conceptSearching

21st Century Government

Mission Support through Technology Innovation

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

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

Twitter: @conceptsearch

August 22nd, 2012

© 2012 Concept Searching

Abstract

The rate at which technology advancements are made means that those government and industry organizations that lack the agility to embrace the advances at an equal rate are being left behind. The military has not kept up with technology capabilities. Challenges embedded in the institutional culture, and a cumbersome regulatory and acquisition environment, have contributed to the problem. The explosion of unstructured content and the use of social networking are infiltrating organizations of all sizes, and will continue unabated, presenting new challenges to the government, which is attempting to analyze, assimilate, and use the information. In addition, achieving organizational agility requires a cultural change that embraces technology which delivers an integrated information infrastructure.

This white paper provides an overview of how an infrastructure metadata framework and the use of flexible technologies address many of the objectives of federal agencies, by maximizing the use and transparency of unstructured content assets. This paper is based on the use of the **Smart Content Framework™** and Concept Searching technology deployments, at clients such as the US Air Force Medical Service, NATO Communications and Information Agency, US Army Medical Command, US Army Records Management and Declassification Agency, the Development, National Transportation Safety Board, Consumer Products Safety Commission, HHS, and the Concepts and Doctrine, Ministry of Defense Centre (DCDC), UK.

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 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.

© 2012 Concept Searching

Table of Contents

Overview **3**

The Business of Government **3**

Transforming Budget Challenges into Innovative Solutions 4
The Human Dimension 5

The Smart Content Framework™ **6**

Technologies 7
Automatic Semantic Metadata Generation 7
Auto-classification 9
Taxonomy 10
The Applications 11
Search 11
Records Management 12
Data Security 13
Migration 14
Security after the Migration Process 14
Collaboration and Enterprise Social Networking 14
Big Data 16
Cloud Computing 17

Summary **18**

Appendix A: Concept Searching Products & Technologies **19**

conceptSearch 19
conceptClassifier 20
conceptTaxonomyManager 20
conceptClassifier for SharePoint 21
conceptContentTypeUpdater 21
conceptTaxonomyWorkflow 21

Appendix B: Summary of Key Features and Capabilities in SharePoint **22**

About Concept Searching **23**

© 2012 Concept Searching

Overview

The ability to harness the power and leverage the value of unstructured content in the federal sector has become a vital component for modernization and improves the ability to systematically capture, organize, and retrieve content assets. The federal sector faces unique challenges, including the transformation of information and operational silos into shared repositories, budgetary constraints, and the impending loss of baby boomers. Although the federal government has matured in technology adoption and is beyond the portal stage, new objectives include modernization to support a global workforce, re-use of intellectual assets, data security, cloud computing, and collaboration.

The rate at which technology advancements are made means that those government and industry organizations that lack the agility to embrace the advances at an equal rate are being left behind. The military has not kept up with technology capabilities. Challenges embedded in the institutional culture, and a cumbersome regulatory and acquisition environment, have contributed to the problem. The explosion of unstructured content and the use of social networking are infiltrating organizations of all sizes, and will continue unabated, presenting new challenges to the government, which is attempting to analyze, assimilate, and use the information. In addition, achieving organizational agility requires a cultural change that embraces technology which delivers an integrated information infrastructure.

This white paper provides an overview of how an infrastructure metadata framework and the use of flexible technologies address many of the objectives of federal agencies, by maximizing the use and transparency of unstructured content assets. This paper is based on the use of the **Smart Content Framework™** and Concept Searching technology deployments, at clients such as the US Air Force Medical Service, NATO Communications and Information Agency, US Army Medical Command, US Army Records Management and Declassification Agency, the Development, National Transportation Safety Board, Consumer Products Safety Commission, HHS, and the Concepts and Doctrine, Ministry of Defense Centre (DCDC), UK[1].

The Business of Government

"Government today spends more on information technology (IT) overhead costs than on the direct costs of mission systems.

Government does not often leverage IT to make things simpler, generate economies of scale, or increase collaboration."

Mark Forman
IBM Center for the Business of Government
Spring/Summer 2012

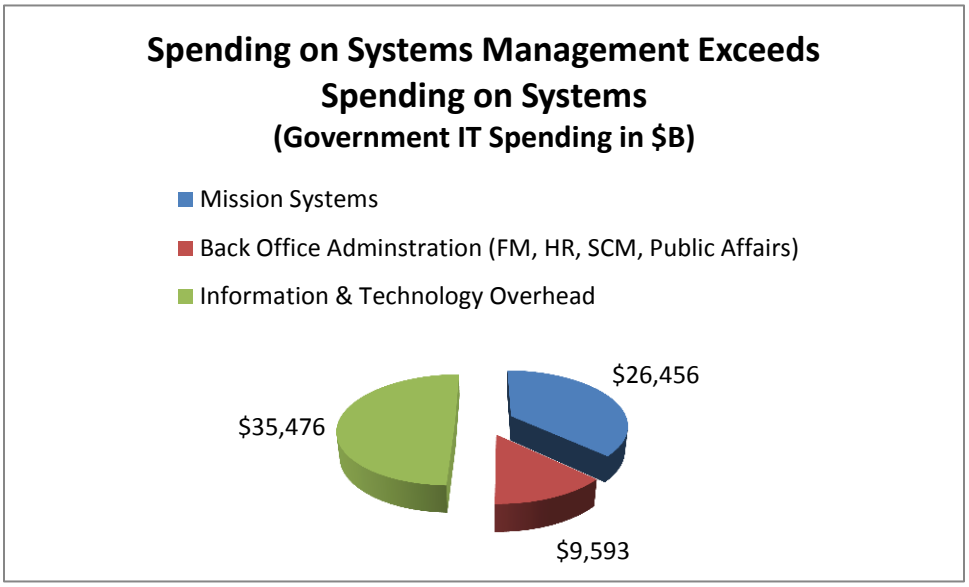
More so than the commercial sector, the government sector faces unique challenges. The impact of budgetary constraints requires thoughtful decisions and a strategic plan to modernize the current way of doing business and improve organizational performance. Quite simply, government needs to change how it does business and evaluate IT tools that are powerful and cost effective. Some of the challenges facing government entities include:

- Proprietary systems – developed with no effort to capitalize on technology innovations or interoperability with similar systems or capabilities
- Narrow optimizations – that limit the opportunities to reuse or adapt capabilities for other systems
- Closed designs – that prohibit visibility throughout the process and limits collaboration and community involvement
- Non-standard architectures

The graphic below illustrates the seriousness of the problem. The typical government approach has been to purchase and customize systems for individual programs at each agency. This approach has resulted in information and operations silos that end up constraining government's efficiency and effectiveness.

© 2012 Concept Searching

A question begging to be asked is why the government is spending more on managing IT than on the missions themselves? Part of the answer lies in the fact that over the past 20 years mission systems have become more complex and customized to meet very specific needs, resulting in information systems that have become even more siloed and complex. Another question is how would the government benefit from new technologies and how can the IT infrastructure be improved in alignment with budgetary considerations?



Source: US Government Accountability Office, Information Technology: OMB Needs to Improve Its Guidance on IT Investments, GAO-11-826 September 2011

Traditional approaches have delivered benefits in the previous environment of large monolithic systems development. These approaches are no longer viable options for today's environment, where rapid development cycles and capabilities can be quickly developed and deployed with the understanding that they will interoperate with the organization as a whole. Achieving interoperability in a net-centric environment is fundamental to realizing the full potential of transformation.

Transforming Budget Challenges into Innovative Solutions

Federal government agencies face unique obstacles specifically due to budgetary challenges and identifying sources of change for both policy and/or operational improvements is a priority. The real challenge is to transform these ideas into action. The objective is to not only achieve cost reductions, but to also capture technology and adopt innovative practices that make government more productive and transparent.

Often, the acquisition of a technology solution is to improve a specific function (i.e. search, records management, data security, etc.). An agency may have SharePoint, Oracle, or a host of platforms, none of which are used to meet their potential. Agencies should seek to maximize their current investment in what they have, by using technologies that can serve multi-purposes and integrate with any 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 solution that can be accomplished in steps. It is highly recommended that the metrics for success are relatively straightforward and can be achieved, even if incrementally.

© 2012 Concept Searching

"Another significant challenge involves maintaining our operational excellence in an era of budget reduction. The focus on budgets is a recurring theme here, but it's how we maintain all the good stuff that we want to do even in a time of fiscal constraint. This uncertainty has required us to change the way we do business and become smarter and more cost-effective. "

Erin Conaton
Former Undersecretary
US Air Force

A key component is the reusability of content. The approach should build upon an infrastructure framework where content can be used and reused to meet the needs of multiple and diverse stakeholders. This eliminates redundant development and moves from a large scale project approach to a more focused approach that solves a specific problem by leveraging the framework to deliver new capabilities. This also enables the organization to react rapidly to changing mission needs through real time access to new and existing content and eliminates the delay and costs of purchasing new systems.

The use of an enterprise metadata framework provides flexibility to the organization, is ideally platform independent, and enables the organization to implement governance policies and operations incrementally, with each step directly correlated to mission value. With current budget restraints, the implementation of specific functionality, for example identifying unknown data security exposures, helps align programs, funding and resources. This approach recapitalizes content assets and re-uses the technology assets against the most critical mission drivers, eliminating redundant investments. Since this approach can be done incrementally, the long term strategic view aims at balancing the intertwined dependencies against requirements, service levels, and funding, that result in increased organizational agility and improved mission performance.

The ability to manage unstructured content can contribute significantly to the process of transformation of the government towards a leaner and more cost-effective organization. It can facilitate communication and improve the coordination of authorities at different tiers of government, within organizations, and even at the departmental level. Benefits can enhance the speed and efficiency of operations by streamlining processes, lowering costs, improving research capabilities and improving documentation and record-keeping. However, the real benefit of lies not in the use of technology per se, but in its application to processes of transformation.

This framework also enables agencies to build applications that leverage unstructured content from outside the immediate domain, and build a content asset repository that provides multiple levels of granularity to satisfy the reuse of intellectual assets from multiple stakeholder views.

The Human Dimension

The human dimension is probably an organization's biggest weakness and its biggest strength. From an information access perspective, ultimately people are the end users of information and interpret the information available to make decisions. How they use the information presents the dilemma of how human behavior can enable or undermine interoperability. One of the goals of a net-centric environment is to share information, but end users need to know that information is available and they can access it, based on their security level and the security level of the content. Therefore, the question of usability must be considered at the beginning of systems planning.

During most systems planning, the end user is the most critical success factor, and this means usability design and subsequent training. Unfortunately, it has been proven that the end user is reluctant to change and typically will not enter metadata, will add erroneous metadata, or incomplete metadata. In applications such as records management or identification of PHI, PII, OPSEC the results can prove disastrous.

Net-centric interoperability, as well as information security, is entrusted to the same people. Without relevant information, knowledgeable, global decisions are unlikely. If the user's motivation cannot be changed, governance cannot be accomplished. In the end user's defense they are often over-burdened with cumbersome constraints to accurately tag content. But what they don't see is the value from a global perspective of

© 2012 Concept Searching

the effects of their changes or lack of changes. Security and interoperability are not foremost in their mind when they are involved in their daily routine and do not understand what is gained or what is lost.

Removing the end user from the tagging process removes the ambiguity in the tagging process. It also enables content to be related in a meaningful way without end user involvement. This enhances the value of knowledge far beyond the original design intent, and expands the value of content to be accessed and used by multiple stakeholders who may or may not have known the content existed. This also transforms the content into a knowledge asset as the data can be trusted, reliable, and is correct. This ability to generate metadata enables the sharing of information, but the lack of metadata guarantees that the data will be difficult to share, if at all. Comprehensive metadata that can capture the meaning of content improves decision making across the global organization.

The Smart Content Framework™

"Enterprise Content Management (ECM) is unfortunately running somewhere in the 50 percent plus failure rate. Failure being measured as an overall sense that we did not get what we paid for. We should not be surprised, considering most companies are going about ECM with the same approach that produced disappointing results before."

Dan Hooper
Principal
Integrated Services, Inc.

The ability to provide structure, share, re-use, and find information must support not only the anticipated but unanticipated users who need to discover and access content as well as develop new interoperable capabilities. The ability to retrieve information based on specific needs of the stakeholders requires a comprehensive metadata repository that can extract not only metadata but the concepts within the content. This ability combines and transforms content into usable information by expressing rich relationships in a thoroughly understandable manner. This eliminates stove-piped metadata approaches that prohibit data interoperability.

Agencies are looking for technologies that can solve more than one issue although each may have their own unique challenges. Recurring issues that surface are access to relevant and timely information, data security, records management, and compliance.

The three key issues are facing all agencies are:

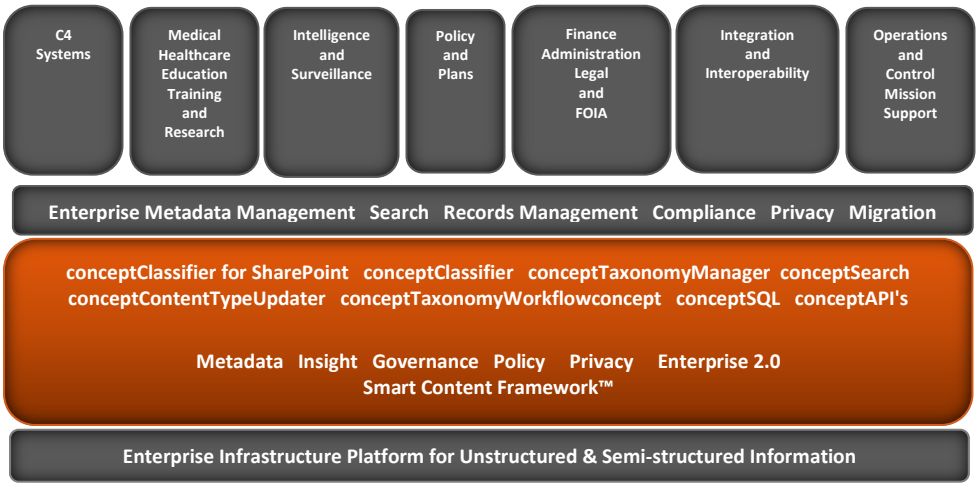
- Data Transparency - Accessibility and understandability in which users can comprehend information both structurally and semantically and readily determine how the content may be used for certain requirements and needs
- Information Assurance - Identification of unknown sensitive information (i.e. PII. PHI, and OPSEC) and controlling both access to and distribution of sensitive organizational information.
- Records Management - Records management and compliance with public law where records are declared in a consistent manner according to organizational guidelines, are routed to the appropriate repository for storages, and are destroyed once their preservation period has expired. Support operations and decision making with data that meets the need in terms of availability, accuracy, timeliness and quality and in ways that encourages horizontal as well as vertical sharing of information within the organization and with other government agencies, private sector organizations, and allied nations, consistent with national security and privacy requirements.

As a result, Concept Searching developed the **Smart Content Framework**™ which is an adaptive architecture strategy that provides the enterprise, infrastructure to mitigate risk, automate processes, manage information, protect privacy, and address compliance issues. The framework is a multi-disciplinary solution delivered through the Concept Searching technologies that encompasses the entire portfolio of information assets. What this means to an agency is the ability to deploy a metadata infrastructure

© 2012 Concept Searching

framework that can help alleviate the above issues, and provide the opportunity to focus on using unstructured content assets to drive transfer of knowledge, expertise sharing, and collaboration, all of which contribute to the agency's agility to perform, and improves decision making.

Underlying the infrastructure framework are technologies that provide the ability to transparently tag content, classify it to organizational taxonomies, preserve and protect information through the automatic identification of records and privacy data, and act as a migration tool. The building blocks that are used in the framework include: Metadata, Insight, Governance, Policy, Privacy, and Web/Enterprise 2.0. The subsequent applications have proven to solve a wide range of challenges.



Many federal agencies are using the **Smart Content Framework™** as a key component in their information governance initiatives. Since the technologies are flexible, the building blocks solve enterprise metadata management, search, records management, compliance, data privacy, migration, social networking, and big data challenges with a single solution. This leverages enterprise information assets, increases organizational performance and protects the organization's investment in technology.

Technologies

Concept Searching has developed innovative platform independent technologies that provide semantic search, auto-classification, and taxonomy management. The Microsoft suite of products is still the only solution that integrates natively with all versions of SharePoint, providing a migration path that leverages the organization's investment in SharePoint.

Automatic Semantic Metadata Generation

The discovery, collection, and management of metadata are essential for the integration of content across disparate systems. The primary issues are the lack of metadata associated with the content and the relating of content in one system that are similar to or equivalent to content in a different system. There is a growing need within the public and private sector to generate far richer metadata and manage it effectively to provide enhanced access to these resources by individuals.

The lack of a common and consistent way to describe or define unstructured content contributes to the inability to share information and results in duplication rather than reuse. This stovepipe approach has created boundaries around the information, making

© 2012 Concept Searching

it a challenge to share information or to make it available to internal users who require the information to complete their tasks.

Common problems in the federal sector are:

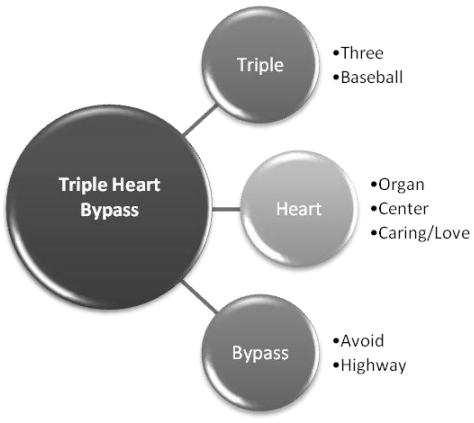
- Metadata is neither maintained nor shared across the enterprise and is inconsistent, incomplete, or nonexistent. It is therefore unavailable for document classification causing:
    - An inability to connect, categorize, and analyze information to improve organizational performance
    - An overabundance of content is surpassing the frameworks and controls in many government entities
    - The effort required to manually analyze and create a managed metadata term store is too labor-intensive, time consuming, and costly for most organizations to do well, if at all

Concept Searching solutions automatically generates semantic metadata based on the concepts within unstructured information. The generation of semantic metadata enables the agency to extract compound terms, acronyms, and keywords from a document or corpus of documents that are highly correlated to a particular concept or meta-tag.

Compound Term Processing automatically identifies the word patterns in unstructured text that convey the most meaning and uses these higher order terms to improve precision with no loss of recall. The algorithms adapt to each customer's content and they work in any language regardless of vocabulary or linguistic style. This capability is still unique in the marketplace.

By identifying the most significant patterns in any text, these compound terms are then used to generate metadata based on an understanding of conceptual meaning. This eliminates the requirement for an individual to read a document and subjectively apply metadata to that document. This ability to identify 'concepts in context' through our unique *compound term processing* technology eliminates inconsistent or non-existent tagging processes, and overcomes different publishing conventions that may exist within the agency.

The graphic below illustrates the results using compound term processing. In this example, the search term 'triple heart bypass' would typically return results containing each of the words in the result set. Using compound term processing, the technology understands that the query (i.e. triple heart bypass) is a concept and will retrieve documents based on the compound terms.



The search results using compound term processing will return documents even if the exact terms are not contained in the document (i.e. coronary artery surgery, heart surgery)

The metadata generation occurs in real time as content is either created or ingested, and can be immediately available to a wide variety of applications. Manual metadata

© 2012 Concept Searching

generation among stovepiped systems add unacceptable cost and time, that compromise effective operations. The ability to overcome incompatible semantics and meanings provides a method of bridging vocabularies across global boundaries.

The benefits to the agency include:

- All content is automatically meta-tagged, regardless of where it resides as it is created or ingested, making it available in real time to users and applications
- Eliminates inconsistent, incomplete, erroneous, or non-existent metadata, which makes the content unusable to the organization
- Eliminates the time to find and/or recreate content (15% of the average knowledge worker is spent duplicating information, 25% of their time searching for information, 40% cannot find the information needed to do their jobs. *IDC*)

Auto-classification

The manual process for classifying documents is both time consuming and labor intensive. Typically, a person associated with the program under which the document was produced must review the document to be classified and search through it to identify material called out in the classification guidelines document produced by the program office. This process can be complicated, due to the sometimes complex conditions, which can lead to a classification decision. For example, certain documents become classified when a series of different technical parameters are present in the document, even though each parameter by itself may not be classified. The manual review process for proper document markings of the security classification may take a few hours to several weeks, depending upon the document length and complexity of the classification guidelines. The key issues that prohibit accurate classification include:

- 80% of Enterprise Data is unstructured *(IDC)*
- Content is not tagged correctly, if at all
- Inability to manage existing content
    - 60% of stored documents are obsolete *(eLaw)*
    - 50% of documents are duplicates *(Equivio)*
- The cost of manually tagging one document is $7.00 (*Hoovers*)

Automatic and/or manual classification to one or more taxonomies prevents the end user from making potentially erroneous classification decision. The technology does provide the ability for knowledge workers, with the appropriate security, to classify content in real time. Content can be classified from within SharePoint and also from diverse repositories, including file shares, Exchange Public Folders, and websites. All content can be classified on the fly, in real time and classified to one or more taxonomies.

The benefits to the agency include:

- All content is correctly indexed and meta-tagged as it is created or ingested
- Elimination of manual tagging unless specifically authorized
- Normalization of content across functional and geographic boundaries
- Integration with the organization's enterprise search engine
- Ability to apply policy consistently across diverse content repositories
- Elimination of costs and errors associated with end user tagging

© 2012 Concept Searching

"The metadata infrastructure provides the critical glue that binds the information infrastructure to the underlying IT infrastructure. Sound information governance practices would take advantage of the metadata infrastructure, to ensure that content and data are managed consistently and adhere to written policies, across on-premise and cloud based environments."

IDC
Digital Universe Study 2010

Taxonomy

One or more taxonomies are critical to any organization that places high value on content/intellectual assets. Taxonomies reflect the operational and business processes that represent the inter-relationships of legacy content, written down as natural language, and represent the intellectual assets found in the expertise of knowledge workers. The taxonomies provide the flexibility to support multiple mission value chains. Coupled with workflows, a rules based approach is an extensible solution that allows the dynamic association with generic content processing sequences. Each taxonomy extracts from the generic sequence to the specifics associated with operational or business needs.

For example, a taxonomy can be created that identifies any confidential information or privacy data that is contained in a document. These descriptors are defined by the organization therefore unique to specific operational needs. As content is created or ingested if the content contains one or more of the descriptors it will be processed differently and moved to a secure repository for disposition.

Taxonomies by nature are organic as they reflect the current state of knowledge by an organization as content is continually changing. Concept Searching's taxonomy development tools address the fluidity of content changes to ensure that the taxonomy remains current and is easily managed. Providing both automatic and manual classification, Subject Matter Experts (SMEs) can utilize rich features such as node weighting, the ability to see the 'concepts in context', the ability to search the corpus in real time, auto-clue suggestion for categorization, and instant feedback on the impact of changes. 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 in government entities by up to 80%.

Some of the benefits that the agency can achieve through automatic metadata generation, auto-classification, and easy-to-use taxonomy tools include:

- Automatic declaration of documents of record based on vocabulary and retention codes
    - Automatically changing Content Types (in SharePoint) and route to the Records Management repository
    - Identification and declaration of records that were not previously identified
- Automatic identification of unknown data privacy exposures
    - Automatically changing Content Types (in SharePoint) and route to a secure repository
- Smart Migration of legacy content from diverse sources
    - Identification of records not previously declared
    - Identification of potential security exposures
    - Elimination of duplicate documents
    - Identification of documents of no value and those that should be archived

© 2012 Concept Searching

- Migration of content to an organizational structure for re-use and re-purposing
- Enable concept based searching
- Notify users of high value content
  - Provide guided navigation via the taxonomy structure (i.e. concepts)
  - Go beyond dynamic clustering with conceptual clustering based on the taxonomies
  - Vocabulary normalization across boundaries

"It is both reasonable and essential to create an environment where every network user can get the information he/she needs, when he/she needs it, through heterogeneous integrated network and information cores. Such cores must be trusted, secure robust and ubiquitous. Net-centric information cores must be designed to evolve with technology and mission changes."

"Creating an Assured Joint DOD and Interagency Interoperable Net-Centric Enterprise" March 2009. Office of the Under Secretary of Defense for Acquisition, Technology and Logistics

The Applications

Federal agencies are facing challenges in managing their content to ensure compliance, reduce organizational risk, and increase preparedness for meeting mission objectives. Not only does the approach have to suit the 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 reuse. Managing information as an asset encourages its collection, dissemination, sharing, and ultimately reduces the cost of business operations.

Within the **Smart Content Framework**™ federal agencies are using the technologies to:

- Improve search outcomes
- Automate Records Management
- Ensure Compliance
- Identify and secure confidential content assets
- Automate migration
- Provide structure to collaboration and social networking tools
- Manage big data to improve organizational performance
- Improve processing time for FOIA and eDiscovery

The results remove the ambiguity in content for a wide range of applications from incompatible data formats as well the incompatible semantics and meanings between applications. For the organization, the translation overhead between stove-pipe systems is removed, reducing costs, time, and improves the operation of applications.

"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

Search

The ability to capture concept based metadata and retrieve relevant search results from within the organization and diverse applications are 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.

The taxonomies align with an organization's mission and contain classes and class clues, allowing search engines to use extracted key words and phrases to identify main concepts. If a user knows the concept but not the correct key words or terminology, he/she can conduct a search to retrieve documents that closely relate to the clues provided by the end user. As an example, using the Medical Subject Heading Taxonomy (MeSH) developed by the National Library of Medicine and an end user conducts a search on "colon cancer", in addition to the directly related documents that will be found, the Concept Searching technology capability used with the MeSH taxonomy will

© 2012 Concept Searching

find related topics to include polyps, cancer learning, virtual colonoscopy, colonoscopy, and colorectal.

Government agencies may have a wide variety of constituents that need access to content to meet different needs. Internal vocabularies are often specific to that agency and may not be easily translated by personnel outside of a particular community rendering the content unusable. Further complicating matters is that within an agency 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 an agency, or inter-agency. 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. Using this approach, from an end user perspective, knowledge workers can locate pertinent information from his or her own individual viewpoint without knowing the exact search terms to use.

The technologies integrate with all Microsoft search products, in fact any enterprise search engine. The semantic metadata is used to populate the search engine index to enable concept based searching. Presenting relevant information to different stakeholders and 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 assists the knowledge worker in offering content that may not have been found. This unified view and access to relevant information from disperse silos within the agency or external sources, can reduce the volume, cost, and time traditionally required to retrieve and find relevant content.

Records Management

Within the SharePoint environment, conceptClassifier for SharePoint is being widely used to facilitate records management processes by eliminating end user non-compliance and inconsistent application of records retention codes, the primary cause of records management failure. The ability to automatically extract conceptual metadata based on the content, identify organizational descriptors associated with the file plan, and apply appropriate records retention codes removes the burden from the end user. The end user no longer has to decide the appropriate classification values. These values are defined by the agency to mirror the records management file plan and unique organizational descriptors.

"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

Most, if not all, of our customers in the federal sector are using SharePoint. In any organizations that use SharePoint, documents are often placed in the wrong location, have inappropriate metadata applied, and lack measures to control access and rights management for individual data assets.

Content types in SharePoint enable organizations to take advantage of the workflow capabilities that can enhance organizational performance while driving down costs. The only obstacle with content type applications is that individuals have to decide which content type applies to every document ingested by SharePoint. For organizations with large content repositories this is no trivial matter.

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

© 2012 Concept Searching

find themselves at increased organizational risk when it comes to storing, preserving, securing, controlling, and exposing information.

In SharePoint, semantic, records retention, and security metadata contained in the term store has to be applied to every data asset in order to ensure data transparency, records declaration, and to control access and apply digital rights management. Leveraging conceptClassifier for SharePoint organizations are able to automatically apply semantic, records retention, and security metadata.

This is accomplished by creating a taxonomy that mirrors the records file plan and organizationally defined descriptors. Uploaded documents are automatically tagged and if the metadata is associated with a specific Content Type, the Content Type is changed and record retention codes are applied for automatic migration to the Records Management System.

Data Security

The importance of data security is two-fold. Primarily security must be enforced at the end user and content level. Secondly, confidential data must be protected from unauthorized use either internally or externally. In a content sharing environment the appropriate security architectures must be defined but also the security of content assets at a very granular level. The organization's traditional security mechanisms and applications must be augmented with additional functionality to ensure that at the content and user level, security is enforced.

"Negligence is definitely a major cause of data breaches in government. Most of these people are not bad people. They don't have enough knowledge or they're rushing. They're under pressure to get something done. But they take the risk and put their organization in peril."

Larry Ponemon
Chairman and Founder
Ponemon Institute

Despite the fact that billions of dollars are budgeted each year by the federal government for Information Technology programs that identify security procedures, data shows that internal data breaches continue to occur an alarming rate (Davis & Waxman, 2006). Even more significant is the revelation that the federal government, under the current setup, is not fully aware of what data are being stored on federal computer systems and thereby, unable to determine exactly what data might be at risk (*Davis & Waxman, 2006*). Organizations used to take, or still currently take a crisis-security policy which is a reactive approach that can prove devastating to the organization. Within the government sector, security and protecting confidential content assets is not an option as the repercussions are too high.

Information Security should lie solely within the Information Technology department of an organization (*McConnell, 2002*). The high percentage of organizations adopting security technologies suggests that organizations may be relying too much on security technologies without accompanying changes in business processes, which take into account the 'people' aspect of the solution. This 'people' aspect has proven to be one of the most significant challenges that are responsible for data breaches.

Concept Searching's approach is to eliminate end user involvement in the process, unless specifically authorized. The solution augments traditional security products and compliance processes within an organization, by discovering where unknown privacy data (PXX) (i.e. any organizationally defined descriptors/content that has been defined as confidential) exists. Fully integrated with all versions of SharePoint, documents containing PXX are automatically identified, and optionally changed to a custom Content Type, routed to a secure server and made available to selected users using Windows Rights Management services for further disposition and analysis.

Fully customizable to identify unique or industry standard PXX descriptors, content is automatically meta-tagged and classified to the appropriate node(s) in the PXX taxonomy based upon the presence of PXX from within the content. Once tagged and

© 2012 Concept Searching

classified, the content can be managed in accordance with internal, regulatory, or government guidelines.

Migration

The issue of legacy data is a real challenge in government. Legacy data must be made available to the organization and it must also be discoverable. Data quality and data cleansing need to ensure the integrity of information. Migrating unstructured content can be a laborious and costly activity. Not a formal building block in the **Smart Content Framework™**, migration of unstructured content is a less used component of information governance.

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. All of these challenges make migration of unstructured content a process that requires thought and careful planning.

The ideal solution is to combine workflow capabilities and enable intelligent automatic classification decisions during and after migration. These decisions enhance organizational performance and drive down costs, but more importantly enforce corporate and legal compliance guidelines.

To migrate document collections effectively you need to search the text content of each document to determine its value. This classification must be done before you can make an intelligent decision about how to relocate items during the migration process. This cannot be done manually as the volume is too high, and the consistency

Migration must also consider the security of the documents as they are moved to their new location. There are two imperatives here; first, to respect the existing security status and apply the same security in the new location and second, to identify sensitive documents that may not currently be in a secure location. Assessing the security needs of these documents requires intelligent interrogation of their content, and then comparison to a number of relevant official taxonomies - PII, PHI, ITAR etc. If a document is automatically classified against one or more of these taxonomies, it must be given the appropriate security profile.

Security after the Migration Process

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.

"As a rule, an organization's knowledge and capability building depends primarily on its human and social capital."

Hitt & Ireland (2004)
Journal of Leadership & Organizational Studies

Collaboration and Enterprise Social Networking

Enterprise 2.0 is technology to bring people together and let them interact, without

© 2012 Concept Searching

specifying how they should do so *(Andrew McAfee, 2009)*. Another way of expressing this is Enterprise 2.0 supports the information organization: the social networks through which work often really gets done *(Rob Cross and Andrew Parker, 2004)*.

Collaboration and social networking is becoming inescapable. In an environment where content is exposed to knowledge workers they can pull relevant content when they need it. Alternatively users and applications can receive alerts when content in which they are interested in or have subscribed to is updated or changed. This also provides immediacy in information availability.

Social networking tools, that encourage collaboration, can link employees, partners, suppliers, and customers to share information, and 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. The tools have also resulted in generating a surge in unstructured content which remains unmanaged.

Enterprise 2.0 can be effectively used to create organizational networks to share knowledge and expertise. Many major players, such as Microsoft, have now joined the race for market share. New capabilities in SharePoint 2013 and the acquisition of Yammer will be changing the landscape. The use of social networking within the enterprise has not been widely accepted, but will become so as the technologies become more sophisticated and can be proven to deliver business value. For the government there are several inherent problems with enterprise social networking. For the most part, unstructured content in this scenario remains unmanaged. Any type of sensitive information needs to be protected; security issues must take into account the end user as well as the data asset. Social networking must also deliver results and not become a waste of time for end users who will eventually abandon the application.

Used correctly, the primary benefit is the ability to foster collaboration and knowledge sharing, either from content, or people expertise. This is can be an extremely useful tool within the government sector facing a surge of retiring baby boomers to transfer knowledge and expertise. Agencies should be looking for ways to capture the expertise and knowledge so it does not become a lost asset to the organization.

Knowledge is a corporate asset. Managing it within an Enterprise 2.0 application provides the ability to present relevant information to potentially different audiences, that effectively results in the sharing of the collective knowledge of the organization. A loosely organized, uncontrolled Enterprise 2.0 environment neither encourages relevant knowledge sharing nor does it drive a return on investment.

Given worsening agency budget situations, there will be a challenge to the way that agencies adopt new technology based services. This will affect adoption of collaboration technologies, social media, and social networking tools by federal agencies. When it comes to incorporating social networking and collaboration into federal operations it is expected to see continued pressure to adopt low-initial-investment solutions. A greater challenge will come with upgrading internal collaboration and social networking tools to internal agency operations. These will require more changes to internal business processes and this is where significant change management and cost challenges will arise.

For example, agencies and individuals increasingly may recognize 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 policy development, acquisitions, rulemaking, and general administration.

© 2012 Concept Searching

Enterprise 2.0 can also cause chaos. Issues of securing confidential data, maintaining security rights of knowledge workers, unauthorized use of sharing documents, and posting of information to public sites, all contribute to the issues that must be addressed. There are several excellent uses of social networking tools, used internally or externally in the organization. They can also achieve benefits to the 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 objective is to provide structure when implementing Enterprise 2.0. The Concept Searching technologies provide improved search outcomes by providing insight into content; can group similar users, concepts, and content together; identify people with expertise, knowledge or interest in a topic; and protect and secure confidential information from unauthorized participants. The end result is consistent understanding of the value and context of information. It also provides confident cross-organizational decision support capability and shared knowledge and enterprise availability of metadata knowledge to increase organizational performance.

Big Data

Big Data deals with structured data, semi-structured or unstructured data, and unstructured content. The first two items are the primary focus of the term Big Data. Unstructured content is typically pigeon-holed into a database which is not the best approach.

One of the fundamental problems is the view that unstructured content must be managed in databases for analysis, in the same way as structured and semi-structured data, which is not the right approach. 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.

Many organizations still struggle with the most basic aspects of managing unstructured content, which include free-form language, emails, documents, and social networking applications. The perceived lack of need for, or seemingly overwhelming challenges of, managing unstructured content has resulted in the inability to manage content and led to poor information governance practices. This has far more immediate and serious implications in terms of compliance and data privacy issues, which can lead to fines, sanctions, and loss of business.

Ensuring that the right information is available to end users and decision makers is fundamental to trusting the accuracy of the information. Once this trust has been established, the content can be managed and used to extend the realm of unstructured content to include massive amounts of information distilled and categorized by conceptual meaning. Organizations can then find the descriptive needles in the haystack to gain competitive advantage and increase business agility.

Concept Searching's technologies and framework analyze and extract highly correlated concepts from very large document collections. This enables organizations to attain an ecosystem of semantics that delivers understandable results. The valuable insight gained can be used to achieve mission objectives, improves the speed and accuracy of decision making, provides a global view of a subject, and perhaps more importantly, identifies the internal knowledge capital and expertise that exists but is rarely used because it cannot be found.

© 2012 Concept Searching

"The cloud will do for government what the Internet did in the '90s. We're interested in consumer technology for the enterprise.

It's a fundamental change to the way our government operates by moving to the cloud. Rather than owning the infrastructure, we can save millions."

Vivek Kundra
Federal Chief Information Officer

Cloud Computing

Many argue that government productivity is now directly tied to how effectively it uses IT. Government should take advantage of new approaches for rapid deployment of IT capabilities by acquiring IT as a service; now commonly referred to as cloud computing. Instead of new capabilities requiring large capital investments and years of sophisticated project management, today's cloud computing services improve agility, cost-effectiveness, responsiveness, openness, and results from government programs. Many cloud computing services are now widely recognized brands, such as Salesforce.com, Google, and ADP Payroll Services.

The 21st-century IT infrastructure is being built around cloud computing, enabling organizations to adopt a new productivity model. Cloud computing incorporates both a continuation of the long-term trend toward automation and commoditization of transactional processes and a newer, rapidly growing trend toward broader access to problem solving tools.

Cloud computing tools enable more rapid, high-quality problem solving to improve productivity. Studies show that problem solving improves when people develop ideas and then use tools to share those ideas in a collaborative environment. As a result of IT infrastructure changes made to implement the 9/11 Commission findings, there has been significant progress in the area of counterterrorism. Other examples include the Recovery Accountability and Transparency Board's use of transparency concepts and tools. Using these tools, citizens identified potential fraudulent behavior, yielding 7,600 complaints from the public that have led to about 1,650 investigations for fraud, waste, and abuse. Both of these advances resulted from the adoption of cloud computing approaches and tools for collaboration, data sharing, and analytics.

To address the changing technology landscape to incorporate the option of cloud computing, Concept Searching has developed a unique integration with Microsoft Office 365 to incorporate the ability to transparently tag and classify content from end users. The *Data Enhancement System* uses Concept Searching's conceptClassifier and conceptTaxonomyManager to automatically classify content to Office 365 to one or more taxonomies. This is all done in a secure environment including transmission using https and the SharePoint site security. Synchronized with the term store in Office 365 documents are automatically classified delivering the benefits of enhancing the organization's SharePoint farms in the cloud in the same way as those that are on-premise, and all done simultaneously. The benefits include the synchronization of taxonomies to multiple term store, enhancing other types of data sources using one standard set of rules, and augmenting the data using the organization's own unique vocabulary and tags.

© 2012 Concept Searching

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

Summary

The solution solves many of the challenges facing government entities. The infrastructure framework assists agencies in reducing costs and increasing productivity. Allowing knowledge workers to effectively query, use and re-use agency wide content improves the speed and efficiency of operations and information can be shared and leveraged throughout the business cycle. The elimination of inconsistent tagging and different publishing conventions across multiple content stores provides access to relevant content from internal and external sources.

For the agency significant benefits can be achieved by removing the ambiguity in content through the identification of concepts within a large corpus of information. Concept Searching's solutions can be the catalyst to improve access to unstructured information, encourage innovation, and deliver real benefits to government entities, their constituents, and stakeholders.

Transforming relevant information into actionable knowledge has three intuitively significant benefits. For leadership, they are able to rapidly organize their organization's explicit and implicit knowledge to facilitate more effective communication and decision-making. For staff, cross-functional operating units are able to push relevant information to interested persons, reduce process timeline, utilize untapped resources, and enhance outcome quality. For the organization, it avails contemporary and relevant information that assists and expedites task performance and decision making advancing individual and group performance via enhanced situational and issue-specific knowledge.

© 2012 Concept Searching

Appendix A: Concept Searching Products & Technologies

conceptSearch

conceptSearch is a unique, language independent technology and is the first content retrieval solution to integrate relevance ranking based on the Bayesian Inference Probabilistic Model and concept identification based on Shannon's Information Theory. Unlike other enterprise search engines that require significant customization with marginal results, conceptSearch is delivered with an out-of-the-box application that demonstrates a simple search interface and indexing facilities for internal content, web sites, file systems, and XML documents. Application developers experience a minimal learning curve and the organization can look forward to a rapid return on investment.

Because of the innovative technology, conceptSearch delivers both high precision and high recall. This is particularly important for organizations that need sophisticated search and retrieval solutions. By weighting compound terms (multi-word) phrases, instead of the typical single words, or words in proximity, the retrieval experience is significantly more accurate and relevant. Much more powerful than single word, multiple words, or Boolean searches, the ability for the search engine to identify concepts enables the organization to improve the search experience for a variety of users.

Key features include:

- Compound terms are extracted when content is indexed from internal or external content sources, enabling the delivery of greater precision of relevant content at the top of the search results.
- Relevance ranking display extracts from the documents based on the query and are returned to the user.
- Search refinement delivers to the end user highly correlated concepts that may be used to refine the search. Taxonomy browse capabilities are also standard.
- Documents can be classified into one or more taxonomy nodes, enhancing the precision of documents returned.
- In addition to static summaries, Dynamic Summarization, a modified weighting system, can be applied that will identify real time short extracts that are most relevant to the user's query.
- Related Topics will return results based on the conceptual meaning of the search terms used. Using the ability to generate compound terms in a search, for example, 'triple' is a single word term but 'triple heart bypass' is a compound term that provides a more granular meaning.
- Based on previous queries, or on extracts retrieved, end users can use the text to perform additional searches to retrieve more granular results.
- The product is based on an open architecture with all APIs based on XML and Web Services. Transparent access to system internals including the statistical profile of terms is standard.
- Easily customized for your requirements.

© 2012 Concept Searching

conceptClassifier

conceptClassifier is a leading-edge rules based categorization module providing our clients with complete control of rules-based descriptors unique to their organization. conceptClassifier provides an easy to implement and maintain categorization descriptor table through which all rules and terms can be defined and managed. This approach eliminates the error prone results of 'training' algorithms typically found in other text retrieval solutions.

conceptClassifier identifies as part of the indexing process, the categories that each incoming document belongs to. Each category is defined by a unique descriptor and is associated with key descriptive words and/or phrases held in the database.

Key features include:

- Rules based categorization module
- Real time classification of individual pieces of content aligned to business structures
- Automatically classifies documents to multiple nodes in multiple taxonomies
- Highly scalable, fast real time classification
- Classifier may be called via web services, or by other related applications
- Based upon identified and extracted concepts this approach has been proven to be more effective than keyword classifiers

conceptTaxonomyManager

conceptTaxonomyManager is a robust and powerful taxonomy management tool that is still unique in the industry. Developed under the premise that a taxonomy solution should be used by business professionals, and not IT or librarians, the end result is a highly interactive and powerful tool that has been proven to reduce taxonomy development by up to 80%.

Key features include:

- Automatic Conceptual Metadata Generation (Unique in Industry)
- Auto-Classification
- Taxonomy Clues used for scoring
- Automatic Clue Suggestion (Unique in Industry)
- Document Movement Feedback (Unique in Industry)
- Taxonomy Workflow
- Boosting Capabilities
- Distributed Taxonomy Management
- Auditing Features
- Industry standard formats and taxonomies such as OWL and MeSH can be easily imported as well as any organizationally defined taxonomy
- Platform Independent

© 2012 Concept Searching

conceptClassifier for SharePoint

conceptClassifier for SharePoint is the only industry solution that delivers automatic identification and extraction of concepts from within content as it is created or ingested, provides intelligent auto-classification, and enables enterprise class taxonomy management fully integrated with the SharePoint server environment and the only solution that runs natively in the term store. conceptClassifier for SharePoint is optimally delivered as a complete platform with all standard features included. It is fully integrated with SharePoint 2007 (MOSS), SharePoint 2010, SharePoint 2013, Microsoft Office, Windows Server 2008 R2 FCI, and all SharePoint search products.

(For more information on the features and functions of conceptClassifier for SharePoint, please see Appendix B.)

Optional Components

There are two optional components for conceptClassifier for SharePoint that can be used to more fully leverage the organizations SharePoint investment.

conceptContentTypeUpdater

conceptContentTypeUpdater works with conceptClassifier for SharePoint to bypass manual processes with the SharePoint 2010 Content Organizer and automatically apply correct content types based on managed metadata properties.

Unlike commercial and private sector deployments of SharePoint, federal deployments have been found to have over 70,000 site collections making it virtually impossible for "Central Admin" players to manage content down at the functional level. This patent pending add-on product is deployed at the operational and tactical levels to provide site collection administrators with the ability to independently manage access, information management, information rights management, and records management policy application within their respective business units and functional areas, without the need for IT support or access to enterprise wide servers.

Automatically generated semantic metadata automates the tagging of content and triggers the content type update, which in turn applies actions on the content, thereby automating and enforcing the application of policies aligned to the organizational goals.

conceptTaxonomyWorkflow

conceptTaxonomyWorkflow is an optional Concept Searching component that can perform an action on a document following a classification decision when the criteria are met. conceptTaxonomyWorkflow is not a general purpose workflow engine, so does not compete with Microsoft's Workflow Foundation or K2 blackpearl/blackpoint. Unlike Concept Searching's conceptContentTypeUpdater, the workflow source type works in all versions of SharePoint as well as for all document types, FILE document types, and HTTP document types.

conceptTaxonomyWorkflow is typically used as a strategic tool for managing migration activities and content type application across multiple SharePoint farms. The module delivers workflow capabilities that enable intelligent automatic classification decisions during and after migration. These decisions enhance organizational performance and drive down costs, but more importantly enforces corporate and legal compliance guidelines.

© 2012 Concept Searching

Appendix B: Summary of Key Features and Capabilities in SharePoint

Feature	Taxonomy Component	Explanation
Ability to build taxonomies/ontologies	Yes	
Auto-classification	Yes	
On the fly classification of internal and external content	Yes	Can also be scheduled if preferred.
Automatic generation of compound terms (phrases) to build the taxonomy	Yes	This feature eliminates developing on a keyword based platform with rules bases to develop the concepts. This significantly reduces time, costs, and the need for highly trained specialists.
Automated Taxonomy Load	Yes	Can be used to add or delete taxonomies from and existing index. This feature is also used to import organizational taxonomies and supports OWL and MeSH.
Hard coded application	No	conceptClassifier is highly adaptable and extensible delivered as APIs that work in any industry and/or market.
Synonyms Supported	Yes	
Synonym Definitions Required	No	Synonyms are supported but not needed as the technology will identify 'like' topics from the organization's own content as opposed to being created manually, which is very subjective.
Controlled vocabularies	Yes	Typical solutions do not support controlled vocabularies using the client's own content and have to be purchased and applied in isolation of the organizations own content, which is a unique organizational nomenclature. Without this capability, the controlled vocabularies must be created manually.
Scalability	Yes	The product can scale in an enterprise environment to handle petabytes of data. When using a statistical modeling approach the fewer the documents the less accurate the modeling.
Native Integration with SharePoint term store	Yes	
Integrated with SharePoint - 2007, 2010, 2013 - Microsoft Office, FAST, SharePoint Search, Windows Server 2008 R2 FCI	Yes	
Integrated with the Refinement Panel	Yes	
Managed by Subject Matter Experts	Yes	Does not require specialists or extensive training. This is very helpful, within an organization as each business unit may have different nomenclature to describe the same topic and the knowledge worker is most likely to identify correct terms for the taxonomy.

© 2012 Concept Searching

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, 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, 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.

© 2012 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