Smart Content Framework™
White Paper

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

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

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

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

June 18, 2012 – revised June 6, 2015

© 2013 Concept Searching
Abstract

The Smart Content Framework™ was developed by Concept Searching to address the need for information governance as it applies to unstructured content. One of the biggest problems tackling the implementation of an information governance strategy is that most companies still do not manage their unstructured content, nor do they use it to improve a variety of business processes such as search, records management, compliance, data privacy, migration, text analytics, or Enterprise 2.0.

This White Paper defines the building blocks that are needed to achieve the ability to not only manage unstructured content, but also implement a strategic and tactical information governance plan. Having several clients who have implemented the Smart Content Framework™ it has proven it to be highly effective and provides significant value in terms of risk reduction and improved business performance.

Author Information

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

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

The Smart Content Framework™ was developed by Concept Searching to address the need for information governance as it applies to unstructured content. One of the biggest problems tackling the implementation of an information governance strategy is that most companies still do not manage their unstructured content, nor do they use it to improve a variety of business processes such as search, records management, compliance, data privacy, or Enterprise 2.0.

The Smart Content Framework™ provides the building blocks to not only manage unstructured content but also leverage content assets to reduce organizational risk, solve business challenges, and improve business processes. The uniqueness of the Smart Content Framework™ is the ability to combine the building blocks with one set of technologies, leveraging an organization’s current IT infrastructure and internal expertise. The flexibility of the technologies enable the organization to address key failures within the management of unstructured content and solve pressing challenges for example, data privacy issues.

This White Paper defines the building blocks that are needed to achieve the ability to not only manage unstructured content but to also implement a strategic and tactical information governance plan. Having several clients who have implemented the Smart Content Framework™ has proven it to be highly effective and provides significant value in terms of risk reduction and improved business performance.

Building Block 1: Enterprise Metadata Repository

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 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.

At a fundamental level, enterprises still struggle with managing content assets. This has always stemmed from the end user’s inability to accurately and consistently tag content
for search, reuse, storage, records identification, and archival purposes. Most organizations focus on relying on the end user for appropriate metadata tagging. Typical software applications still lack the ability to fully extract the concepts within content, nor do they have the flexibility to be applied to multiple business disciplines.

For example, one solution may improve search outcomes, but cannot be used for records declaration and preservation. Only by eliminating the human factor from the process, and providing a technology that can automatically generate semantic metadata through concept extraction, can viable enterprise metadata management be achieved.

The statistics indicate that:

- 80% of Enterprise Data is Unstructured (IDC)
- 60% of Documents are Obsolete (eLaw)
- 50% of Documents are Duplicates (Equivio)

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

Many of the issues organizations are seeking to solve include:

- Ability to develop a single repository of organizationally relevant metadata to be made available to any application that requires its use
- Integration with an organization’s enterprise search engine to improve search results
- Elimination of costs and errors associated with end user tagging
- Normalization of content across functional and geographic boundaries to remove ambiguity in vocabulary
- Metadata managed and changed in one place
- Ability to apply policy consistently across diverse repositories and applications
- Provides flexibility to rapidly make changes to the repository for regulatory compliance where changes are immediately available for use by applications

An enterprise metadata repository is the primary building block in the Smart Content Framework™ that enables the proactive management of content. This first building block is an enterprise infrastructure component tightly integrated with the management of the lifecycle of content.

This fundamental component of the framework provides the technologies to build a consistent information infrastructure that can be shared across different applications and business divisions. This is the key factor that extends the usability of the metadata framework for improving search outcomes, records management, compliance, governance, migration, and protection of security assets.

**Building Block 2: Insight**

The second building block in the Smart Content Framework™ is Insight. The two key performance measures for information retrieval are precision and recall. The ideal solution is to have them balanced. Precision is the retrieval of only those items that are relevant to the query. Recall is the retrieval of all items that are relevant to the query. Higher precision often leads to missing items that may be relevant to the query.
use a different vocabulary. Higher recall often leads to the retrieval of too many items that may be unrelated to the query. Regardless of the enterprise search solution, the delivery of meaningful results depends on the ability to effectively index and classify content, and to develop taxonomies to better manage the content. The search engine itself provides the features, functions, and interface, while the classification structure delivers relevant results. Transforming content into knowledge assets can lead to better decision making and business agility, but only works when the content can be found.

The advantages of searching, and finding, highly granular and relevant information include:

- Provides the ability to find and deliver the most relevant and granular results from large, heterogeneous repositories
- Provide access to relevant knowledge assets that typically would not be found
- Reduces duplication of content
- Make content available for re-use and re-purposing instead of recreating it
- Removes ambiguity in search
- Compliance and security of content assets
- Improves any interactive metadata application such as search, eDiscovery, litigation support, FOIA, text analytics, social tagging, and collaboration

**Application Use Case: Enterprise Search**

A sound information governance strategy includes enterprise search. These information access technologies are a powerful business tool, yet surprisingly are often not scrutinized or evaluated during acquisition. For many organizations, content exists in numerous locations, on diverse repositories and replicated across various silos. Most end users are unable to find relevant information to support business objectives resulting in the inability to find, reuse, and repurpose information. All of this leads to impaired decision making and decreased organizational agility.

Manual tagging is subjective at best and often lacks any alignment to the enterprise goal or mission. Studies have concluded that the same individual will tag the same content differently in the morning and the afternoon. This subjectivity is immediately applied to search results, resulting in inaccuracy, subjectivity, and relevant information never being found.

The results of impaired search results and organizational costs include:

- Estimates indicate that end users spend 2.5 hours per day to find information necessary to do their jobs *(IDC)*
- 85% of relevant documents are never retrieved in search *(IDC)*
- 23% of executives feel that the top drawback of enterprise search is that too much irrelevant data will be found *(Information Week Analytics, 2011 Search Survey)*

Traditional search products force the end user to ask the right question, using the right combination of keywords, and repeating that iteration until the content is found. This not only hampers the user’s ability to find relevant information and reduces their productivity, but also impacts the organization financially. Information cannot be used because it cannot be found, information that cannot be found has no value, and information that is hard to find is rarely used.
Many of the issues organizations can solve include:

- Provide insight into content regardless of where it resides
- Reduce duplication of content
- Make content available for reuse and re-purposing instead of recreating it
- Ensure compliance and security of content assets

**Application Use Case: Text Analytics**

Many organizations still struggle with the most basic aspects of managing unstructured content, which includes 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 has led to poor information governance practices.

One of the challenges is that data is machine driven, whereas unstructured content is driven by people, which makes the nuances, insights, relationships of disparate content, sentiment, and knowledge capital much more difficult to extract. Unstructured content is continually in a state of flux and changes rapidly. Intelligent search assists an organization in finding relevant information based on the end user requirements.

Text analytics leverages the Smart Content Framework™ to capture the meaning of content at a more granular level. Text analytics can be used to extract and group patterns, relationships, and sentiment about a specific topic.

The issues facing organizations include:

- The majority of organizations do not manage their unstructured content and few have adopted metadata, auto-classification, and taxonomy tools
- Approximately 69% of the data most organizations keep can, and should, be deleted (Lorrie Luelli, Of counsel at Ryley Carlock & Applewhite, PC Information Governance)
- 80% of corporate information is unstructured (IDC)
- 86% of respondents admit that unstructured data is important to their organization, yet only 11% have clear procedures and policies for managing unstructured data in place (Unisphere Survey)

The ability to enrich content through concept identification and taxonomies mirroring the topics enables organizations to capitalize in real time on unstructured content, to simplify their business processes, drive positive business outcomes, and transform unstructured content into business assets.

The advantages are:

- Ability to capture content regarding a specific topic
- Ability to analyze in real time sentiment from diverse sources (blogs, wikis, Twitter, email)
- Provides consistency and standardization through elimination of manual tagging
- Eliminate content ‘noise’ and gain valuable insight into content

The valuable insight gained can be used to identify competitive advantages, customer sentiment, service and support, or any topic where the organization needs to gain a better understanding.

“Losing control of content: One might argue that sharing content is, by definition, giving control of it to others. But lots of companies spend good money trying to create a message and to build a brand. Every word on the company website and in collateral publications is vetted and edited to maintain a consistent message. When you open up the conversation, for better or worse you lose control of that message, at least in ways you have previously defined it.”

Ron Miller
Enterprise 2.0 Definition and Solutions CIO Magazine
Application Use Case: Business Social Tagging and Enterprise 2.0

Enterprise 2.0 is technology to bring people together and let them interact, without specifying how they should do so (Andrew McAfee, 2009). Another way of expressing this is Enterprise 2.0 supports the information organization: the business social networks through which work often really gets done (Rob Cross and Andrew Parker, 2004).

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.

There are several excellent uses of business social networking tools, used internally or with external constituents. The tools can also achieve benefits in applications such as project collaboration, awareness of organizational knowledge, employee induction and training, expertise location, communities of interest, collective intelligence, and innovation management.

One of the basic problems is control, 54% of CIOs forbid the use of social networking tools (Sharon Gaudin, 2009). This results in either losing control of content or a lack of control. It has been shown, time and again, that the majority of security breaches are caused internally either by accident or maliciously. This cannot only wreak havoc but prove costly to the organization. Organizations spend a considerable amount of time and money building a strong and consistent brand. For better or for worse, that brand can be jeopardized by opening it up to uncontrolled communication. The problem of management support over the past few years has turned around and many executives understand the value of business social applications and tools.

Unfortunately, the other problem, that still exists, has been user acceptance. Unless the tools are easily integrated into day-to-day business and provide value to the individual they are typically not used. Therefore, integration is a key factor for success. The E2.0 adoption council reports that 72% of organizations reported resistance from users, and only 32% overcame it.

Enterprise 2.0 can be effectively used to create business networks to share knowledge and expertise. The use of the technologies provides structure for social networking applications. The primary benefit is the ability to foster collaboration and knowledge sharing, either from content, or people expertise.

The use of these technologies to address Enterprise 2.0:

- Improves search outcomes by providing insight into content
- Groups similar users, concepts, or content together
- Automatically tags content based on concepts as well as incorporating folksonomies
- Identifies people with expertise, knowledge or interest in a topic
- Can be embedded in the workflow of everyday activities
- Protects and secures confidential information from unauthorized participants

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.

**Building Block 3: Risk**

Corporate risk is increasingly becoming a top priority for senior management. Information that is compromised has now become a corporate responsibility even to the level of holding executives personally liable. The demarcation of who is responsible for the protection of privacy data is becoming blurred. Each business function may have a unique view of what is confidential, such as legal, human resources, and product development. It can no longer be the responsibility of one person, and must encompass all stakeholders and end users to identify and secure privacy assets to protect the organization.

Risk has to be determined by the organization. Part of the problem has been the inability to identify risk factors, analyze them, put the proper policies in place, and take action on them. Risk is different for every organization, whether it be regulatory, intellectual property protection, cyber security, eDiscovery, data retention, even the use of information in unintended ways. The benefits of this approach include:

- The ability to effectively identify and validate the ‘risk’ factor
  - Cost versus Benefit (an organization may want to assume risk in certain instances)
  - Provides the ability to identify risk (known and unknown) while weighing information value
  - Proactively addresses and reduces risk factors through the use of business processes and technology
  - Integrated into the organization’s enterprise objectives or functional objectives

Through the ability to capture and analyze unstructured and semi-structured content, the organization can identify and quantify risk.

**Application Use Case: Privacy and Protection of Confidential Information**

A data exposure does not have to be catastrophic or even public. The bottom line is that it is the responsibility of the organization to set the policies, and the responsibility of the accountable stakeholders, and ultimately the end users, to protect and hold confidential certain information assets. Regardless if a breach is contained within the organization, or worst case scenario exposed to the world, 76% of organizations will face investigations, forensics, law enforcement, and actions to protect the victim from harm.

The root of the problem goes back ultimately to the end user. Statistics indicate that:

- 70% of data breaches are due to a mistake or malicious intent by end users (Ponemon Institute)
- 88% are attributed to negligence (Wharton Information Security Best Practices Conference)
- Average cost per exposed record is $197 and ranges from $90 to $305 (Ponemon Institute)
- 62% of data breaches were attributed to a significant internal error

Using the core technologies to identify risk provides the proactive identification and protection of unknown privacy exposures before they occur, as well as monitor in real time organizationally defined vocabulary and descriptors in content as it is created or ingested.
The organizational benefits include:

- Reduces organizational costs associated with data exposures, remediation, litigation, and fines and sanctions
- Eliminates the risk associated with end user non-compliance issues
- Eliminates manual metadata tagging and human inconsistencies that prohibit accurate identification and protection of unknown privacy/confidential data assets
- Protects the organization by identifying and securing unknown data privacy/confidential information and preventing the portability and electronic transmission of secured assets
- Protects the organization in real time as potential exposures are identified
- Available in on-premise, hybrid, or cloud environment and can identify potential exposures from diverse repositories

Regardless of the size of the organization or the industry, data privacy should be a high priority to ensure that content is proactively identified and protected. Whether it is an internal or external breach of confidential information, the stakes are too high not to address this issue.

With stricter regulatory oversight and the potential of significant remedial costs and fines, organizations can also expect to face civil and criminal penalties escalating the already high costs of breaches.

Both the public and private sector continue to build complex data flows reaching far beyond the confines of the organization. Information is routinely communicated to clients, vendors, partners, consultants, state and local government agencies, making the prevention of data breaches and identity theft complex and challenging.

**Building Block 4: Policy**

Building Block 4 in the Smart Content Framework™ is Policy. The building block includes the ability to utilize metadata and develop and deploy processes that identify records, privacy information, and intellectual assets. This automation of the process is done transparently, without user involvement, to 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 defining the policies with automatic enforcement. This also addresses the issue of intelligent migration where organizations are seeking to implement an infrastructure where content is relevant, protected, archived or deleted during the migration process.

The application of policy must be deployed from an enterprise perspective and address the entire portfolio of information assets. It consists of a twofold approach; the policies must be created and then followed. Inhibiting success has been the inability to monitor for compliance based on the enterprise policies.
The overarching enterprise governance structure allows staff to work in the most efficient and effective way possible by giving them access to information assets in a controlled and secure manner. A key component is ease of use and transparency. If governance or information management controls are too difficult, they will fail.

Organizations typically take one of two routes; either they keep all information, or force the deletion of information according to policy or length of time. The first approach results in an overwhelming amount of content, still growing at an exponential rate, that is unmanaged, and the knowledge of what the content contains cannot be identified. In the second approach, the organization is deleting corporate memory, not only intellectual assets but also records needed for potential non-compliance issues. In both scenarios, the organization is at increased risk and incurs costs associated with the inability to apply governance to all forms of content within the organization.

The costs to the organization include:

- Less than 50% of content is correctly indexed, meta tagged or efficiently searchable (IDC)
- Highly trained information specialists will agree on meta tags only 33%-50% of the time (Cyril Cleverdon)
- Average cost of manually tagging one item runs from $4-$7 per document and does not factor in the accuracy of the meta tags nor the repercussions from mis-tagged content (Hoovers)

This building block makes available the technologies to assist business professionals who are responsible for ensuring information quality, maintaining the lifecycle of information, addressing the retention and disposition of records, securing and protecting privacy assets, and establishing standards when dealing with information assets including unstructured and semi-structured content.

The benefits to the organization include:

- Early detection of non-compliance issues for remediation
- Reduce costs associated with the inability to find information
- Ability to make better decisions
- Eliminate costs associated with tagging and mis-tagging
- Ability to develop and enforce policies removing the end user from the decision making process

Sound information governance and the deployment of transparent policies can assist organizations in managing and balancing regulatory requirements, the risks and costs, while ensuring that knowledge workers can be productive. It enables not only the management of content to reduce risk, but also the ability for the organization to make better business decisions and reduce risk.
Application Use Case: Records Management

In many organizations, enforcement of policies falls to records management, which is typically not viewed as a strategic business group. This results in a disconnect between senior level management and the actual enforcement of policy. The ability to develop policies is not that challenging; where organizations run into the biggest hurdles is in the enforcement of the policies. Unfortunately, successful records management and policy enforcement must include not only records management professionals but also human resources, compliance, legal, and ultimately the end user, to implement a comprehensive information governance plan which ensures adherence to policies.

The costs being incurred by the organization include:

- 67% of data loss in records management is due to end user error (PRISM International)
- Large organizations lose a document every 12 seconds (PRISM International)
- It costs an organization $180 per document to recreate it when it is not tagged correctly and cannot be found (IDC)

Traditionally the lifecycle of a record began at creation to eventual disposal. Although it still does, it has become more complex with issues such as privacy; increased regulations, compliance issues, and the content explosion have impacted the ability of organizations, specifically records management professionals, to keep pace. Today, a more comprehensive approach for managing information is necessary to ensure compliance with policies.

Since the technologies can identify exposures, non-compliance, and risk the results can achieve many benefits in records identification and declaration to the organization including:

- Reduce potential non-compliance issues and mitigate corporate risk
- Reduce costs associated with litigation and eDiscovery
- Eliminate costs associated with manual tagging

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

Application Use Case: Intelligent Migration

Migrating unstructured content can be a laborious and costly activity. Migration of unstructured content is a less used component of information governance, however, a critical one. It is relatively straightforward to mass move content from one repository to another. Some tools provide the ability for human intervention to improve the metadata individually when moving. 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 planning. The issues facing organizations when migrating unstructured content include:

- Poor classifications must be corrected before migration
- Volumes are too high for manual classifications
- Consistency of human review is unreliable and too costly
• Protecting and preserving confidentiality of content when migrated can be difficult to achieve

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. The benefits of this approach:

• Identifies records that have never been declared
• Identifies any confidential/secure content that was not previously identified
• Extracts intelligent metadata before migration to migrate content through auto-classification to the appropriate repository, with semantic metadata applied
• Retains security on content and identifies content that wasn’t secured that should have been secured

Migration can be difficult to accomplish successfully. The advantage of this building block is to provide organizations with an intelligent migration approach delivering the benefits of organized and managed content aligned with the organizational needs long after the migration.

Building Block 5: Action

Action as a building block in the Smart Content Framework™ is the execution and interactive management of the policies and subsequent processes that ensures all unstructured and semi-structured content is processed in a manner that achieves the information governance objectives.

The advantages include:

• Fulfills the defined organizational policies that reduce risk and enable effective management of all semi-structured and unstructured content
• Enforceable and adopted by business users
• Facilitates and improves business processes
• Quantifiable and able to be measured

Information governance must be an organizational effort with support from a wide variety of functional groups, including end users. The Concept Searching technologies address many, if not most, of information governance technology challenges. However, our technologies are only the enabler to achieve the organization’s objectives.

Summary

The need for an information governance strategy for unstructured content is a priority. Content is accumulated from a variety of sources and must be in managed and aligned to the information governance organizational strategy and tactics. The result is the ability to manage the quality of information as well as its lifecycle while at the same time reduce organizational risk and improve business performance.

Many organizations are using the Smart Content Framework™ as a key component in their information governance initiatives. Since the technologies are flexible, the building blocks solve search, text analytics, eDiscovery and FOIA, content management, records management, compliance, migration, data privacy, and Enterprise 2.0 challenges with a single solution. This leverages enterprise information assets, increases organizational agility and protects clients’ investment in technology.
About Concept Searching

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

Concept Searching is the only platform independent statistical metadata generation and classification software company in the world that uses concept extraction and compound term processing to significantly improve access to unstructured information. The Concept Searching Microsoft suite of technologies runs natively in SharePoint 2007, SharePoint 2010, SharePoint 2013, FAST, Windows Server 2008 R2 FCI, Office 365, and in Microsoft Office applications.

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

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