



# THE BENEFITS OF COMPOUND TERM METADATA IN E-DISCOVERY IN A MICROSOFT SHAREPOINT ENVIRONMENT

This document discusses the importance and value of semantic metadata generation during the e-Discovery process and tools available to reduce costs and meet electronic discovery challenges.

CONCEPT SEARCHING



## Introduction

The cost of e-Discovery is the single largest cost in litigation today. For organizations that have massive volumes of potentially relevant documents the economic repercussions due to litigation can be devastating. Despite the potential financial impact, 52% of organizations do not have an e-discovery plan and 53% can not comply with all of the requirements of the Federal Rules of Civil Procedure (FRCP)<sup>1</sup>. With the burgeoning amount of digital information within an organization, any e-Discovery initiative must deal with large amounts of unstructured information.

Not only is FRCP compliance an issue with corporations and law firms, the true problem lies in e-Discovery practices that are inefficient, time consuming, and costly. Most organizations have already addressed the issue of records management and preserving documents but have not addressed possible improvements to the identification, collection, analysis, and review process of e-Discovery where significant benefits can be achieved.

## The Metadata Issue

Metadata by definition is 'data about data'. Syntactic metadata can be automatically created by the application, for example the creation date of a document, the author, properties, versions, etc. Metadata can also be added by the end user to further define the content within the document. Driven by an organization's records management policies metadata is often required to be added by users, which is haphazard at best and in many organizations users refuse to take an active role in managing information leaving an organization with no governance and possible litigation exposures. From the perspective of e-Discovery legal teams are forced to find and review all documents relevant to the litigation and in the process eliminate all the irrelevant documents which is extremely labor intensive.

Semantic metadata generation identifies how data items are related as well as the meaning of the content. For example, if a document contains the words balance sheet and income statement semantic metadata generation would identify that the document is about accounting and/or finance, even though neither of those words appeared in the document. The legal team can see the concepts within the context of the document and appropriately decide if the document is relevant. Additionally the identification of potential keywords to aid in the negotiations at the 'meet and confer' conference can be simplified. This ability to capture the relationships between multiple documents, automatically understand the concepts and group the results together not only reduces the e-Discovery time during the identification, collection and the review phase but also provides a comprehensive approach to better manage unstructured information within the organization.



## FRCP Metadata Amendments

In 2006 new FRCP amendments became effective that clarify the use of metadata in e-Discovery. Organizations that do not proactively develop policies and implement processes to address not only their electronic data but also their metadata run the risk of incurring penalties, sanctions and potential loss of business and reputation. Processes to manage electronic data and the associated metadata can also reap benefits for the organization including reduced costs and facilitating the e-Discovery process.

The amendments define guidelines only, as opposed to concrete rules. Based on the approach of the Sedona Conference Working Group the principles place the burden of dealing with electronic data on the organization to identify the most effective way to preserve and produce the data. The most significant implication is that the treatment of electronic documents and paper documents from a legal perspective are equivalent. Failure to comply can result in sanctions, default judgments, potential monetary fines, and the risk losing cases that may have been previously won. There are five amendments that establish guidelines for the treatment of metadata.

Rule 34(A) created a new category of Electronically Stored Information (ESI). This new amendment removes the distinction between electronic and paper documents. In regards to metadata, it becomes 'discoverable' under the FRCP and organizations must be able to preserve and produce ESI.

Rule 34(B) provides the requesting party to specify the format in which ESI is produced, preferably in the native file format in which the data was created. Traditionally litigants provide electronic documents in the TIFF format, a scanned copy of the printed document, PDF which is considered more convenient and usable, or native file format which is the actual format the application generates to produce the document. The advantage of using native file format is obviously a reduction of costs as scanning or conversion is not necessary. However, in native file format all metadata will be accessible as well as information that may contain embedded data such as a formula in a spreadsheet.

Rule 26(b)2 determines categories for accessible and inaccessible ESI. This effectively eliminates the ability to argue that metadata is not accessible and the generation of ESI will involve the consideration of metadata. Inaccessible refers to the economic (or other) burdens of generating ESI, not in terms of being hidden or inaccessible. Therefore all metadata or embedded data will be considered as accessible.

Rule 37(f) created safe harbor protection for organizations that can demonstrate policies and procedures as to the management of computer systems. Some experts believe the language is broad enough to include the management of metadata. Although unclear of the result, organizations with metadata policies and procedures will be able to present arguments that they fit into the safe harbor provisions.

Rules 16(b) and 26(f) requires that discussion and understanding of ESI issues for both parties must happen within 45 days of filing a case. These meet and confer rules include the discussion of the handling of ESI and electronic discovery issues, including metadata.



## The Costs

Unmanaged data, not even considering e-Discovery issues, is pervasive and increasing rapidly. Approximately 80% of all data within an organization is unstructured and over 90% is unmanaged<sup>2</sup>. When faced with potential litigation, regulatory issues, or compliance the identification of pertinent and relevant information can become a costly endeavor. According to Gartner Group, the average Fortune 500 Company will respond to 6-10 discovery requests per year at a cost of \$1.6 million each. Factors to consider include:

### Document Decisions per Hour

The review process has a significant impact on total e-Discovery costs. The ability for an attorney to quickly and accurately determine a document's relevance and/or non-relevance can generate significant cost efficiencies. According to KPMG, utilizing a document analytic tool the attorney can review approximately 1,000 pages per hour electronically as opposed to 200 pages per hour reviewing documents electronically without a tool<sup>3</sup>.

### Cost Per Document

How much is spent on each document, including both the technology cost and the attorney review cost?

### Elimination of Repetitive and Irrelevant Data

Reducing expenses related to irrelevant and repetitive material presents one of the largest opportunities for reducing costs. As much as 70% of emails and corporate documents are duplicates<sup>4</sup>. The challenge during the e-Discovery process is to be able to identify not only repetitive data but also irrelevant data. The more quickly an organization can reduce the size of the corpus the costs associated with the identification and review process will be significantly less.

### Review Completion

Understanding the amount of material left for review can be important to managing expectations and evaluating the review strategy, especially in a large scale document review.

## The Concept Searching Approach

Traditional information retrieval systems use 'keyword searches' of text and metadata as a means of identifying and filtering documents in e-Discovery. These keyword searches can include the use of simple words or combinations of words and often use Boolean operators to further refine the information retrieval. Although the ability to perform keyword searches against large quantities of documents is a useful tool, there are still inherent issues. Keyword search captures only 33% of relevant information resulting in the retrieval of potentially a large amount of documents that are not weighted nor ranked based upon their relevance. Each document is considered to have an equal importance and equal probability of relevance, therefore each would require manual review. Boolean operators and other techniques can be used to increase the number of relevant documents and a Boolean argument is often created to achieve more relevant results. Although commonly used, these approaches are limited by their dependence on matching specific language entered by the legal professional to retrieve the desired topic of interest.

How to search for and find the appropriate and relevant documents during the identification stage is hampered by the search specialists' ability to think of every known term that would be applicable. Often the different parties will use different words, depending on their role. It is estimated that legal professionals are less than 20% to 25% accurate and complete when searching and retrieving information from a heterogeneous set of documents<sup>5</sup>.



## Precision and Recall

Precision and recall are the two key performance measures for information retrieval. The ideal solution is to have them balanced. Precision is the retrieval of only those items that are relevant to the query. Recall is the retrieval of all items that are relevant to the query. Higher precision often leads to missing items that may be relevant, but may use a different vocabulary. Higher recall often leads to the retrieval of too many items that may be unrelated to the query.

Some systems allow the practitioner to modify either precision or recall. By adjusting the system to retrieve more documents the recall is increased but at the expense of retrieving more irrelevant documents, decreasing precision. By increasing recall the identification and review process will be lengthier due to potentially irrelevant documents being identified. By increasing precision, important relevant information may be missed that could jeopardize the litigation result.

## The Solution

Concept Searching's suite of software solutions address both recall and precision and automate many of the laborious and costly tasks e-Discovery teams currently utilize during the e-Discovery process. The technologies are unique as 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 access to unstructured information. The tool set provides semantic metadata generation, automatic classification and taxonomy management.

## Semantic Metadata Generation

Automatic semantic metadata generation enables an organization to extract compound terms, acronyms, and keywords from a document and corpus of documents that are highly correlated to a particular concept or meta tag. Automatically identifying the most significant patterns in any text, these compound terms are then used to generate metadata based on an understanding of conceptual meaning. When these compound terms, acronyms, and keywords are prevalent within a particular document that document is automatically meta-tagged eliminating the requirement for an individual to read that document and subjectively apply metadata to the properties of a document. The ability to identify 'concepts in context' ensures that the e-Discovery process can identify all potentially discoverable data.

## Taxonomy Management

The ability to automatically generate semantic metadata from unstructured content is extremely valuable. However, the organization of the content and presentation to the e-Discovery team must also facilitate the e-Discovery process. 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. Concept Searching's taxonomy management tool has been proven to reduce the time to build and subsequently maintain taxonomies by 80%. Providing both automatic and manual classification, Subject Matter Experts (SME's) 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. The taxonomy provides the structure for the grouping of like documents together and enables a more targeted, accurate, and efficient e-Discovery process which translates into reduced costs and improved productivity.



## Automatic Classification

The automatic classification function classifies content that has been tagged with the highly relevant metadata associated with the organizational taxonomies. This eliminates all costs and human intervention associated with manually tagging documents for classification and results in information that is categorized in real-time. This speeds the identification and collection of business, legal, and regulatory records from multiple sources and removes the burden on the e-Discovery team to identify all relevant and discoverable content. This also enables the identification of new information to be captured and identified early in the discovery process and immediately made available to the discovery team.

## One Click Save

Additional functionality exists to enable legal professionals to automatically classify content from within the traditional Microsoft Office interface. This can be done automatically, or optionally the legal professional can add manual adjustments to the classification to provide further refinement if required. This ensures that organizational content is consistently classified and available during the e-Discovery process.

## Navigation

Presenting relevant information to the legal professional through effective search is enabled via taxonomy based navigation or through faceted navigation utilizing Microsoft Search products and the familiar interface. Taxonomy based navigation dramatically improves the search experience<sup>6</sup>. Faceted navigation is a logical extension of the taxonomy. The legal professional 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 concept and assists the legal professional by offering content that may not have been found. This unified view and access to relevant information across disperse silos of information can reduce the volume, cost, and time required for the e-Discovery process in the identification, collection, review and analysis phases.

## Microsoft Integration

The robust technology framework is fully integrated with Microsoft Office SharePoint 2007, Microsoft Search Server 2008 and Microsoft Search Server Express. Extending the capabilities of Microsoft SharePoint Products, the technology can be integrated into Microsoft Office, the Business Data Catalog and the Microsoft Records Center to improved the classification and retrieval process for organizations that require access to unstructured information. Full integration with Microsoft Office SharePoint Server 2007, Search Server 2008, Search Server Express and Microsoft Office is accomplished through SOA compliant services, Web Parts, and Open XML. The software does not need a separate index, automatically populates Microsoft SharePoint properties with compound term metadata, and fully respects inherent Microsoft SharePoint security.

## Summary

The challenges and escalating costs for e-Discovery will continue to increase. Concept Searching provides effective tools to reduce costs and alleviate electronic discovery challenges. The ability to reduce the time in the identification, collection and review process enables attorneys to spend more time on higher value issues of the litigation, realizing substantial cost savings for the organization. The tools also help ensure the organization is in compliance with the FRCP metadata amendments. Through electronic automation, content can be meta-tagged, classified, and presented to the legal professional in a manner that enables them to more rapidly identify relevant information based on content not keywords. Significant benefits can be achieved by removing the ambiguity in content and the identification of concepts within a large corpus of information. The Concept Searching software delivers expediencies and reduces costs in several phases of e-Discovery offering an effective solution that overcomes many of the challenges found in the traditional electronic discovery process.

## 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).



## References

<sup>1</sup> Osterman Research

<sup>2</sup> Doculabs

<sup>3</sup> 'A Revolution in e-Discovery', The Persuasive Economics of the Document Analytic Approach, KPMG

<sup>4</sup> Law Technology News, January 2004

<sup>5</sup> David C. Blair & M.E. Maron, 'An evaluation of retrieval effectiveness for full-text document retrieval system,' 1985.

<sup>6</sup> Hao Chen & Susan Dumais, "Optimizing Search by Showing Results in Context".

**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)