



# UTILIZING COMPOUND TERM PROCESSING TO ADDRESS RECORDS MANAGEMENT CHALLENGES

CONCEPT SEARCHING

This document discusses some of the inherent challenges in implementing and maintaining a sound records management process through the use of compound term metadata and ways to enforce governance at the desktop.



## Introduction

Although government, legal, and healthcare entities have been required to develop a comprehensive records management discipline, the general management of corporate records across diverse industries has traditionally been poorly implemented. Records management was often viewed as a low priority administrative task however new statutes, scandals, and litigation preparedness are increasingly becoming issues that impact the boardroom. Managing enterprise risk and ensuring compliance is becoming a high priority for publicly traded companies who face more regulatory initiatives and increased public analysis.

Records Management must be a corporate initiative with stakeholders at all levels of the organization. Companies who adopt the attitude to keep everything and not address the management of non-record as well as record materials can also incur a potentially devastating impact on the enterprise. In addition, enterprises may still be held accountable for non-record information with catastrophic consequences.

A successful records management system must suit the organization's workflow, is easily adaptable by users and can be integrated into their daily activities, ideally transparently. From a management perspective the system must achieve the records management goals of the organization and effectively integrate governance, risk and compliance into a single corporate wide objective.

Concept Searching's products can be implemented along with a Records Management or an Enterprise Content Management solution to facilitate the implementation and on-going management by providing true concept search, automatic compound term metadata generation, automatic classification, and taxonomy management.

## The Challenges

Records Management represents an explicit corporate memory for the organization. Electronic records unlock the content, provide content control, enable more effective sharing of information and contribute to knowledge network flows. These solutions support evidence-based policy making by providing reliable evidence of past actions and decisions; but to achieve this, the content must be managed to retain integrity and authenticity. A key consideration in planning for a Records Management implementation is the ability to accommodate current and future records keeping needs including content types, media types, storage requirements, business processes and policies. Meeting current and future legal and audit requirements must also be taken into consideration.

Perhaps an overlooked consideration of a Records Management solution is to enhance the retrieval of information within the organization. Do the records need to be consistently searchable, are the document categories defined for easy retrieval, and how will metadata be used to enable relevant retrieval? Although it seems intuitively obvious, it is remarkable that organizations generate prodigious amounts of content without having a plan for access and reuse. Addressing these issues ultimately enables employees to do their jobs more efficiently and effectively and is a critical component for the successful implementation of records management solution.



## Whose Job Is It?

Many organizations that have implemented a records management solution have identified end user adoption as a key barrier to success. Difficult interfaces, complex processes, and disparate tools make it hard for users to embrace the records management process. End users are often reluctant to change their work habits which can pose a challenge in consistently maintaining accurate records. As a result, too many documents are never subjected to enterprise policies, resulting in widespread noncompliance issues. Automating the records management processes to facilitate end user acceptance removes much of the burden from the staff. Although a records management solution can provide the method for document retention, workflow, and maintenance the intellectual control of the end user is ultimately responsible for the assignment of documents with the appropriate descriptors, and are burdened with the task of providing sufficient information that provides corporate meaning to the entity.

Many Records Management solutions are now using Federated Records Management to automatically declare records simply by knowing the location of the content and what parameters make content eligible to become a record. Although this approach is effective in ensuring records are generated without user intervention, the drawback of course is that there is little ability to determine the content. The crux of the issue is that the content is still unmanaged regardless of its record status. Although a document may be declared a record, the content or essence of the record is still not captured. This approach does overcome issues with the governance of records management and end user adoption as the classification can take place without user intervention. However this poses other issues when the organization needs to find and use the information with system assigned descriptors and no information regarding the content. With the burgeoning amount of content organizations need to access relevant information quickly and reduce the unproductive time spent trying to find the record when needed.

## The Use of Metadata

Although not a new concept, metadata is a critical element of records management. In Records Management metadata has the primary purpose to define data so the system can properly apply rules for disposition. The basic elements of metadata are a structured format and a controlled vocabulary. Metadata in information technology as a whole is fundamentally important and in records management it can define record keeping and the records retention schedule. Most software applications automatically create metadata and associate it with files. For example Microsoft Word will automatically create metadata such as title, author, file size and provides the option for the user to add metadata. The challenge with metadata is making it more useful to the end user and to the organization. For search and retrieval the more descriptive metadata that can be generated not only reduces operational costs and improves productivity but enables the organization to react more quickly and make better informed decisions.

## Classification & Taxonomies

The automatic classification of new and existing content from potentially diverse repositories can assist with records management. Ideally, this capability involves a rules based tool that can understand language and the context of those records not residing in the Records Management repository and categorize them accurately without user intervention and ingest the content so retention policies can be assigned.



A taxonomy provides the capabilities to automatically identify, analyze, and classify unmanaged corporate content and build the hierarchical structure to classify that content to the appropriate taxonomy node(s). In the records management planning process, a taxonomy can identify the content, classify it, and provide insight to the organization in defining retention policies.

According to Gartner over 70% of organizations have six or more content repositories – the ability of Record Management products to extend their reach through federation with other content management repositories, archives and applications as well as supporting compliance and discovery is essential. The taxonomy can create virtual centralization of content across disparate repositories and content management systems and can classify content from these repositories to the corporate taxonomy. An added advantage is the ability to identify the relationships between content, regardless of where it is stored that may require the same retention policies.

The management of metadata through the taxonomy establishes processes on the collection and control of metadata. The use of metadata in a corporate taxonomy results in the extraction of business value from content that can be optimized so that content becomes an asset to be managed for competitive advantage.

## The Concept Searching Approach

Concept Searching is the only statistical semantic metadata generation and classification software company in the world that uses concept extraction and compound term processing to significantly improve management of unstructured information. The tool set provides advanced search, compound term metadata generation, automatic classification and taxonomy management.

The technologies provide the framework to enable organizations to facilitate the implementation and management of their compliance initiatives. The software has the ability to automatically identify new as well as existing unstructured content and generate compound term metadata (concepts) within the content and classify the documents to a taxonomy or multiple taxonomies. The compound term metadata can then be used by the Records Management solution to further define the record, provide insight on duplicate information, and evaluate the content to determine if it should be a record.

## Compound Term Processing

The generation of metadata based on concepts means that compound terms as well as keywords are extracted from a document or corpus of documents that are highly correlated to a particular concept. By identifying the most significant patterns in any text, these compound terms can then be used to generate non-subjective metadata based on an understanding of conceptual meaning. Compound term metadata can be automatically generated either when the content is created or ingested. The metadata can be stored and used to provide additional descriptors in the records repository.

Metadata ultimately must be understandable by end users. So although metadata is machine processable and can be used in a variety of application services, if a person unfamiliar with the resource cannot understand what the metadata record describes, the record becomes unusable. The unique process to generate compound term metadata enables information to be retrieved that is useful and usable and supplements the system generated metadata. This delivers “the ability to perform search over diverse sets of metadata records and obtain meaningful results. Therefore multiple users could retrieve the same document based on business needs even though their query was completely dissimilar. This improves the records as they become self-explanatory and meaningful to the end user. This is particularly useful in potential compliance or discovery issues that may occur in the future when the documents need to be quickly identified and reviewed.



## Automatic Classification

The Concept Searching automatic classification function classifies content that has been tagged with the highly relevant compound term metadata associated with the corporate taxonomies, or a Records Management taxonomy. This eliminates all costs and human intervention associated with manually tagging documents for classification and results in information that can be categorized in real-time. This speeds the identification and collection of unstructured content from multiple sources and removes the burden from the records management staff to identify all relevant content, identify relevant content that may have not been found, and irrelevant documents that should not be part of the records management system.

## The Need for Taxonomies

Due to the explosion of electronic content, a records management solution must include a comprehensive approach to identify, organize and retrieve content assets. As new content is created or ingested new records need to be created. Implementing a framework that enables identification and management of unstructured content can facilitate the records management process.

Concept Searching's robust automatic classification and taxonomy management tools are designed to provide as much depth or hierarchical granularity that is needed within the organization. Since the automatic semantic metadata generation has the ability to identify concepts as opposed to keywords, documents with the same concept can be classified against multiple nodes within a taxonomy or multiple taxonomies. From an end user perspective, knowledge workers can locate pertinent information from his/her own individual viewpoint without knowing the exact search terms to use. The easy-to-use taxonomy and automatic classification features function as a labeling mechanism to quickly create the foundation that can be altered to suit the unique requirements of the organization.

The ability to automatically generate semantic metadata from unstructured content is extremely valuable. A taxonomy (or classification structure) provides a hierarchical view of topics that have been grouped together because they share the same quality or characteristic. Because of the semantic metadata generation, documents can be grouped in the taxonomy based on their relationships and relevance based on concepts. Pointers to the documents may exist in multiple categories as one document may contain multiple concepts. Traditional taxonomy tools often require significant investments in time, expertise, and money to develop and maintain. Features such as automatically generating compound term clues from the document corpus, dynamically showing the effect of changes on the taxonomy, and class weighting influenced by parent, child, and sibling can reduce taxonomy development and on-going maintenance by 66%-80%. Providing both automatic and manual classification, Subject Matter Experts (SME's) or Records Managers can utilize rich features such as node weighting, ability to see the 'concepts in context', ability to search the corpus, auto-clue suggestion for categorization, and instant feedback on the impact of changes.



## Governance at the Desktop

The most important criteria for a Records Management system is end user ease-of-use. The typical approach when implementing a Records Management system is to enforce new rules and, in the process, reduce productivity and create adoption issues. Failure to embrace the solution, lack of end user adoption and inconsistent adoption across the enterprise are factors that contribute to increased risk exposure and significantly dilute the success of the project. Ideally the content should be captured at creation without end user involvement.

Unlike Federated Records Management that provides the ability to automatically declare a record, Concept Searching can automatically classify content based on concepts from within the Microsoft Office applications. Optionally, the end user can add manual adjustments to the classification to provide further refinement if required. Either stand-alone or coupled with federated records capabilities compound term metadata can be added to the record and, if needed, modified by the Records Manager. Not only does this eliminate end-user adoption issues but ensures that all newly created content is correctly tagged with far richer metadata that can be used by the Records Management processes.

The feature set can be used in conjunction with the Microsoft Records Center to provide the ability to develop organizational, functional, geographic, and program related taxonomies from validated sources to facilitate the classification of an organization's electronic records, applying retention metadata and providing automatic upload, verification and ability to browse records from the original location and ensure lockdown by the Records Center.

## Integration with Microsoft Search

End users need to identify content in the context of what they are seeking. Information is only useful if users can find the information when they need it. The search engine must look beyond the ambiguity of natural language and identify the content fragments they require to solve the problem they are facing at that moment. The ability to search on concepts as opposed to keywords for relevant information within an organization can greatly improve the search experience.

Presenting relevant information to diverse end users through effective search is enabled via taxonomy based navigation or through faceted navigation. Taxonomy based navigation dramatically improves the search experience<sup>4</sup>. Faceted navigation is a logical extension of the taxonomy. The end user controls the search experience and the search results present 'facets' of documents grouped together based on the concepts identified. These facets extend the search process as documents will be grouped by concepts and assists the end user by offering content that may not have been found. This unified view and access to relevant information across disperse silos of information can increase productivity and enable knowledge workers to effectively query, use and re-use organizational content.



## The Benefits

The ability to improve the implementation and on-going management of a records management solution can be augmented through the use of Concept Searching's technologies. The tools complement records management and can help address many of the issues organizations face in meeting legal and regulatory compliance.

These benefits include:

- ◆ Provides transparent governance at the desktop eliminating end user adoption issues
- ◆ Protects enterprise record integrity as the burden of adding metadata is removed from the end user
- ◆ Ensure the records long term usefulness and reduces the costs and time to manage the records through compound term metadata generation
- ◆ Facilitates and enables the retrieval of relevant information and identifies highly correlated content that normally would not be found reducing organizational risk
- ◆ Enables accessibility to records in a way that meets the needs of the organization and other stakeholders, increasing productivity and improving decision making capabilities
- ◆ Assists in the evaluation of records that are not needed because they have no value, reducing costs and streamlining organizational and operational change
- ◆ Creates virtual centralization through the ability to link disparate content repositories enabling the effective management of organizational content and identification of similar records reducing costs, and risk exposure
- ◆ Reduces the time and cost of finding the 'correct' business record
- ◆ Improves the consistent development of corporate memory

## Summary

Deriving the benefits of a Records Management solution to an organization requires a change in individual and corporate culture. A key benefit of records management is the potential return on investment resulting in the ability to harvest the knowledge assets of the organization as well as ensure the organization is compliant and can effectively manage risk. Organizational content represents business assets to the organization and are of vital importance in not only the day-to-day tactical activities but also critical in achieving the organization's strategic goals and objectives. The approach of managing information as an asset encourages its collection, dissemination, and sharing and can reduce the cost of business operations and encourage a responsiveness to change.

Managing content goes hand-in-hand with a Records Management solution. To achieve the benefits promised organizations must move beyond the generation of traditional metadata generation or metadata generation based on only keywords. Creating compound term metadata and utilizing automatic classification, taxonomies, and concept based search can be the impetus to improve information access and consistent information use. This approach can deliver business benefits and drive organizations toward realizing the full potential of their information assets.

## References

<sup>1</sup>Sprehe, J.T., "Enterprise Records Management: Strategies and Solutions", September 2002.

<sup>2</sup>E-Government Policy Framework for Electronic Records Management, Public Record Office, [www.pro.gov.uk/recordsmanagement/](http://www.pro.gov.uk/recordsmanagement/).

<sup>3</sup>Priscilla Caplan, "Metadata Fundamental for All Librarians".

<sup>4</sup>Hao Chen & Susan Dumais, 'Optimizing Search by Showing Results in Context'.

## About Concept Searching

Founded in 2002, Concept Searching's software products deliver advanced search, auto-classification, taxonomy management and advanced metadata tagging solutions from the desktop to the enterprise. Concept Searching is the only statistical metadata generation and classification software company in the world that uses concept extraction and compound term processing to significantly improve access to unstructured information.

Headquartered in the U.K. with offices in the U.S. and South Africa, Concept Searching solves the problem of finding, organizing, and managing information capital. For more information about Concept Searching's solutions and technologies please visit [www.conceptsearching.com](http://www.conceptsearching.com).

Europe  
9 Shephall Lane  
Stevenage  
Herts SG2 8DH, UK  
P: 44 1438 213545  
[info-uk@conceptsearching.com](mailto:info-uk@conceptsearching.com)

Americas  
8300 Greensboro Drive  
Suite 800  
McLean, Virginia 22102 USA  
P: 1 703 531 8567  
[info-usa@conceptsearching.com](mailto:info-usa@conceptsearching.com)

South Africa  
15 Conifer Road  
Tokai, 7945  
Cape Town, South Africa  
P: 27 21 7125179  
[info-sa@conceptsearching.com](mailto:info-sa@conceptsearching.com)