Why You Need Metadata-Driven Records Management

Michael Paye
Chief Technology Officer
Concept Searching
mikep@conceptsearching.com

Robert Piddocke
VP Channel and Business Development
Concept Searching
robertp@conceptsearching.com

www.conceptsearching.com
marketing@conceptsearching.com
Twitter @conceptsearch
Michael Paye – Chief Technology Officer at Concept Searching has been the driving force behind many of the company’s recent innovations, including the SharePoint Add-in and hybrid search products. He has a wealth of experience across the Microsoft platform and related technologies, and oversees all product development.

Robert Piddocke – Vice President of Business and Channel Development at Concept Searching has over fifteen years of business development experience. With extensive Microsoft product knowledge and information management understanding, he has implemented hundreds of site and enterprise search solutions, and has authored two books on SharePoint Search.
Agenda

- Introduction to Concept Searching
- Introduction to records
- Managing records manually
- The future of records in a Microsoft world
- First steps to records success
- Demo
- Key takeaways
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 latest 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.
What’s in a Record?

- A record is a record, regardless of physical form or characteristics
- Medium doesn’t matter – content does
  - Blogs, wikis, tweets, Facebook, text messages, voice messages, images, video, LinkedIn, YouTube, Flikr
- Metadata can be non-existent, requiring manual application, which is time consuming and subjective
- Organizations tend to keep all social media, which is costly and inefficient and impacts eDiscovery
- Written policies on the use of social media are not prevalent
- Education and training for information being shared outside the organization is required
- The organization is legally responsible for all social media that is generated from within the organization, which may be user personal

76 percent of businesses regard communications by social media as formal business records, but 46 percent have no idea that they carry legal responsibility for their content through social media channels.

Business managers are in the dark about social media legal responsibility.
Growing Risk

Why keep records?
- Internal investigations
- Privacy regulations affecting clients’ personal and financial information
- Internal or external audits
- Litigation and eDiscovery
- Impact of GDPR
- Partner communications
  - Once it’s gone, it’s gone
- For government – mandated

Your cloud provider does not protect what is in your files – you do
Zone 1: 80%
- ‘Read and delete’ variety
- Automatically deleted after a set period of time

Zone 2: 15%
- Emails with corporate value but which have yet to be declared as records (working)
- Either they get declared as records within two years or they get deleted

Zone 3: 5%
- Emails defined as records that need to be moved into an enterprise records repository, and have some kind of policy placed on them for retention, disposition, eDiscovery
- These are email records that need to be managed properly
A records management system includes the following elements

- **Content analysis** that describes and categorizes content that can become records, provides source locations, and describes how content will move to a records management application.
- **File plan** that indicates where each kind of record should be retained, the policies that apply to them, how long they must be retained, how they should be disposed of, and who is responsible for managing them.
- **Compliance requirements document** that defines the rules an organization’s IT systems must follow to ensure compliance, and the methods used to ensure enterprise team member participation.
- **Method for collecting records** that are no longer active from all record sources, such as collaboration servers, file servers, and email systems.
- **Method for auditing records** while they are active.
- **Method for capturing records' metadata** and audit histories, for maintaining them.
- **Process for holding records** (suspending their disposition) when events such as litigations occur.
- **System for monitoring and reporting** the handling of records to ensure employees file, access, and manage them in accordance with defined policies and processes.
How Is it Supposed to Work?
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.
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
Metadata Tagging – a Solution

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
Taxonomies

• 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 natural language processing (NLP)
Auto-classification
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 Classifier 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.

Building a Records Management Concept Index – Example
Is Office 365 in Your Future?

• Complete collaboration solution
  • Chat, email, calendar, voice, video, meetings, sites, content management, enterprise social
  • Core component – Microsoft Teams

• Benefits
  • Supports different collaboration styles
  • Reduces complexity of infrastructure
  • ‘One version of the truth’
  • Consolidation of administration
SharePoint Gotchas

- File plan and taxonomy are one unit
- Two-fold problem with attempt to classify (match to record policies) content in SharePoint
  - Can’t mix content types and locations, you must chose one or the other
  - Using other data to classify, such as SharePoint columns, managed metadata or even site names
- SharePoint has weak content rules engine
- Cut off and retention
  - Without custom code, SharePoint uses only date properties (think created date, modified date or custom dates) to determine when retention should begin
  - Impossible to act on an entire file/case without building some type of workflow that would populate the date property on each item
- Disposition – SharePoint has does not have a review process to solicit approvals before allowing the appropriate disposition to take place
- Physical records – SharePoint does not have any concept of physical records
- Auditing
  - SharePoint limited to the basics of create, edit, modify, check in/out, and search
  - Significant performance implications
  - Logs saved for only 90 days
- Reporting – SharePoint limited to very basic audit information and provides it only as cryptic data in spreadsheet format
So Why Do You Need Concept Searching?

- **Compound term processing** technology that identifies ‘concepts in context’
- **Automatic intelligent metadata generation** as content is created or ingested
  - Retrieves information you didn’t know existed
  - Eliminates end user tagging
- **Rule-based engine** that eliminates the need for training sets and highly specialized human resources
- **Auto-classifies content** to one or more taxonomies
- **Content optimization**
  - Identifies records that were never declared
  - Privacy and sensitive information exposures
  - Records incorrectly declared
  - Indexes and cleans up corpus of content, diverse content stores, file shares, and social

Compliance
Data Protection
GDPR
Data Discovery and Classification
File Analytics
Text Mining and Analytics
Migration
Search and Information Transparency
eDiscovery
Collaboration
Information Governance
Let’s Muddy the Waters with Labels

1. User creates new group for collaboration
2. Group identity created in Azure Active Directory
3. Group experience populated in app of choice

Office 365 Application

Azure Active Directory
Identity, Resource URLs, Owners, Members

Office 365 Application
Office 365 Labels

- Office 365 labels are a part of the Advanced Data Governance (ADG) suite of tools
- Security and Compliance Center

Office 365 Advanced Data Governance Features

- **Labeling**: Classify data across your organization for governance, and then enforcing retention rules based on that classification.
- **Retention**: Ensures that you retain content as long as required but no longer than that.
- **Supervision**: Define policies that capture email and 3rd-party communications in your organization so they can be examined by internal or external reviewers.

Labels work well for two things

- **High-level classifications**, such as designating documents as confidential rather than public, are quite helpful – this approach can help people use the content correctly and help prevent data leaks
- We can use labels to **identify sensitive data**, such as a US social security number or an Australian tax file number
Office 365 Labels Considerations

• It takes up to **seven days** to automatically apply a label – testing has to wait seven days.
• Applying labels to search, you can’t use site or content properties. Can result in **mislabeled documents**. More accurate to use site and content properties to label content automatically.
• You cannot automatically apply labels to **sensitive data** that resides in Exchange or Groups. This automatic labeling works only with content in **SharePoint** or **OneDrive for Business**.
• You can apply **only one label** to content.
• There is no **hierarchy** available for labels, or a way to weight their importance. The label that was **created first** will be applied.
• Office 365 can only classify content located in **Office 365**. Applications such as Salesforce, Box, Dropbox, and G Suite require another records management system.
• Office 365 provides only **generic functionality** that doesn’t meet local standards.
• You can’t automatically declare content as a record. Users must **manually declare records**.
• You can only trigger **label retention** based on the date the document was created, last modified, labeled, or on an Office 365 event.

![Office 365 Won’t Auto-Apply if the Label is a Record](image)
There are more than 14,000 laws and regulations related to information management, many of which can be challenging to enforce across an enterprise-sized 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 with 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
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 keywords, proximity, full text (low-precision).
- Cleanses file shares, SharePoint, Exchange, and any repository.
- Identifies and protects privacy and sensitive information 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, and processes to a records management application.
- Provides secure collaboration based on the content within documents to be shared.
- Knowledge management, research, text mining and analytics.
Thank You

Michael Paye
Chief Technology Officer
Concept Searching
mikep@conceptsearching.com

Robert Piddocke
VP Channel and Business Development
Concept Searching
robertp@conceptsearching.com

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