Why You Need Intelligent Metadata and Auto-classification in Records Management

Graham Simms
Director of Delivery and Consulting Services
Concept Searching
grahams@conceptsearching.com

www.conceptsearching.com
marketing@conceptsearching.com
Twitter @conceptsearch
Graham Simms – Director of Delivery and Consulting Services at Concept Searching has over 20 years’ experience in the IT industry, and received his Bachelor’s degree in Computer Science from the Victoria University of Manchester, where he studied under Professor Steven Furber, developer of the ARM processor. He is an expert in information retrieval and automated document classification solutions, and has developed taxonomies and classification systems for several companies, including AT&T, BP, DAI, and the US Air Force.
Agenda

- Who we are and what we do
- Why we all agree that metadata is key to records management
- Why manual metadata doesn’t work
- Auto-classification as a solution
- Pros and cons of auto-classification types
- Case studies
- Conclusions
The Global Leader in Managed Metadata Solutions

- Company founded in 2002
  - Product launched in 2003
  - Focus on management of structured and unstructured information
  - Profitable, debt free

- Technology Platform
  - Delivered as a web service
  - Automatic concept identification, content tagging, auto-classification, taxonomy management
  - Only statistical vendor that can extract conceptual metadata

- 9 years KMWorld ‘100 Companies that Matter in Knowledge Management’
  9 years KMWorld ‘Trend Setting Product’

- Authority to Operate enterprise wide US Air Force, NETCON US Army, and Canadian SLSA

- Client base: Fortune 500/1000 organizations in Healthcare, Financial Services, Manufacturing, Energy, Professional Services, Pharmaceutical, Public sector and DoD

- Microsoft Gold Certification in Application Development

- Member of SharePoint PAC and TAP programs

- Suitable for all versions of SharePoint on-premises and SharePoint Online, including the vNext dedicated platform and the government cloud
Why Are We Different?

It’s all about metadata

- Unique IP **compound term processing**
- Identifies multi-word terms that form a complex entity
- Ambiguity inherent in single words is eliminated
- Works in any language, regardless of grammar or linguistic style
- Generates non-subjective metadata based on an understanding of conceptual meaning
These valuable documents should be stored for five years.

This job got so much easier when I realized that nobody ever asks for anything back.
The Changing Role of Records Management
Innovation or Disruption?

• Vocabulary and assumptions based on paper fail to scale
• Confidence in records management, eDiscovery, security, privacy, and compliance remains low
• Records managers must develop new lifecycle roadmaps to include on-premises, cloud, and mobile
• The technical challenges tie the hands of records managers to make strategic decisions
• Older records management applications have not kept pace
• Preservation of older records should be a concern
• Social media may contain legal and regulatory risk, provides little to no metadata
• Email
Unable to Connect

We encountered an error while trying to connect.

Details: "OData: The feed's metadata document appears to be invalid."

Retry  Edit  Cancel
What Does Metadata Impact?

- Content Optimization
- Intelligent Migration
- Semantic Search & Information Transparency
- Records Management
- Privacy and Sensitive Information Protection
- Secure Collaboration
- Text Mining
- eDiscovery, Litigation Support, FOIA
- Knowledge Management & Research
- Compliance
- Mergers & Acquisitions
- GDPR

conceptClassifier
How Is it Supposed to Work?
“Information which is not communicated is valueless, and information that cannot be found is similarly useless.”

Robek, Brown, and Stephens
Records Management, Fourth Edition
A manual metadata approach will fail 95%+ of the time

**Inconsistent:** Less than 50% of content is correctly indexed, meta tagged or efficiency searchable

**Subjective:** Highly trained information specialists will agree on meta tags 33% to 50% of the time.

**Expensive:** Average cost of manually tagging one item runs from $4 - $7 per document and does not factor in the accuracy of the meta tags nor the repercussions from mistagged content *(Hoovers)*

**Malicious Compliance:** End users will consistently select the first choice on a drop down list.

**No Perceived Value:** What's in it for me syndrome.
Metadata Tagging – the Problem

Ineffective Capture → Manage → Store → Preserve → Deliver

Manual Meta-tagging Problems

• Created from a subjective frame of reference
• May not be in line with corporate governance
• Limits document transparency in an ECM environment
• Repercussions from noncompliance, impacts eDiscovery, potential privacy or sensitive information exposure, degrades enterprise search
• Cost in-effective
Effective Capture → Manage → Store → Preserve → Deliver

Solution to Manual Meta-tagging Problems

- Taxonomy management – organizational file plan/folder structure
- Automatic metadata generation – produce highly relevant corporate metadata
- Automatic document meta-tagging – eliminate all manual meta-tagging costs
- Auto-classification of all documents – organize all content to organizational standard
“There is a debilitating disconnect between the proliferation of electronic information and the constant need to quickly and accurately find all of the information and expertise that is essential for work every day. From top to bottom, enterprises have failed to take seriously the high cost of being grossly inadequate at finding information, data, documents, experts. Instead they have settled for low performance, low-return techniques to… sort of handle Search.”

Julie Hunt
Search Consultant
• Hierarchical representation of entities of interest in an organization
• Primary tool to provide structure to unstructured data
• Front end and/or back end functionality
• Actualized through metadata
• Business Taxonomies
  • Tend to be less rigid and constrained
  • Usability – minimize clicks
  • Content driven
  • Allows flexibility and redundancy
• Provides a single methodology for classification (categorization)
• Provides for entity extraction using NLP
Auto-classification
Types of Classification Metadata

**Intrinsic** – information that can be extracted directly from an object (file name, size)

**Administrative/Management** – information used to manage the document (author, date created, date to be reviewed)

**Descriptive** – information that describes the object (title, subject, audience)

**Semantic** – ability to extract concepts from within content and generate the metadata (intelligent metadata)
**Content Based** – weight given to a particular subject in a document determines the class to which the document is assigned

**Request Based** – sometimes referred to as indexing, classification in which the anticipated requests from users influence how documents are classified

**Policy Based** – classification that is aimed at a particular audience or user group
Automatic Document Classification

**Supervised** – some external mechanism, such as human feedback, provides information on the correct classification

**Unsupervised** – also known as document clustering, where the classification has no reference to external information

**Semi-supervised** – where parts of the documents are labeled by an external mechanism and some by human intervention
Auto-classification Systems – Statistical

Taxonomies and thesauri are the foundation of an auto-classifier. They provide the vocabulary against which rules are built and ‘teach’ the machine how to ‘understand’ and categorize content

**Statistical**
- Often use Bayes theorem: measures ‘degrees of belief’ (or ‘degrees of aboutness’)
- Use frequency and location to determine important or useful concepts
- Feed the system example text for the specific category
- Statistically identifies and extracts significant keywords and patterns
- Document training sets
  - Match word/concept patterns to categories
  - Often need sets of 50+ documents, or more
  - Poor document choice can cause pollution/noise
- Drawbacks
  - Effort required to create the training set
  - Relies on the availability of keyword-rich text
  - Hard to determine problems
Rule-Based

- Rely on Boolean (and, or, not) categorization rules to find either a positive or negative evidence of a match to a category
- More control over behavior – More work!
- Success depends on quality of rules
  - Example: (Google OR SalesForce) NOT LinkedIn
- Drawbacks
  - Dependent on the richness of the taxonomy and collection of synonyms/keywords
  - Creating and/or tweaking the rules for each category – can be onerous

Most popular taxonomy management suites include auto-classification modules
- With few exceptions, taxonomy tools are generally rule-based systems
**Linguistic**
- No commitment to a taxonomic tree, based on parts of speech and their relationships, typically not scalable
- Related to parts of speech, syntactic parses, or semantic interpretations

**Machine Learning**
- Subfield of computer science (CS) and artificial intelligence (AI) that deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions

**Semantic Networks**
- Refers to a set of relationships between concepts and words, including parts of speech and real-world relationships
- These can include rules of various types – not just Boolean
Pros and Cons of Most Widely Used Classification Techniques

- Most widely used are statistical and rule-based
- Several are a combination of both statistical and rule-based

<table>
<thead>
<tr>
<th></th>
<th>Statistical</th>
<th>Rule-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work involved in building good training sets</td>
<td>Work involved in building exhaustive rules (mitigated by taxonomy tools)</td>
<td>If there’s a problem, go back to the rule set and tweak</td>
</tr>
<tr>
<td>If there’s a problem, can be difficult to diagnose and rectify/retrain</td>
<td>Machine learning can augment accuracy or lead to pollution (accuracy can wax and wane)</td>
<td>System doesn’t evolve without new rules, but high degree of control (accuracy mostly increases)</td>
</tr>
</tbody>
</table>
Auto-classification Systems – What Do They Do?

**Document Preparation**
- Split into language blocks (paragraphs, headings), formatting, layout

**Parsing**
- Entity extraction
- NLP: parts of speech, phrases
- Terms, variants

**Weighting**
- Frequency
- Location in text, phrase
- Proximity
- Combination
- Format of text

**Classification**
- If threshold reached
- Can influence search results

Not all classification solutions are created equal!

This is where rules vs statistics come into play…
Concept Searching has a unique approach to ensure success

- Concept Searching’s unique statistical concept identification underpins all technologies
- Multi-word suggestion is explicitly more valuable than single term suggestion algorithms

Concept Searching provides automatic concept term extraction

- conceptClassifier will generate conceptual metadata by extracting multi-word terms that identify ‘triple heart bypass’ as a concept as opposed to single keywords
- Metadata can be used by any search engine index or any application/process that uses metadata
Unique to concept **TaxonomyManager**
- **Compound term processing** technology that identifies ‘concepts in context’
- Automatic intelligent metadata generation as content is created or ingested
- Rule-based engine that eliminates the need for training sets and highly specialized human resources
- Automatic taxonomy node clue suggestion
- Dynamic screen updating to immediately see impact of changes in the taxonomy
- Document movement feedback to see cause and effect of changes without re-indexing
There are more than 14,000 laws and regulations related to information management, many of which can be challenging to enforce across an enterprise-size IT infrastructure.
When Does Records Management Become Information Governance?

- **Records Management** involves the implementation of a process or system for directing and controlling an organization's information (records).

- **Information Governance** is the strategy or framework for controlling information (records) in a way which encourages compliance, mitigates legal risks, and aligns to corporate governance policies.
  - Holistic approach to manage information in all its formats and forms, regardless of where it is stored or how it was acquired.

- **Beyond Records Management** – transitioning to information governance
  - Risk assessment
  - Legal mitigation
  - Defensible audit
  - Process

"So called ‘data breaches’ are thefts of information and, as such, they are first and foremost a traditional records management problem. Until organizations understand this and include records management as a critical component of their long-term cybersecurity strategy, data breaches – and the disastrous consequences they bring – will continue unabated.”

Don Lueders

© Concept Searching 2018
The Value of Semantic, Multi-term Metadata

Advantages

• Ability to develop a single repository of organizationally relevant metadata to be made available to any application that requires the use of metadata
• Elimination of costs and errors associated with end user tagging
• Normalization of content across functional and geographic boundaries to remove ambiguity in vocabulary
• Metadata managed and changed in one place
• Ability to apply policy consistently across diverse repositories and applications
• Provide flexibility to rapidly make changes to the repository for regulatory compliance where changes are immediately available for use by applications
• Works bisynchronously with the SharePoint Term Store, reading and writing in real time
Situation:
- Nonprofit public benefit corporation
- Highly regulated
- Relied heavily on website to address the unique requirements of diverse audiences, updated daily
- Unable to implement content lifecycle management

Challenges:
- Erroneous tagging of documents
- Poor information retrieval
- Needed to improve site visitor experience
- Management of the complexity and amount of content and data
- Leverage SharePoint investment

Products Used:
- conceptClassifier platform
- conceptClassifier for SharePoint
- conceptTaxonomyWorkflow

Achievements
- Automates document workflow for storage, preservation, access, and usage controls and eliminates end user tagging
- Assists in the management of content by identifying records as well as content that should be archived or contains sensitive information
- Facilitates the retrieval of records as well as highly correlated content that typically would not be found
- Ensures compliance with industry and government mandates enabling rapid implementation to address regulatory changes
- Native integration with SharePoint and the Term Store, maintains GUIDs
Identification of Sensitive and Privacy Information

Situation:
• US Government Military Agency

Challenges:
• Erroneous tagging of documents
• Prevent data exposures of privacy, sensitive, and confidential information

Products Used:
• conceptClassifier platform
• Hosted by Serco

Outcome:
• The solution scans both the Internet and intranet, and compiles a report of PII erroneously present on unprotected portals

Achievements
• Ability to proactively take action on the potential data breaches
• Removes from unauthorized access
• Prevents portability
• Eliminated end user tagging
• Not transmittable view email without specific authorization
• No breach for 11 years
• Increased productivity – 2,500 record codes
• Processed information faster and achieved higher accuracy
Automatic Tagging, Policy, and Governance

Situation:
• Budget of $6.9 Billion
• Over 60,000 users
• Runs 75 hospitals and clinics providing care to more than 2.6 million beneficiaries

Challenge:
• Data Privacy
• Intelligent Migration
  • Before and after
• Records Management
  • 72,000 Site Collections, 5,300 retention codes, classify 200,000 documents per hour with minimum resources (Proof of Concept)

Solution:
• conceptClassifier for SharePoint platform

Benefits:
• Automatic tagging based on organizational vocabulary and descriptors
• Automatic routing and the ability to change the SharePoint content type
• Eliminated manual tagging, removes from unauthorized access and portability
• No security exposures or breaches in 5 years, since deployed
What Are the Results?

- All information is automatically tagged resulting in the classification of unstructured data to organizational taxonomies.
- All information is retrievable using concepts (high-precision) instead of key-words, proximity, full text, (low-precision).
- Cleanses file shares, SharePoint, Exchange, and any repository.
- Identifies and protects privacy and sensitive information exposures in real time.
- GDPR compliance.
- Insight engine feeds search engine index enabling concept-based searching.
- Reduces the risks and costs associated with eDiscovery.
- Works interactively with records management applications to identify and classify records based on the file plan, automatically classifies them to a SharePoint content type, processes to records management application.
- Provides secure collaboration based on the contents within documents to be shared.
- Knowledge management, research, text mining and analytics.

© Concept Searching 2018
Thank You

Graham Simms
Director of Delivery and Consulting Services
Concept Searching
grahams@conceptsearching.com

www.conceptsearching.com
marketing@conceptsearching.com
Twitter @conceptsearch