ECM Practitioner Course Outline

Foundations
1. Introduction
2. Technologies & Functionality
3. Information Architecture

Tools & Instruments
4. Create & Capture
5. Metadata
6. Taxonomy
7. Security & Control
8. Process & Automation
9. Findability
10. Delivery & Presentation
11. Trends & Directions

Futures

Agenda
- Defining findability
- Search overview
- Basic search techniques
- Presentation and interfaces
- Findability and security
Defining findability

- Findability is the art and science of making content easy to find
- Findability moves beyond simple search to include elements of browsing and discovery

Source: AIIM Market IQ on Findability (2008)
Browsing

- User interface-oriented
- Dependent on metadata and/or taxonomies

Effective browsing

- Dependent on
  - Structure
  - Labeling
  - Location of the content
- “Virtual folders” represent different classifications
  - Allows for multiple paths to the same content
  - Ideally content should be cross-referenced, but not duplicated
Search

- A tool for finding information via user specified terms and keywords
- Advanced display techniques can blur the line between search and browse
- Search is not a magic bullet or effective panacea for lack of information organisation
  - Better-organised information will yield more effective search results

Agenda

- Defining findability
- Search overview
  - Basic search techniques
  - Presentation and interfaces
  - Findability and security
Overview of search functionality

- Allows an administrator to identify specific content to be indexed, searched, and displayed to authorised users

Core features:
- Automatic Indexing
- Query processing
- Presentation of results

Approaches to search implementation

- Application search
  - Search provided within a single software application, designed to provide localised search services to application users
  - E.g., e-mail, desktop

- Enterprise search
  - Systems intended for use within an organisation by employees seeking information held internally by the organisation in a variety of formats and locations, including databases, document management systems, and other repositories
  - Homogeneous search engine, heterogeneous repositories
  - Federated
  - Universal
Application search

- Many applications come with Search embedded in the software

- Searches only the information managed by that particular tool (e.g., Outlook), repository (e.g., a Records Management System), or system (e.g., Desktop Search)

Value of application search

- Engrains search inside the user’s daily work environment (i.e., the business application)
- Potentially leads to “actionable” content
- Typically understands and respects the application’s security and access model
Drawbacks to application search

- Search subsystem may not be as feature-rich as best-of-breed alternatives
- Search results are typically limited to a single repository, and may not encompass information across a complete business process

Enterprise search

- The enterprise is not a monolith
  - Multiple information repositories
  - Multiple search engines (i.e., application search)
- Need to search across information domains from a single query interface
- Three approaches to enterprise search
  - Homogeneous search engine, heterogeneous repositories
  - Federated
  - Universal
Federated search example

- Example: “Merlot”
  - Federated engine for education resources on the public web (www.merlot.org)

Challenges of enterprise search

- More difficult to implement than application-based search
- Needs to resolve the intersection of multiple access controls with multiple indices
- Different index and query approaches across search systems may skew results
- Potential performance problems
  - Results must be transferred, de-duped, merged, and ranked
Agenda

- Defining findability
- Search overview
- Basic search techniques
- Presentation and interfaces
- Findability and security

Basic search techniques

- All search techniques fall into two basic approaches
  - Rules-based
  - Statistical
- Many tools combine both
Rules-based search

- Based on a pre-defined set of parameters
  - Queries and results based on those rules
- Predictable and controllable
- Speed

Statistical search

- Utilise mathematical algorithms to create an understanding or profile of the content and the user queries
  - Query profiles matched to content profiles for similarity
- Many different algorithms are used, typically proprietary and not modifiable
- Algorithms can be applied in a variety of settings, including different languages and content types

\[ p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\sum_{\theta'} p(x|\theta', p(\theta'))} \]
Parametric search

- Parametric search (a.k.a. “fielded search”) adheres to pre-defined attributes present within a given data source
  - For example, a search for a women’s size 8 red shoe with a 3 inch heel
  - The parameters in this example are gender, size, colour, and heel size
  - The parameter values (populated by predefined vocabularies) are women’s, 8, red, and 3 inches
- High precision, limited flexibility

Keyword search

- A form of parametric search based on one or more fields, containing user-declared words or phrases that represent concepts within the content
- Keywords may or may not physically be within the content
  - Can be applied to any type of content
- Requires human indexing
- Inflexible
Full-text search

- Based on an automatically generated index of all words in a corpus of information, alphabetically arranged with pointers back to where each word can be found in the corpus
- Query for the word “lettuce” returns all documents that have the word “lettuce” within them
  - Would not retrieve documents that contain the words “salad,” “romaine,” or “radicchio” as alternatives to lettuce
- Various levels of granularity available in the index
- Flexible, but also inflexible

Concept and fuzzy search

- Provides retrieval based on broad appreciation for word meanings, semantic relationships
  - E.g. a query for the word “fast” automatically locates documents containing related concepts such as “quick,” “speedy” and “rapid”
  - E.g. walk = walks, walked, walking, walker
- Concept clustering provides holistic analysis of a document and indexes it at a concept or topic level
Multilingual search

- Can be critical for multinational organisations
- Results can be targeted based on user location or authentication
- Involves various capabilities
  - Interface
  - Search terms
  - Stopwords

Agenda

- Defining findability
- Search overview
- Basic search techniques
- Presentation and interfaces
- Findability and security
Navigation means moving through a corpus of information by traversing directories or following hypertext paths (links)

“The Scent of Information” in the navigation paradigm

- When a user wants to find something, they are on a hunt
- They’ll be most successful when they pick up a strong “scent”
- Navigation design must enhance your information’s scent
- Label nomenclature and design are paramount

Source: Jared Spool (UIE)
Five facets: Wine.com

Tag clouds
A clustering example

Relevancy ranking

- "3,897 results" - which are the best?
  - Date
  - (Weighted) Term summing
  - Omni-term skewing
  - Term density
  - Term proximity
  - Popularity/voting
  - Best Bets
  - Connection tracking

Source: cwi.nl & Inxight

© AIIM | All rights reserved
Agents

- User defined omni-present queries
- Run in background
- Provide immediate notification

Agenda

- Defining findability
- Search overview
- Basic search techniques
- Presentation and interfaces
- Findability and security
Levels of security

- Repository
- Cabinets
- Workspaces
  - Folder Groups
  - Folders
  - Subfolders
    - Documents
    - Elements

Securing findability-specific content

- Securing access controls within a search environment should extend beyond the content itself
  - Index
  - Queries
Enhancing security through findability

- Data Loss/Leak Prevention (DLP)

Summarising enterprise search

How Enterprise Search Subsystems Work Together

1. Content Indexing: 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 automation.
3. Query Processing: Query Parser - Accepts search queries and encodes them for optimization by that system.
4. Query Engine: Places query over index and finds documents matching search criteria.
5. Post-Process: Sorts documents, e.g., by relevance, and applies other logic to the results, such as categorization, clustering, and recommended 'best bets.'
6. Format: Streams out and formats results, usually within some sort of template.

Source: CMS Watch
What you have learned

- Findability is a critical component to an ECM strategy
- Findability is comprised of multiple technologies and techniques
- Wide variety of interfaces available
- Present challenges and enhancements to security