.aspx

search within channel results · most popular documents

"passive search" users can monitor new additions to channel · most recent additions · Ocean Nautilus Headlines

exactly like best bets but restricted to channel content

content for channel note: can be documents, headlines, etc...

users can sign up to receive periodic emails notifying them of new content in channel

NOTIFICATIONS FOR: Jane Doe
user profile information · my current notifications · alert content

all alerts are combined into one email
user settings like email address, frequency, etc...
shows user exactly what the email will look like
list of subscribed to channels with options to add delete

Personal View

http://intranet.abccompany.com/mypersonalview.aspx

designed to be combined with corporate intranet (can be used as stand alone app)

controls to change layout and add additional channels — <Add Channels> <Change Layout> <Change Colors>

if standalone, intranet content or navigation controls can be added to personal view page

gadget · gadget · gadget · gadget · channel · channel · channel · channel

user can edit each module

modules can be organized using various templates (limited to a single page)

DSGN David Gossett	EDM LOGIC — DIAGNOSTIC	PRIVACY
PROG · DATA · WEB · HRDW · APPR	ENTERPRISE SEARCH CONSULTING	
DATE 08-April-2009	DESC Blueprint that demonstrates the need for a diagnostic "needs analysis" phase before enterprise search is considered. Shows in general terms the benefits enterprise search brings to an organization and the right way to implement a search engine.	
COMP VERSION 10.0	SIZE 36/66 · CODE EDM SEA · DRAWING NO EDM.SEA.2.WEB · ARCH David Gossett	
DEVELOPMENT PURPOSES ONLY	SCALE · SPEC Blacklines · SHEET 1 OF 1	

PRIVACY
All information provided with this presentation is protected by United States copyright law and may not be reproduced, distributed, displayed, or published without prior written permission. All other trademarks, product names, and company names or logos cited herein are the property of their respective owners. Do not alter or remove any trademark, copyright, or other notice from this or copies of the content.