The Trend Towards Intelligent Metadata Solutions
White Paper

Prepared by:
Concept Searching
8300 Greensboro Drive
Suite 800
McLean
VA 22102
USA
+1 703 531 8567

9 Shephall Lane
Stevenage
Hertfordshire
SG2 8DH
UK
+44 (0)1438 213545

info-usa@conceptsearching.com
http://www.conceptsearching.com
Twitter:@conceptsearch
Concept Searching Blog

Martin Garland
President
+1 (703) 531 8567
marting@conceptsearching.com

January 10th, 2013

© 2013 Concept Searching
Abstract

In the fourth quarter of 2012, Concept Searching conducted an analysis of enterprise clients that have deployed Concept Searching solutions or are planning to deploy a metadata solution to help identify executives’ key priorities in 2013. For new clients, improving search is always on the list, however, the increase of legislation and the proliferation of unmanaged social networking and free form communications are escalating compliance and governance to board-room issues. The study also indicated that migration is a high priority and is no longer a copy and paste exercise, but that information compliance and governance are high priorities in the migration process.

This White Paper focuses on the results of the study and describes the current use of Concept Searching technologies to resolve issues of migration, compliance and governance, specifically in the areas of intelligent content migration, sensitive information breaches, records identification, text analytics, social networking, FOIA, eDiscovery, and information transparency.

Author Information

Martin Garland has over 20 years’ experience in search, classification and Enterprise Content Management within the broader information management industry. His keen understanding of the information management landscape and his business acumen provide a solid foundation for guiding organizations to achieve their business objectives using best practices, industry experience, and technology. Martin’s expertise has been instrumental in assisting multi-national clients in diverse industries to understand the value of managing unstructured content to improve business processes.

He has focused on sales, marketing and general management, and has expertise in both startup and turnaround operations throughout Europe, the US and Asia Pacific. One of the founders of Concept Searching, Martin is responsible for both business strategy and North American and International operations.
Overview

Concept Searching has been successfully deploying intelligent metadata enabled migration and compliance solutions for several years that build on semantic metadata tagging, auto-classification, and taxonomy management. What began to emerge was a lack of information governance planning that incorporated and addressed unstructured and semi-structured information. In early 2012, Concept Searching introduced the Smart Content Framework™ to provide its clients with an information governance roadmap to incrementally address a wide range of business processes with one set of technologies, as well as the implementation of policy and governance for the on-going management of unstructured information.

In the fourth quarter of 2012, Concept Searching conducted an analysis of enterprise clients that have deployed Concept Searching solutions or are planning to deploy a metadata solution to help identify executives’ key priorities in 2013. For new clients, improving search is always on the list, however, the increase of legislation and the proliferation of unmanaged social networking and free form communications are escalating compliance and governance to board-room issues. The study also indicated that migration is a high priority and is no longer a copy and paste exercise, but that information compliance and governance are high priorities in the migration process.

This White Paper focuses on the results of the study and describes the current use of Concept Searching technologies to resolve issues of migration, compliance and governance, specifically in the areas of intelligent content migration, sensitive information breaches, records identification, text analytics, social networking, FOIA, eDiscovery, and information transparency.

Concept Searching’s Smart Content Framework™ and technologies have been delivering intelligent metadata enabled applications to a wide variety of Fortune 1000, Government, and Public Sector enterprises for over twelve years. Regardless of size or industry, any organization that places high value on knowledge assets or faces strict regulatory guidelines, can benefit from an industry unique solution able to capture the meaning of content and use associated metadata, to not only proactively manage the content but to also deploy intelligent metadata enabled solutions that are integrated with the organization’s information governance policies. The outcome delivers quantifiable business objectives, while mitigating corporate risk and reducing costs.

“...The worlds of data governance and content governance are not at all aligned. How do people access data when they are trying to answer a question through reporting or business intelligence? They ask a very structured question and they get the right answer.

Now, how do people access information with content? They go do a search and they get 100,000 potential answers and they are OK with that, flipping through it until they find the answer for them. There are very different cultures in how these kinds of information are managed.”

Rob Karel, Principal Analyst for Data Management Forrester Research, Inc.
The Concept Searching Foundation

The Concept Searching technologies include a Service Oriented architecture (SOA) based search and classification technology, a browser based taxonomy management technology, and a tightly integrated feature set that operates with any platform. Industry unique compound term processing technology enables the rapid creation of semantic metadata, which can be classified to organizationally defined taxonomies. The tagging and auto-classification of content can be aligned to business goals and the semantic metadata generated can be easily integrated with any third party application or platform that can interface via web services.

From a business perspective, one set of technologies can be used for intelligent metadata enabled solutions that result in improved information transparency, compliance, governance, preservation and storage of documents of record, securing sensitive information, text analytics, and social networking.

To support the unique needs of organizations, the Smart Content Framework™ was developed by Concept Searching to address the need for information governance as it applies to unstructured content. Most companies still do not manage their unstructured content, nor do they use it to deploy intelligent metadata driven business applications. This is illustrated by the typical approach that is based on the deployment of siloed applications. The Concept Searching approach is to build an information governance plan where multiple stakeholders can achieve their objectives and the underlying technologies and be used and reused for a variety of business processes.

The Smart Content Framework™ provides the building blocks to not only manage unstructured content but to leverage content assets to reduce organizational risk, solve business challenges, and improve business processes. The flexibility of the technologies enables the organization to address key failures with the traditional management of unstructured content and more rapidly achieve business objectives.
Compound Term Processing

The Smart Content Framework™ recommends the development of an infrastructure framework that can be used, and reused to support any application that requires metadata. This is the first and most critical step in developing a sound information governance plan.

The Concept Searching technologies are based on the still unique use of compound term processing. Instead of identifying single keywords, compound term processing identifies multi-word terms that forms a complex entity and identifies them as a concept. By forming these compound terms and placing them in the search engine’s index, or making them available to any application that requires metadata, the outcomes are highly accurate, because the ambiguity inherent in single words is no longer a problem. As a result, a search for “survival rates following a triple heart bypass” will locate documents about this topic even if this precise phrase is not contained in any document. A concept search using compound term processing can extract the key concepts, in this case “survival rates” and “triple heart bypass” and use these concepts to select the most relevant documents.

Identification of concepts within a large corpus of information removes the ambiguity in search, eliminates inconsistent meta-tagging, and automatic classification and taxonomy management based on concept identification simplifies development and ongoing maintenance.

The key benefit of compound term processing is that it is an adaptive technology that can generate conceptual metadata at source, is not based on keywords, proximity, or algorithms that can’t be changed, and can capture very specific criteria for business applications using the metadata. The end result is a rich set of conceptual metadata that reflects the unique terminology and vocabulary of the organization, thereby rendering it valuable in search, records management, compliance, legal and eDiscovery, and data privacy. Concept Searching is unique among all commercially available technologies in incorporating a compound term processing engine.
Intelligent Metadata Infrastructure

The role of metadata has been transformed from being an afterthought to a fundamental requirement for organizational growth, profitability, and risk reduction. The term itself is abstract and still not widely understood by business users. Yet metadata has played a critical role in IT investments for many years, including Knowledge Management, Data Warehousing, Data Mining, Business Intelligence, and Customer Relationship Management.

The traditional use of metadata is becoming much broader in scope yet more focused on solving critical business problems. With the unprecedented growth of unstructured content, the rise of social media, the cloud, and more stringent compliance mandates, organizational risk is on the increase. New concerns of potential data breaches, records management challenges, litigation, and the age old problem of poor search, organizations must face these new issues, and many are now key priorities at the board room level.

Creating metadata repositories and taxonomies that are optimized for the organization is challenging, as each participant in the process, and every end user, may have a different way of expressing the same or similar descriptors (metadata). The goal is to give people not only the right information, but information distilled from a variety of distinct content, making available usable knowledge.

A comprehensive approach requires more than syntactic metadata, and requiring end users to add rich metadata is haphazard and subjective at best. Since the Concept Searching solutions are not restricted to keyword identification, compound term metadata or intelligent metadata can be automatically generated, either when the content is created or ingested. The generation of metadata based on concepts extracts compound terms and keywords 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.

An Intelligent Metadata Infrastructure is a proven approach to solve these challenges. The framework provides the platform to ensure all content is managed in a consistent manner and adheres to organizational policies regardless of where the content reside.

Eliminating Information Silos

Typically, the acquisition of a technology solution is made to improve a specific business function. For example, search, records management, or data security. An organization may have SharePoint, Oracle, or a host of platforms, none of which are used to meet their potential. A better approach is to seek to maximize the current investment in what they have, by using technologies that can serve several purposes and offer interoperability with any enterprise infrastructure platform. A more comprehensive yet broader approach is needed, that focuses not only on achieving the immediate tactical objectives, but also views the introduction of technology as a strategic enabler for information governance that can be accomplished in clearly defined steps. The
traditional approach, still being used, is to acquire software to fulfill a specific business function.

The issues in this scenario include:

- Inability to create an enterprise metadata framework
- Cannot capture, use or reuse intelligent metadata at source
- Inability to automatically apply information governance as content is created or ingested
- Expertise and skill sets required for the setup and maintenance of each application
- Higher costs, initially and on-going
- Inability to ensure forward compatibility with other applications
- Increased organizational risk

An Intelligent Metadata Approach

The development of an enterprise metadata framework provides flexibility, is ideally platform independent, and enables organizations to implement governance policies and operations incrementally, with each step directly correlated to business value. From this overarching infrastructure framework, intelligent metadata enabled solutions can be deployed to address the unique challenges of organizations. These challenges may include search, compliance, information lifecycle management, sensitive information
protection, migration, text analytics, and Enterprise/Web 2.0.

The Intelligent Enterprise Metadata Repository is the framework that will process the metadata and auto-classify it to one or more taxonomies. Any application that uses metadata can access the conceptual metadata, and the taxonomies can be easily managed using the metadata and organizationally defined descriptors.

There are significant benefits with this approach including:

- Auto-classification of content aligned to corporate goals
- Automatic intelligent content routing as defined by an organization’s information governance policies
- Automatic identification, routing, lock down and notification of sensitive information at source mitigating risk
- Real-time generation of intelligent metadata at source improving information transparency
- Intelligent metadata can be retrieved by any application that requires the use of metadata
- Works with any other application that is SOA compliant, for example, a security application, but the requirement to identify unique organizational content that should be secured is needed, or with any search application to improve search results
- Deployed as a single process lowering costs and delivering a quantifiable ROI
- Reduces organizational complexity, eliminates manual content tagging ensuring consistency
- Elimination of expertise and skills required for each application
Intelligent Search

Based on the responses in the Concept Searching study, search is still the primary purpose for implementing our technologies, either using its conceptSearch product, or in conjunction with any other search engine. Organizations must recognize that search is not a stand-alone technology or application but must be integrated with business processes and corporate objectives as a key infrastructure component.

The term Insight, as defined in Concept Searching’s Smart Content Framework™, goes beyond the typical search application and the overused ‘findability’ term popular today. Regardless of the enterprise search solution, the delivery of meaningful results depends on the ability to effectively index and classify content, and to develop taxonomies to better manage the content. With the unabated growth of unstructured content and the introduction of diverse environments such as on-premise, cloud, or hybrid, the need to access relevant information is falling far behind organizational needs. Content can exist in multiple repositories, and if not managed will increase organizational risk, particularly in the area of securing unknown data exposures.

Manual tagging is still the primary approach to tag content, which is subjective at best and often lacks any alignment to the enterprise goals or mission. This subjectivity is immediately applied to search results, resulting in inaccuracy and the inability to find relevant information. The extraction of metadata through unsophisticated tools can extract system defined metadata, through keywords and proximity, extensive rule building, or artificial intelligence, which provide no way to adapt to meet unique organizational needs. All these approaches have significant failure rates.

A key component in an intelligent metadata enabled information transparency solution is the ability to automatically generate semantic metadata, which Concept Searching terms ‘concepts in content’, that is based on one or more words that can be considered a concept. In conjunction with conceptTaxonomyManager, Subject Matter Experts can test, validate, and manage the multi-word terms aligned to their internal vocabularies.

Intelligent Records Management

Concept Searching has delivered intelligent metadata enabled search solutions since 2003, integrating its own Concept Searching platform, as well as FAST, Autonomy, SharePoint Search, Google Search Appliance, and Solr. Concept Searching’s advanced search platform is currently in use with many Enterprise commercial, Federal and DoD clients.
The ability to facilitate records management processes with the introduction of conceptual metadata generation has proven to be a valuable component in the business process, with 25% of Concept Searching clients using it for that purpose. In the broader marketplace, 27% are looking at the value of intelligent metadata generation to improve these processes.

Records management, as an intelligent metadata enabled application, includes the ability to identify records, privacy information, intellectual assets, and fully automate the process transparently, without user involvement, as well as handle the appropriate disposition of the content. This includes discovering where the content resides, cleansing the content through organizationally defined concepts and descriptors, identifying the relationships within the content and then applying the policies for automatic enforcement and routing to the appropriate repository.

Proof of compliance and data protection can only be accomplished through the automatic enforcement of organizational policy, to ensure consistency, and can be implemented transparently, resulting in improved record keeping, monitoring, and auditing processes.

### Intelligent Data Privacy

Data privacy is high on the objectives of CIOs for 2013. With new mandates, compromised content has now become a corporate responsibility, even to the level of holding executives personally liable. According to PRISM International, 67% of data loss in records management is due to end user error. This, of course goes back to the end user being responsible for correctly identifying the record with the appropriate tags to ensure processing. The Concept Searching analysis revealed that 29% of clients are using the technologies for the identification of privacy/confidential Information and 51% are looking for a solution that addresses data privacy.

Coupling the Smart Content Framework™ with the Concept Searching technologies, an intelligent metadata enabled application, which includes the ability to proactively identify and protect unknown privacy exposures before they occur, can be achieved. As content is ingested or uploaded, real-time monitoring of organizationally defined vocabulary and descriptors can occur, and the content can be processed and protected according to established organizational guidelines.

### Intelligent Migration

Concept Searching has delivered intelligent metadata enabled records identification solutions to commercial and DoD clients since 2009, with records solutions as diverse as SharePoint, FileNet, Hummingbird and OnBase.

Concept Searching has delivered intelligent metadata enabled data privacy solutions since 2008. One DoD client has not had an internal or external data security breach since deployment in 2008.
Surprisingly, across the board migration is a high priority. The migration process is now requiring more sophisticated techniques to ensure compliance objectives are met. Simply moving documents from one repository is not enough, as content that was typically unmanaged will remain unmanaged, continuing to expose an organization to risk. Information cannot be managed from inception to deletion without comprehensive metadata associated with the content, and incorporating the multiple channels and origination points from which content was received.

Responses from the Concept Searching client base revealed that 33% have used the technologies for migration, with 53% looking for a solution that can intelligently address the migration process.

Migration of unstructured content can be a laborious and time consuming project. The challenge is that documents can exist in multiple places at the same time, different revisions of the same document exist, some documents should be deleted, and others should be archived. There may be records that were never declared, as well as confidential or privacy information that will not be identified when migrated. The ability to mass move content is relatively straightforward. However, from an information governance approach, mass moving content results in the same problem of mis-managed content.

To migrate document collections effectively the text content of each document needs to be searched to determine its value. This cannot be done manually, as the volume is too high, and the consistency of human review and decision making is unreliable as well as costly. Migration must also consider the security of the documents as they are moved to their new location.

As an intelligent metadata enabled solution, migration tasks are simplified and the accuracy of the migration is greatly improved. As content is migrated it is analyzed for organizationally defined descriptors and vocabularies, which will automatically process the content to the appropriate repository for review and disposition.

**Intelligent Data Analytics**

The current maturity of text analytics tools rates the certainty of the information at less than 70%. Vendors of all kinds are jumping on this bandwagon with the intent to open new markets. Big Data as it relates to unstructured and semi-structured content does not fit into the current models. In regard to the ability to use text analytics across the corpus of information, 23% of clients are using the solution for text analytics, which would appear to have been an extension of the capabilities fundamental to the technologies as they are already installed. However, in the broader market, only 16% felt that the capabilities for text analytics are important.

Big Data and more specifically, a focus on text analytics, deals with semi-structured or unstructured data. Unstructured content in the context of Big Data has been pigeon-holed into the database approach, which can’t work. Data is machine driven, whereas unstructured content is driven by people, which makes the nuances, insights, relationships of disparate content, sentiment, and knowledge capital much more difficult to extract. Unstructured content is also continually in a state of flux and changes rapidly.

There are two issues that are consistently stumbling blocks, which organizations
typically do not know how to solve. These are end users’ inability to correctly tag content for reuse and an organization’s inability to enforce policy. Concept Searching’s technologies analyze and extract highly correlated concepts from very large document collections from diverse repositories and applications, such as Twitter and blogs. This enables organizations to attain an ecosystem of semantics that delivers understandable results. The valuable insight gained can be used to identify competitive advantages, customer perception, regional trends, and, perhaps more importantly, identify the internal knowledge capital that exists but is rarely used because it cannot be found.

**Intelligent FOIA and eDiscovery**

eDiscovery and Freedom of Information Act (FOIA) requests are costly, typically inefficient, and time consuming. Although organizations have addressed the issue of records management and preserving documents many have not addressed possible improvements to the identification, collection, and analysis of relevant content where significant benefits can be achieved. Based on Concept Searching’s study, 21% of current clients use the technology for eDiscovery or FOIA. It is important to note that Concept Searching is not a legal application, but can integrate with legal and eDiscovery applications to improve the ability to capture the relationships between multiple documents, automatically understand the concepts, and group the results.

Traditional information retrieval systems use ‘keyword searches’ of text and metadata as a means of identifying and filtering documents in eDiscovery and FOIA. 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. 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. Although Boolean operators are commonly used, these approaches are limited by their dependence on matching specific language entered by the knowledge professional to retrieve the desired topic of interest.

How to search for and find the appropriate and relevant documents is hampered by search specialists’ ability to think of every known term that would be applicable. In eDiscovery, the different parties may use different words, depending on their roles. 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.

In 2011, FOIA requests surpassed 597K. The US government currently creates thousands of classified documents with over 20 million classification decisions made in a single year. In addition, there are currently classified documents that will become declassified according to the law which allows release after a pre-determined time period. The process for classifying documents is both time consuming and labor intensive. Document review and searching through the document to identify material called out in the classification guidelines is arduous and sometimes complex. Proper document marking of the security classification may take a few hours to several weeks.

The challenges and escalating costs for eDiscovery and FOIA continue to increase. Concept Searching provides intelligent tools to reduce costs and alleviate many of the challenges. Content can be semantically meta-tagged, classified, and presented to knowledge professionals in a manner that enables them to more rapidly identify relevant information based on concepts, and process the content accurately. Significant
benefits can be achieved by removing the ambiguity in content and the identification of concepts within a large corpus of information. This approach delivers expediencies and reduces costs, offering an effective solution that overcomes many of the challenges found in the eDiscovery and FOIA processes.

Intelligent Social Networking

Interestingly, with all the industry focus on social networking, and the acquisition of Yammer by Microsoft, clients, as well as the broader marketplace, do not currently seem to have any plans for introducing social networking into their organizations. Although unknown how this will evolve, social networking tools, that encourage collaboration and can link employees, partners, suppliers, and customers to share information, are becoming useful tools for business communication. The primary business benefits of these collaboration and social tools are also accompanied by inherent weaknesses. There are several concerns, such as security, unauthorized use, and communication noise. As an intelligent metadata enabled application, structure for social networking applications can be defined. The management and delivery of social networking applications driven by metadata provides the ability to present relevant information from diverse repositories, based on the concepts or attributes of information to different audiences.

The other benefit of course is to identify and deliver much more granular results to specific end users who would find the information valuable. Social networking, as it exists today, is based primarily on the end user’s ability to tag content correctly, but it also applies very subjective information based on the end user perception which the tools then use for ranking, which is somewhat haphazard. A building block in the Smart Content Framework™ enables as much or as little control to be integrated with social networking tools and is used to find relevant content that is based on the specific user.

Summary

The pulse of the marketplace has indicated the need for better approaches and tools to improve business processes. It is clear the marketplace is looking to utilize and access relevant content when needed, and at source eliminating the need for secondary applications. This holistic approach improves and streamlines business processes, reducing costs and risk. Intelligent metadata enabled solutions driven by Concept Searching’s Smart Content Framework™ and flexible core technologies have shown to improve and solve many business challenges in this way, including intelligently enabled:

- Search
- Records management
- Data privacy
- Migration
- Data analytics
- eDiscovery and FOIA
- Social networking

The first graphic illustrates how Concept Searching clients have deployed the technologies. Since all clients are at different stages of deployment and use, the findings represent the priority of their requirements at the time of acquisition of the technologies. As clients build the Smart Content Framework™ within their own organization, additional functionality can be rapidly implemented. A good example is the US Air Force, a Concept Searching client for eight years, currently using the technology to address all the solutions shown in the graphic.

![Graph showing technology priorities](image)

Although search is always one of CIOs’ objectives, other challenges dealing with compliance, governance, and data analytics are also top objectives. Concept Searching has delivered these intelligent enabled applications for several years, whether as a standalone solution, or in conjunction with third party applications to provide a best of breed solution.

The second graphic illustrates the priorities of organizations seeking not only a search solution, which is not depicted, but the priorities they need to address as corporate objectives.

The three highest solution objectives are intelligent migration followed by data privacy and protection, and records identification. eDiscovery, FOIA, and text analytics are more specific to certain types of organizations but are still high on CIOs’ priorities.
The application of an information governance strategy for unstructured and semi-structured content is becoming an objective for many CIOs, but is still in its relative infancy. Forward looking companies have realized that the key to an effective strategy must include the ability to identify, use, and reuse intelligent metadata to ensure compliance, reduce risk, and costs.

Utilizing the Concept Searching Smart Content Framework™ and technologies as the infrastructure platform, many organizations have reaped the benefits of managing content with a single set of technologies, to address their unique organizational challenges.
About Concept Searching

Founded in 2002, Concept Searching provides software products that deliver conceptual metadata generation, auto-classification, and powerful taxonomy management from the desktop to the enterprise. Concept Searching, developer of the Smart Content Framework™, provides organizations with a method to mitigate risk, automate processes, manage information, protect privacy, and address compliance issues. This infrastructure framework utilizes a set of technologies that encompasses the entire portfolio of unstructured information assets, resulting in increased organizational performance and agility.

Concept Searching is the only platform independent 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. The Concept Searching Microsoft suite of technologies runs natively in SharePoint 2007, SharePoint 2010, SharePoint 2013, FAST, Windows Server 2008 R2 FCI, Office 365, and in Microsoft Office applications.

The building blocks of Concept Searching’s Smart Content Framework™ are being used by organizations from a diverse number of industries including the US Army, the US Air Force, the UK MOD, Baker Hughes, Deloitte, Logica, NASA Safety Center, OppenheimerFunds, Point B, Perkins+Will, Parsons Brinckerhoff, Burns & McDonnell, DAI, MarketResearch.com, the US Department of Health & Human Services, Transport for London, the London Fire Brigade, the National Transportation Safety Board, and Xerox.

Headquartered in the US with offices in the UK, South Africa and Canada, Concept Searching solves the problem of finding, organizing, and managing information capital far beyond search and retrieval. The technologies are being used to improve search outcomes, in records management, to identify and secure sensitive information, improve governance and compliance, add structure to Enterprise 2.0, facilitate eDiscovery, and intelligent migration. For more information about Concept Searching’s solutions and technologies please visit http://www.conceptsearching.com.