

Capitalizing on Unstructured Content in the Energy Sector 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](https://twitter.com/conceptsearch)
[Concept Searching Blog](#)

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

May 2013

© 2013 Concept Searching

Abstract

Energy companies are facing challenges in managing their content to ensure compliance, reduce organizational risk, and increase preparedness for meeting corporate objectives. With end user adoption cited as the single most critical failure point in many line-of-business applications, tools that can automate the process can significantly reduce non-compliance issues and improve the quality of the information for retrieval and re-use. Managing information as an asset encourages its collection, dissemination, sharing, and ultimately reduces the cost of business operations.

This White Paper explores the use of Concept Searching's Smart Content Framework™ and intelligent metadata enabled solutions as an enterprise infrastructure for unstructured information governance that reduces risk, and ensures compliance. The framework addresses information complexity and provides solutions that can effectively increase information transparency and provide the ability to manage content regardless of where it resides. By providing immediate and actionable visibility into all relevant content decision making can be improved, risk minimized, and costs reduced.

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.

Table of Contents

Abstract	1
Author Information	1
Overview	3
A Single Platform to Capture, Use, and Re-use Knowledge Assets	4
Compound Term Processing	4
The Role of Metadata	5
Minimizing Risk in the Energy Sector	6
Managing Content, Regardless of Where It Resides	7
The Role of Intelligent Search	8
Intelligent Records Management, eDiscovery, and Litigation Support	8
Data Governance, Privacy, and Cyber Security	9
Multinational Content Asset Protection	10
Vocabulary Normalization	10
Intelligent Collaboration, Reducing Project Risk and Business Complexity	11
Changing Workforces and Inability to Capture Digital Intellectual Assets	11
Migration	12
Big Data and Text Analytics	13
Summary	14
About Concept Searching	15

“Most oil and gas companies would agree that the most significant challenge for their enterprise is management of information. Oil and gas companies continue to work to be able to create intelligence from the massive amount of technical and business data, both structured and unstructured that they have collected.”

Roberta Bigliani and Jill Feblowitz
IDC Energy Insights
July 2011

Overview

The impact from the change in the economic climate, increased pressure from shareholders, more stringent regulatory requirements, environmental issues, unanticipated process failures, and the ability to maintain competitive advantage is increasingly more difficult for the Energy sector. Aging baby boomers will continue to retire, requiring the need to capture and retain the knowledge assets of a skilled workforce that will not be easily replaced. Facing unprecedented changes in the industry, the need to ‘excel’ is much more difficult and has been replaced with the necessity to improve organizational performance and minimize risk to succeed.

All these challenges have the same fundamental requirement - improved utilization of enterprise information. The result must achieve the ability to access and distribute accurate and trusted information to other systems, employees, suppliers, partners, and third parties, on demand and typically globally, with a common vocabulary. Improving business performance requires information flows between systems and processes to be captured and improved through technology. Through the efficient management of all unstructured and semi-structured content, energy companies can more readily improve business processes, mitigate risk, and quickly adapt to regulatory requirements.

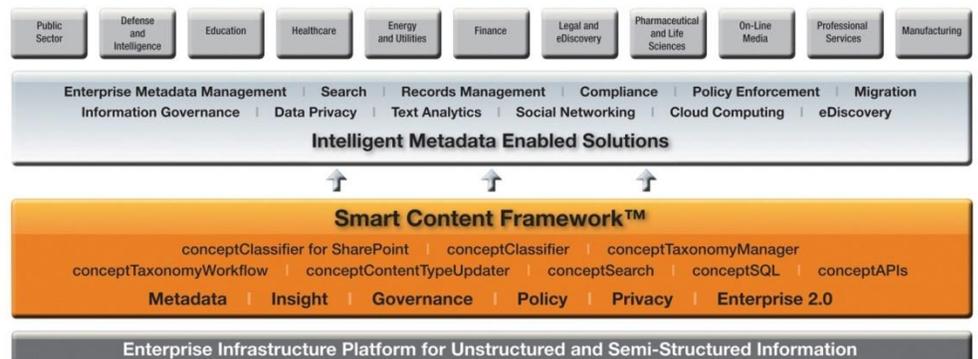
Energy companies are facing challenges in managing their content to meet corporate objectives. Not only does the approach have to suit an organization’s workflow and culture, it must be easily adaptable by end users, who are typically reluctant to change. With end user adoption cited as the single most critical failure point in many line-of-business applications, tools that can automate the process can significantly reduce non-compliance issues and improve the quality of the information for retrieval and re-use. In cases such as litigation support and eDiscovery that cause unexpected disruptions to business, significant manpower and manual methods to identify all relevant information is often required. Managing information as an asset encourages its collection, dissemination, and sharing, and ultimately reduces the cost of business operations.

This White Paper explores the use of Concept Searching’s Smart Content Framework™ and intelligent metadata enabled solutions, and how this enterprise infrastructure for unstructured information governance reduces risk and ensures compliance. The framework addresses information complexity, and suggests solutions that can effectively increase information transparency and provide the ability to manage content, regardless of where it resides. By providing immediate and actionable visibility into all relevant content, decision making can be improved, risk minimized, and costs reduced.

A Single Platform to Capture, Use, and Re-use Knowledge Assets

Most companies purchase software to achieve a single solution. The trend is changing, where companies are seeking one integrated solution that enables stakeholders to collaborate across functions, locations, and with external stakeholders.

Before organizations can maximize the use of content assets, a framework needs to be in place. From there, opportunities to improve a variety of challenges can be achieved. The Smart Content Framework™, developed by Concept Searching, outlines the building blocks that need to be in place for organizations to harness the power of their information capital. The framework and the technologies provide the ability to transparently identify and tag content with semantic metadata, and then classify it to organizational taxonomies aligned to business goals.



Regardless of size or industry, any organization that places high value on knowledge assets or faces strict regulatory guidelines can benefit from this unique solution to capture the meaning of content and use semantic metadata, to not only proactively manage content but also deploy intelligent metadata enabled solutions that are integrated with the organization's information governance policies. The outcome achieves quantifiable business objectives, while mitigating corporate risk and reducing costs.

This enables the effective management of content and results in:

- Improved search
- Automatic records identification
- Information security protection
- Intelligent migration
- eDiscovery and litigation support
- Real-time and relevant collaboration and productivity enhancements
- Text analytics

Compound Term Processing

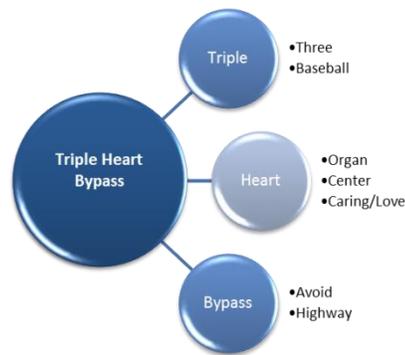
The Concept Searching technologies include a Service Oriented architecture (SOA) based search and classification technology, a browser based taxonomy management technology, and a tightly integrated feature set that operates with any platform.

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.

Instead of identifying single keywords, compound term processing identifies multi-word terms that form 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.

“Sound information governance practices and tools would enable organizations to align their data retention, acceptable use and communication, data privacy, records management, and information security policies, processes, and technical controls.”

Worldwide Governance, Risk, and Compliance Infrastructure 2010–2014

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 identify very specific criteria for business applications that use the metadata. The end result is an enterprise metadata repository containing a rich set of intelligent metadata that reflects the unique terminology and vocabulary of the organization. Concept Searching is unique among all commercially available technologies in incorporating a compound term processing engine.

The Role of Metadata

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.

Creating metadata repositories and taxonomies that are optimized for an 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 useable knowledge and delivering insight into content.

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 is 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.

Minimizing Risk in the Energy Sector

The energy sector faces a broad range of risks, from economic, environmental, geopolitical, societal, and technological. Managing risk in many cases is nearly impossible, as uncontrolled events such as process failures, hurricanes, terrorism attacks, and even global pandemics can disrupt a global organization without any warning.

At a more basic level there is also the process of managing risk on a day-to-day basis. This would include managing known risks and identifying unknown risks that can be proactively managed. In the realm of unstructured and semi-structured content, most organizations are highly vulnerable because content assets remain largely unmanaged and change rapidly. Risk is increased during process failures, as identifying, securing, and researching have high manpower requirements, and can prove very costly unless the organization can quickly mitigate or correct the impact from the failed process.

With a potential global workforce and extensive partnerships, accessing unstructured corporate knowledge that is relevant, timely, and accurate is a business necessity. The increasing need to share information must be aligned with security and information governance practices. The goal is to maximize profitability. To do this, energy companies must effectively manage a complex infrastructure of knowledge assets, employees, suppliers, and partners. Multiple cumbersome processes create bottlenecks, drive up costs, and slow the discovery, production, and distribution of product.

Organizations in the energy sector face considerable risks from unmanaged information. The issue stems from the lack of an enterprise content structure that can be used to address the diversity of stakeholders, regardless of where they reside and the specific knowledge they are seeking. Information Lifecycle Management processes and Enterprise Content Management systems may be in place, but there is still no visibility of understanding the concepts found within unstructured data that exist in a variety of formats and are ingested into an organization from different sources.

The types of challenges include:

- Inability to manage content existing in global locations, on diverse repositories, and replicated across various silos
- Inability to find trusted and relevant information regarding health and safety, asset maintenance, compliance guidelines such as OSHA, key information for decision making

“Gartner predicts by 2016, 20% of CIOs in regulated industries will lose their jobs for failing to implement the discipline of information governance successfully.”

-
- No capability of ensuring accurate records management, regulatory compliance, or improve eDiscovery, and litigation support processes
 - Potential weaknesses in compliance, information governance, privacy, and cyber security protection
 - Multinational content asset protection, authorization to assets, and compliance
 - Management of project risk and increased business complexity
 - Lack of real-time collaboration capabilities, vocabulary normalization, and the inability to ensure that the most relevant and recent information is available internally and with partners and suppliers
 - Changing composition of workforce and the inability to capture digital intellectual assets
 - Ability to scale, process, and manage petabytes of content

Although diverse challenges, they impact the entire enterprise from an infrastructure level down to a specific business process. In almost all instances, compliance is at the forefront, and all vital documents, as well as content throughout the enterprise and across enterprise boundaries, must be identified to minimize risk. Inundated with terabytes of data, both structured and unstructured, the ability to process and access the data and a better way to manage knowledge assets should be a priority.

Managing Content, Regardless of Where It Resides

All content that is ingested from a variety of sources, such as scanned or faxed content, content from other web sites, social feeds, blogs, or diverse repositories often lacks relevant intelligent metadata, rendering information assets unusable and inaccessible in the search, collaboration, and even the decision making process.

Leveraging information capital regardless of where it resides encourages the use of legacy content and the accumulated organizational knowledge that is often hard, if not impossible to find. Using intelligent metadata generation and taxonomies the organization has the ability to aggregate relevant content to gain insight into past projects, best practices, and access the human expertise that has been recorded in all forms of unstructured content.

Collaboration among internal and external stakeholders is becoming a valuable business tool and in some cases a necessity. The ability to make new or changed information available in real-time to stakeholders globally reduces project risk for all parties. All participants have access to the most relevant and up-to-date information in a consistent vocabulary, that reflects the internal nomenclature but easily understood by all.

The ability to capture concept based metadata and retrieve relevant search results from within an organization and diverse applications is the real currency of interoperability. Providing syntactic as well as semantic metadata delivers the ability to represent and share the meaning of content in an unambiguous manner.

Within the energy sector, time is often a critical success factor. The optimal solution, as described in the Smart Content Framework™ approach, is to build and deploy an enterprise metadata repository that serves as the focal point for managing unstructured content, regardless of the end purpose. Once deployed, an organization can reduce time, effort, and costs to address other business processes such as records management, data privacy, and migration. One advantage of the technology is that it was designed for use by Subject Matter Experts, not the IT team, highly trained information scientists, or highly priced consultants.

“By itself the search function has limited value. The real value of search and information access technologies is in the ongoing efforts needed to establish effective taxonomies, to index and classify content of all kinds, in order to provide meaningful results.”

Tom Eid,
Research Vice President
Gartner Group

The Role of Intelligent Search

With the explosion of unstructured content and its unabated growth, enterprise search has become inadequate, not only for the knowledge worker but also renders relevant content inaccessible when applying it to finding trusted and relevant information, such as asset maintenance, compliance guidelines like OSHA, up-to-date manuals, and vital information that is used in day-to-day business activities.

Specifically in the energy sector, companies typically have a wide variety of constituents that need access to content to meet different needs. Internal vocabularies are often specific to the organization and may not be easily translated by personnel outside a particular community, rendering the content highly ambiguous. Further complicating matters is that within a company there may be varying solutions for identifying and storing electronic documents. The inconsistency of these systems hampers the ability of users to find relevant information, specifically when searching across multiple silos of content within the company. Although knowledge workers need unified and universal access to information, at a more granular level they need to be able to find exactly and only the content they need.

The fundamental problem has always been, and still is, the inability of end users to consistently and accurately assign or create metadata. While enterprise search technologies may differ, the end result has been the same. Organizations are still facing the issues of wasted time, unproductive knowledge workers, and the inability to find business critical information. The need for a robust search solution is not an option or an afterthought, but is now a key business requirement.

conceptSearch, or any search engine, can use the automatically generated semantic metadata that has been placed in the search engine index to identify the concepts the end user is seeking. Presenting relevant information to different stakeholders and delivering effective search results is further enabled via taxonomy based or faceted based navigation. The knowledge worker controls the search experience and the search results present facets of documents grouped together based on the concepts identified. This extends the search process, as documents will be grouped by concept and will assist the knowledge worker in offering content that may not have been found. Using this approach, from an end user perspective, knowledge workers can locate pertinent information from an individual viewpoint, without knowing exact search terms to use.

This unified view and access to relevant information, from dispersed silos within the organization or external sources, can reduce the volume, cost, and time traditionally required to retrieve and find relevant content.

Intelligent Records Management, eDiscovery, and Litigation Support

Records management, eDiscovery, and litigation support can be the unknown Achilles' heel in many organizations. Without sound unstructured information governance policies, corporate risk escalates. In global organizations, compliance and mandates can vary widely from country to country and even in the US from state to state. From an eDiscovery and litigation support perspective, emails, tweets, and even any comment posted by an individual employee to an external site is now viewed as a corporate record and should be treated as such.

Effective information governance as a corporate responsibility must involve all the stakeholders, to define the appropriate policies and the least onerous approach for implementation.

“It is simply not realistic to expect broad sets of employees to navigate extensive classification options while referring to a records schedule that may weigh in at more than 100 pages.”

Forrester Research/
ARMA International Survey

“Malicious insiders account for 10% of the costs associated with a cybercrime and represent 38% of the types of attacks.”

Ponemon Institute
2012 Cost of Cyber Crime

With compliance requirements changing rapidly and end user adoption cited as the biggest source of failure in records management, there is a great need to simplify and manage the process. As an intelligent metadata enabled solution, the outcome 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.

This proof of compliance and data protection can only be accomplished through the automatic enforcement of organizational policy, which ensures consistency and can be implemented transparently, resulting in improved record keeping, monitoring, lifecycle management, and auditing processes.

Regulatory guidelines associated with records management, information security, and e-Discovery drive the requirement for workflow. Organizations without automated processes that enable records declaration, data transparency, and information security find themselves at increased organizational risk when it comes to storing, preserving, securing, controlling, and the inadvertent exposure of confidential information.

Concept Searching accomplishes this at a tactical level by creating a taxonomy that mirrors the records file plan and organizationally defined descriptors. Uploaded or ingested documents are automatically tagged, and if the metadata is associated with an information governance policy, the appropriate record retention codes are automatically applied and the documents moved to the Records Management application or any defined repository.

Data Governance, Privacy, and Cyber Security

Although most data base applications typically provide comprehensive security and use rights, unstructured content is much harder to manage. Confidential or privacy data exists in emails, documents, spreadsheets, PDF's, presentations, instant messaging, even tweets. With the flow of unstructured content being transmitted far beyond the confines of the organization, it is increasingly difficult to identify unstructured content that could increase organizational risk if exposed. Rectifying data breaches and publication of confidential information, whether internal or external, is not only extremely costly but negatively influences the image and brand of the organization.

What are the typical challenges?

- Lack of tools to identify all possible privacy data exposures at the time of content creation, modification, or ingestion
- Lack of end user compliance to segregate content from the network and ensure that uploaded privacy data is not available for general access and protected accordingly
- Lack of governance to enforce the accurate metadata tagging of documents based on content by end users
- No standard process that addresses all aspects of data privacy that are unique to the organization
- Inability to automatically identify privacy and confidential data risks in real time
- Inability to ensure protected data assets are subject to portability and security controls

Most security applications will address the identification of information, such as a social security number. conceptClassifier operates stand alone or in adjunct with third party security solutions and provides organizations with the ability to rapidly develop a taxonomy that contains unique vocabularies and descriptors that are specific to the organization's security concerns. The sources of risk included in the identification are:

- Document management systems and document libraries
- File servers, email servers, fax servers
- Scanned content
- Unstructured content generated through collaboration among internal and external sources

As content is created or ingested, the content containing the organizationally unique vocabularies will be automatically identified and routed to a secure repository for disposition. conceptTaxonomyWorkflow, an adjunct product, is typically implemented as a strategic and tactical tool, managing privacy as well as migration activities and content type application across multiple SharePoint farms, as well as from diverse repositories and in non-SharePoint environments.

“Only 50% of oil and gas organizations have a documented and approved information security strategy in place, less than the cross-industry average of 58.7%. This is clearly an area of action for the industry, especially considering that 45% of oil and gas industry respondents do not know how many security events have occurred in the past 12 months, and 50% of them do not know the nature of those events.

IDC Energy Insights

Multinational Content Asset Protection

Energy organizations often cross global boundaries, and additional security and compliance mandates are critical to the process of managing unstructured information. For example, the Patriot Act allows the US government to subpoena all data stored within the country, the EU Data Protection Directive does not allow personal information to be transferred to any outside country, the Massachusetts Breach Law specifies that citizens' private information must be protected and has specified strict guidelines around storage, access, and transmission of personal information. Different regulations also apply as to where the content is actually stored. The energy sector is already well aware of the challenges in compliance with multinational content asset protection. Not only must they address changing compliance mandates, they must also react in a timely manner to prove compliance.

Within the Smart Content Framework™, the deployment of intelligent enabled metadata solutions is easily extended to address specific challenges surrounding any application that requires the use of metadata. In the area of compliance, a taxonomy can be developed to address the compliance mandates. For example, one client has implemented over two thousand information governance rules to accurately identify, process, and classify content using conceptClassifier and conceptTaxonomyWorkflow. One of the key benefits in this scenario was the ability to rapidly develop the taxonomy, as well as make changes to address changing compliance regulations.

Vocabulary Normalization

Unlike many solutions, Concept Searching supports controlled vocabularies, an important consideration for any global organization. For those organizations that cross country and language boundaries, vocabulary normalization imposes order to aid in finding relevant information, reduce the duplication of effort when information can't be found, and increases the understanding of information between different languages and even between different divisions who have divergent vocabularies. These controlled vocabularies improve the accuracy of search results by removing the inherent ambiguity in natural language. For example, the word “football” can apply to both soccer and rugby. Therefore, a search for “football” will retrieve documents that are about several completely different sports. Or, the use of the word “resume” as opposed to “CV”.

A controlled vocabulary resolves this problem by metadata tagging in a way that the ambiguities are removed. The end result is a controlled vocabulary that is consistent with the organizations nomenclature, increasing the accuracy of search and clarity of the results in a way that it is understood by the end users.

Intelligent Collaboration, Reducing Project Risk and Business Complexity

The need to effectively communicate and collaborate with not only colleagues but also a wide range of external partners is increasingly important and valuable for energy companies. Projects can consist of geographically distributed teams that require real-time distribution of updated or new content, as well as the ability to collaborate on projects. The end goal of collaboration is the uniting of a flexible workforce while capturing best practices and knowledge, to reduce risk and become part of corporate memory for future access. Collaboration can also include the use of blogs and wikis to further disseminate relevant information and encourage the contribution of individual knowledge to the corporate knowledge base.

Companies and individuals are increasingly recognizing the inefficiencies of using email as a collaboration tool, given how poorly it performs in situations requiring collaborative work on single documents in situations such as project management, acquisitions, real-time response to risk, and general administration. The ability to share information across global locations in real time reduces the possibility of error, reduces redundancy in emails, and locates the most current document. This enables better decision making, facilitates project management, and can increase innovation.

For example, at a functional level for project management, projects can be accelerated by aggregating all relevant content, reducing validation time of the correct version of documents and document handover. This reduces the time spent organizing all appropriate documents for projects, and eliminates resource time to assemble and manage all documents, which results in fewer errors and the elimination of incorrect documents that can take the project off track.

Effective collaboration can also achieve benefits to an organization in applications such as project collaboration, awareness of organizational knowledge, employee induction and training, expertise location, communities of interest, collective intelligence, and innovation management.

The Concept Searching technologies provide improved collaboration outcomes by providing insight into content; grouping similar users, concepts, and content together; identifying people with expertise, knowledge or interest in a topic; and protecting and securing confidential information from unauthorized participants. The end result is consistent understanding of the value and context of information. They also provide confident cross-organizational decision support capability, shared knowledge, and the enterprise-wide availability of knowledge, to increase organizational performance.

Changing Workforces and Inability to Capture Digital Intellectual Assets

Many industries are facing the loss of knowledge workers and are challenged with capturing the inherent expertise gained from years of experience. Energy companies should be looking for ways to capture the expertise and knowledge so it does not become a lost asset to organizations as workforces transition. Organizations need to capture the existing knowledge, and knowledge workers need to be able to access these intellectual assets to collaborate globally with diverse stakeholders, find organizational resources, identify experts, and leverage business insight to make informed decisions.

“Over the next 10 years more than 50 percent of the energy workforce will be eligible for retirement, causing decades of operational, financial, engineering, and management experiential knowledge to be lost.”

Microsoft

This is not an easy task and the approach will vary by organization. For example, this approach can be used to integrate with social networking to find human resources with a particular expertise and knowledge. The result is the ability to deploy sophisticated knowledge applications to find relevant content by project, research, documentation, that is associated with a specific person. This requires the automatic culling of information to meet the requirements of the individual end users through the delivery of relevant content in the right context, as well as to the right people at the right time.

The primary benefit is the ability to foster collaboration and knowledge sharing, either from content, or people expertise. This can be an extremely useful tool within the energy sector facing a surge of retiring baby boomers to transfer knowledge and expertise.

Migration

Organizations are requiring more sophisticated techniques to ensure compliance objectives are met and a typical loop hole is in the migration process. Simply moving documents from one repository to another 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.

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 straight forward. However, from an information governance approach, mass moving content results in the same problem of mismanaged content.

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.

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. If manually processed, the security rights of the documents as they are moved to their new location must be applied. General migration tools cannot safeguard document confidentiality because they do not make intelligent taxonomy workflow decisions based on the text content of the document. If this security profiling is not performed during migration, then many of these documents will be easy to surface using enterprise search, breaching the relevant document security obligations.

Using the taxonomy workflow process, these documents will be safely routed to the record application, or some other appropriate secure location with the correct access rights, protecting and preserving documents during the migration process. Information governance best practices should be applied to the migration of unstructured content. This also provides organizations with a highly effective way to clean up the irrelevant or unnecessary documents, as well as to identify records that may not been declared or have potential privacy exposures.

“At the 2012 Compliance, Governance and Oversight Counsel (CGOC) Summit, a survey of corporate CIOs and general counsels found that, typically, 1 percent of corporate information is on litigation hold, 5 percent is in a records-retention category and 25 percent has current business value. This means that approximately 69 percent of the data most organizations keep can—and should—be deleted.”

“42 per cent of executives say that data analysis has slowed down decision-making, the vast majority (85 per cent) believes that the growing volume of data isn’t the main challenge, but rather being able to analyze and act on it in real-time. As organizations increasingly look to the output from analytics to automate decision making, data quality is seen as a major hurdle to this with two-thirds (67 per cent) claiming they struggle with data inaccuracy on a daily basis.”

“The Deciding Factor: Big Data and Decision Making” Survey
Economist Intelligence Unit Report
Commissioned by Capgemini

Big Data and Text Analytics

Even with a holistic approach, big data solutions typically parse content upon entry to the system or place the content in well-defined structures and impose metadata to ensure consistency and data types. Similar to placing a round peg in a square hole.

Before adopting a big data strategy specifically for unstructured data, organizations must first develop the infrastructure and implement technologies that can extend to the realm of big data, specifically text analytics, which is focused solely on unstructured and semi-structured content. This causes a shift from departmental to business process repositories (i.e. taxonomies), which encourages the sharing and integration of unstructured content assets, and creates access to relevant content on demand, based on security and user rights.

It is estimated that 80% of corporate data is unstructured (IDC) and growing at twice the rate of structured data (IBM). This factor can significantly influence decision making, based on the inability to access relevant content. On the other hand, the ability to harness the meaning of unstructured content in real time improves decision making, as organizations can proactively react with greater certainty to rapidly changing information. Energy companies must be able to access relevant content, regardless of how or where it was acquired and incorporate it into the overall big data solution.

What are the inherent issues with big data for unstructured content?

- Unstructured content is surpassing structured data and must be proactively managed
- Typical big data solutions do not provide the identification of nuances, sentiment, or knowledge assets found within unstructured content
- Inability to rapidly respond to information for decision making in real-time
- Lack of effective decision making as all the pertinent information cannot be found or easily extracted

The ability to process unstructured information for improved decision making begins with a taxonomy, or multiple taxonomies, to identify content in real-time that impacts decision making. This information can include any internal source of content, content that is ingested from other sources such as fax, scanned content, social tools, emails, and web sites. The advantage of conceptClassifier and the taxonomy development tools is that they can be easily deployed and managed by Subject Matter Experts. This provides business users a high degree of flexibility as the taxonomies can be rapidly changed to address new organizational requirements. This provides a 360 degree corporate view of unstructured content regardless of where it resides.

Ensuring that the right information is available to end users and decision makers is fundamental to trusting the accuracy of the information. Once this has been accomplished, the content can be managed and used to extend the realm of unstructured content, beyond improving business processes such as search, records management, and data privacy. Organizations can then find the descriptive needles in the haystack to gain competitive advantage and increase business agility.

Concept Searching’s intelligent metadata enabled solution and framework for text analytics analyzes and extracts highly correlated concepts from very large document collections. This enables organizations to attain an ecosystem of semantics that delivers understandable results.

Summary

Energy companies that proactively apply information governance to unstructured content can reap many rewards through improved organizational performance and reduced risk. Not only does structured content need to be managed, but energy companies urgently need to manage their unstructured content assets.

Concept Searching provides the ability to improve search, collaboration, compliance, records management, data privacy, migration, and text analytics. Leveraging the unique technology, based on compound term processing and with the ability to automatically generate semantic metadata, any application that uses metadata can be significantly improved. This eliminates the need to purchase applications that address a specific challenge, removes interoperability issues, and reduces costs associated with manpower and technology.

The primary benefits include:

- Improve search outcomes through the delivery of relevant content pertinent to the end user query
- Identify and secure potential confidential or sensitive information exposures
- Identify records according to international laws and regulatory requirements and automatically process accordingly
- Reduce human resources currently needed to process, organize, and manage unstructured content
- Automate business processes and improve decision making
- Cleanse and easily manage or dispose of content during migration
- Capture critical undocumented and documented information from an aging workforce
- Implement an enterprise information governance approach that maximizes collaboration, streamlines business process, improves operational efficiencies and the same time protect content that must be secured
- Reduce costs related to the manual application of business processes through automation of classifying unstructured content
- Increase the value of knowledge assets that already exist but can't be found and enable re-use and re-purposing of existing information

The Smart Content Framework™ and intelligent metadata enabled solutions provide an unprecedented level of power that rapidly turns business challenges into intelligently automated processes, spanning diverse organizational applications, and addressing the unique needs of organizations in the energy sector.

IDC estimates that return on investment for extending an organization's knowledge infrastructure ranges from a minimum of 38% to as much as 600%.

About Concept Searching

Founded in 2002, Concept Searching is now the industry leader in advanced semantic metadata generation, auto-classification, and taxonomy management resulting in intelligent enabled metadata solutions. The award winning products are the only statistical metadata generation and classification technologies that use compound term processing to generate intelligent metadata from unstructured and semi-structured data. The use of compound term processing, or identifying 'concepts in context' enables organizations to more effectively find, organize, and manage their information capital.

Concept Searching's Smart Content Framework™ utilizes a set of technologies and best practices that encompass the entire portfolio of unstructured information assets, resulting in increased organizational performance and agility. The intelligent metadata enabled solutions are being used to improve search, records management, protection of privacy data, migration, text analytics, and Enterprise/Web 2.0. The solutions are deployed in diverse industries, Fortune 1000 companies, and smaller companies with strict regulations in regards to compliance, data privacy, and information governance.

Concept Searching is a Microsoft Gold ISV and Microsoft's only managed ISV partner in the metadata enabled migration and compliance application sector. Although platform independent, the Concept Searching Microsoft suite of products uses a single code base able to be deployed in SharePoint 2007, 2010, 2013, and Office 365, providing clients with the choice of on-premise, cloud based, or hybrid environment to best meet their needs. The Microsoft products fully integrate with Windows Server 2008 R2 FCI, and the former Microsoft FAST products.

Headquartered in the US with offices in the UK, Canada 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 <http://www.conceptsearching.com> and follow us on [Twitter](#) and [LinkedIn](#).

Microsoft Partner
Gold Independent Software Vendor (ISV)

© 2013 Concept Searching



Americas
+1 703 531 8567
info-usa@conceptsearching.com

Europe
+44 (0)1438 213545
info-uk@conceptsearching.com

Canada
+1 703 531 8567
info-canada@conceptsearching.com

Australia
+61 (0)2 8006 2611
info-australia@conceptsearching.com

New Zealand
+64 (0)4 889 2867
info-nz@conceptsearching.com

Africa
+27 (0)21 712 5179
info-sa@conceptsearching.com

Marketing and PR
International: +1 703 531 8564
Europe: +44 (0)1438 213545
marketing@conceptsearching.com

Follow us on Twitter
[@conceptsearch](#)

www.conceptsearching.com