

## Managing Unstructured Content in the Cloud 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

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

Twitter: @conceptsearch

October 23<sup>rd</sup>, 2012

© 2012 Concept Searching

---

## Abstract

This white paper focuses on the challenges and benefits of managing unstructured content in a cloud based environment, as well as in a combined deployment of on-premise and cloud. Since Concept Searching technologies are platform independent, the integration of on-premise and cloud solutions removes many of the obstacles facing organizations as they evaluate cloud options. Utilizing Concept Searching's **Smart Content Framework™** and technologies improves the ability to manage unstructured content, in either or both environments, and addresses the major concerns of security, compliance, privacy protection, and synchronization of content.

## 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.

---

## Table of Contents

<b>Abstract</b> .....	<b>1</b>
<b>Author Information</b> .....	<b>1</b>
Table of Contents.....	2
<b>Overview</b> .....	<b>4</b>
<b>Cloud Computing - The Basics</b> .....	<b>4</b>
Service Delivery Models.....	4
Software as a Service (SaaS) .....	4
Cloud Development as a Service (DaaS) .....	5
Cloud Platform as a Service (PaaS) .....	5
Infrastructure as a Service (IaaS) .....	5
Emerging Deployment Models .....	5
Private Cloud.....	5
Public Cloud .....	5
Hybrid Cloud .....	6
Community Cloud .....	6
<b>Key Challenges in Cloud Adoption</b> .....	<b>6</b>
Security and Data Privacy .....	6
Regulatory and Compliance Issues .....	7
Synchronization of On-Premise and Cloud Applications .....	7
Global Considerations.....	8
<b>Key Benefits of Cloud Adoption</b> .....	<b>8</b>
Reduce Costs.....	8
On Demand Access to Information.....	8
Improved Collaboration and Communication .....	9
Ability to Focus on Core Capabilities .....	9
<b>The Advantages of the Smart Content Framework™</b> .....	<b>9</b>
<b>Technologies</b> .....	<b>10</b>
Automatic Semantic Metadata Generation.....	10
Auto-classification.....	11
Taxonomy .....	12
<b>Turning Challenges into Solutions</b> .....	<b>12</b>
Non-Intrusive Information Governance .....	12
Security .....	13

---

Data Privacy .....	13
Compliance .....	14
Collaboration and Knowledge Sharing.....	14
Big Data in the Cloud .....	15
Synchronization of On-premise and Cloud Infrastructures .....	15
Migration .....	16
Replication .....	16
Synchronization .....	16
Microsoft Office 365 Integration .....	17
<b>Summary .....</b>	<b>17</b>
<b>Appendix A: Concept Searching Products &amp; Technologies.....</b>	<b>18</b>
conceptSearch .....	18
conceptClassifier.....	18
conceptTaxonomyManager .....	19
conceptClassifier for SharePoint.....	19
Optional Components.....	20
conceptContentTypeUpdater .....	20
conceptTaxonomyWorkflow .....	20
<b>About Concept Searching .....</b>	<b>21</b>

---

## Overview

Cloud computing continues to evolve into a more attractive option and in some cases a necessity for organizations of all sizes. In any cloud scenario, the security, reliability, and accuracy of content assets must be maintained and managed as if it was on-premise. According to 'Gartner Hype Cycle for Cloud Computing 2012', "the best results are being attained by enterprises that focus on a very specific strategy and look to cloud-based technologies to accelerate their performance. Leading with a strategic framework of goals and objectives increases the probability of cloud-based platform success. Those enterprises that look to cloud platforms only for cost reduction miss out on their full potential."

This white paper focuses on the challenges and benefits of managing unstructured content in a cloud based environment, as well as in a combined deployment of on-premise and cloud. Since Concept Searching technologies are platform independent, the integration of on-premise and cloud solutions removes many of the obstacles facing organizations. Utilizing Concept Searching's **Smart Content Framework™** and technologies improves the ability to manage unstructured content, in either or both environments, and addresses the major concerns of security, compliance, privacy protection, managing big data, and synchronization of content.

"The "cloud" model initially has focused on making the hardware layer consumable as on-demand computer and storage capacity. This is an important first step, but for companies to harness the power of cloud, complete application infrastructure needs to be easily configured, deployed, dynamically-scaled and managed in these virtualized hardware environments."

K. Sheynkman, Co-Founder  
Elastra Corporation

## Cloud Computing - The Basics

According to the official National Institute of Standards and Technology's (NIST) definition, "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."<sup>1</sup> In addition, they have defined the Essential Characteristics, Service Delivery Models, and Deployment Models as illustrated below.



## Service Delivery Models

There are four Service Delivery Models that have currently been defined for cloud computing. Depending on the organizational needs and requirements, these will dictate the type of Service Delivery, or a mix of models. These are described below.

### Software as a Service (SaaS)

Software as a Service (SaaS) includes applications that are licensed to customers, such as Microsoft Office 365, Google Apps, and Salesforce.com. End users access the application through a thin client interface such as a web browser. According to IDC, it is predicted that SaaS will be adopted by most companies in the next few years at some level or the other, especially in content management, collaboration, document

---

<sup>1</sup> *The NIST Definition of Cloud Computing (NIST Special Publication 800-145) by Peter Mell and Tim Grance*

---

management, and customer management applications.<sup>2</sup> Specifically for the Small and Medium Business (SMB), it can reap the benefits of world-class infrastructure and enterprise class features without the capital investment. On the other hand, the cost of conformity is the cost of flexibility. When the business changes the flexibility to add features or functions that are considered critical by the organization isn't always available.

#### **Cloud Development as a Service (DaaS)**

Development as a Service (DaaS) is defined as a set of tools and APIs provided for creating customized applications such as code editors, source control systems, and batch scripts. A good example is Salesforce.com which encourages development of new applications and features from third party developers. Companies can then purchase the feature or application resulting in a customized application tailored to their requirements. Another example would be Microsoft Windows Azure.

#### **Cloud Platform as a Service (PaaS)**

Platform as a Service (PaaS) provides the hosting for client developed applications or third party applications that are developed using Java and .Net. Examples include Microsoft Windows Azure, Salesforce.com, Amazon Web Services, Sun, and Rackspace. The advantages of this delivery model is considerable cost savings and faster deployment and potentially better security as it is often built into the application. The biggest challenge in this scenario is the difficulty in migrating existing applications from internal data centers to the cloud.

#### **Infrastructure as a Service (IaaS)**

Infrastructure as a Service (IaaS) provides the computing resources such as processing, storage, a network to run full virtual servers. In this scenario, the customer has control over operating system, storage, and deployed applications. This provides significant benefits for large enterprises to reduce the operational cost and support for large data centers. For small companies or startups this provides the ability to concentrate on core competencies without worrying about the IT infrastructure. Examples include Amazon Web Services, Sun, and Rackspace.

### **Emerging Deployment Models**

The four types of Deployment Models that are used most frequently include the Private Cloud, Public Cloud, Hybrid Cloud, and the Community Cloud. They are described below.

#### **Private Cloud**

The Private Cloud is often used and suited to organizations that require a high level of data protection. According to Werner Vogels, VP and CTO, Amazon, "CIOs know that what is sometimes dubbed 'private cloud' does not meet their goal as it does not give them the benefits of cloud: true elasticity and capex elimination." The Private Cloud is minimal risk due to single ownership and strong shared mission goals and legal/regulatory requirements.

#### **Public Cloud**

The Public Cloud model has emerged as a cost effective option for Small and Medium Businesses (SMBs). This model is also used by companies who want to 'try before they buy' and use a cloud option before committing to a specific cloud model. The Public Cloud is the highest risk due to lack of security control, multi-tenancy, and data

"Public cloud services are generally not providing as much customization as customers want, but the cloud model is gaining popularity both among users who want to sidestep their companies' IT departments, and from small businesses that want to get out of the IT business."

Tim O'Brien  
Director, Platform Strategy  
Microsoft

---

<sup>2</sup> IDC Worldwide Software as a Service 2010-2014 Forecast: "Software Will Never Be the Same"

---

management. There is little Service Level Agreement (SLA) control and a lack of common regulatory guidelines which currently exists throughout the industry.

### Hybrid Cloud

The Hybrid Cloud is a mix of the Private Cloud and the Public Cloud, which provides more flexibility but is not as expensive as the Private Cloud model. In this model, the risk is dependent on the combined solution model. The combination of a Private and Community Cloud is the lowest risk where the combination of Public is the greatest risk.

### Community Cloud

The Community Cloud model is expected to address the requirements of governments and their agencies. These are typically dedicated by a user industry group that shares the same concerns such as security, policy, and compliance requirements. The Community Cloud is typically viewed as moderate risk due to the multi-tenancy, but poses less risk than the Public Cloud due to shared legal, regulatory, and compliance standards. In the US Federal Government, examples include the Defense Information Systems Agency (DISA) Rapid Access Computing Environment (RACE) and the National Aeronautics and Space Administration's Nebula.

Community Clouds represent a lower information security risk profile than a public cloud environment and fewer legal and regulatory compliance issues, but they carry certain risks associated with multi-tenancy.

### Key Challenges in Cloud Adoption

Cloud computing poses several challenges for organizations. The decision is not just moving to the cloud but an accurate analysis of the pros and cons. There is still a lack of awareness of the issues outside of a small group of experts. Organizations do not always know the questions to ask. Cloud computing is still in its infancy in terms of following protocols and standards. This will not be solved overnight. Finally, there is a gap between what the customer is expecting and what actually exists in the cloud computing solutions market.

### Security and Data Privacy

Cloud computing is being aggressively marketed by major vendors such as Microsoft, Google, and Amazon as a solution with significant benefits. Although there are many benefits there are also inherent risks such as a lack of regulations, standards, and data security. Major breaches at Google, Salesforce.com, Twitter, and Amazon have all proven that there are hidden costs and repercussions from compromised data.

Initially, the applications and data that were processed in the cloud were non-sensitive, but as the market is maturing all organizations will face the possibility of compromised data assets. The different types of security considerations are listed below.



“Cloud Computing: Fact versus Fog” Grail Research

One of the most challenging aspects is where the data is going to end up. It's not unusual for a cloud vendor to store data on servers managed by another company. In reality, there can be two or more degrees of separation between your company and your company data. Cloud service providers tend to be vague about their architectures

“Cloud Services have shifted from a year ago. We did a focus group around 12 months ago and they pretty much took the mickey out of the cloud. It was seen as unrealistic and CIO's weren't considering it. What's even more of a surprise is that in a short period of 12 months, we've seen cloud go from a bit of a joke to a number two priority on the plate of CIOs today, and a very serious consideration that they are taking on board.”

Paul Harapin  
Director, ComputersOff.org and  
Ex-MD, VMware

“Security has been identified as the most significant issue associated with cloud computing adoption.”

Melvin Greer  
Chief Strategist, Cloud Computing  
Lockheed Martin

---

and where the data is stored. Concurrently, state and federal regulations govern the management of health-related and other personal data, and from the legal aspect will not accept an 'I don't know' as an answer to questions where data is being stored.

According to IDC, to aid in the understanding of the degree of security in the digital universe, they have classified information that requires security into five categories, each requiring successively high levels:

- Privacy only – such as an email address on YouTube
- Compliance driven – such as emails that might be discoverable in litigation or subject to retention rules
- Custodial – account information, a breach of which could lead to or aid in identify theft
- Confidential – information the originator wants to protect, such as trade secrets, customer lists, confidential memos etc.
- Lockdown – information requiring the highest security, such as financial transactions, personnel files, medical records, military intelligence etc.

Without adequate information on the security and compliance profile of the data, including its ownership, access controls, audits and classification, cloud initiatives can fall short of expectations and put sensitive data at risk. Understanding the data owners, the authorized users, and user activity is critical to garnering organizational input, which in turn, is critical to defining the security and compliance profile of the data for internal datacenter and for the cloud.

“I am 100% responsible and accountable for all technology and every shred of data that moves in and out of my company, and don't want IT to be seen as 'the say-no people' but end users may not foresee the difficulties meshing new products with existing technology. On-premise, we have technology standards. Nothing like that exists in the cloud. If business users adopt these things we CIOs are challenged in IT to figure out how to integrate them with the rest of our world.”

Don Goin, CIO  
Santander Consumer

### **Regulatory and Compliance Issues**

Governance issues have not been fully addressed yet by cloud vendors. Addressing these issues falls not only on the customer of cloud services, but also the cloud service or application provider. For example, the Federal Financial Institutions Examination Council (FFIEC) published a first-ever statement on the risks of outsourced cloud computing for banks and other financial institutions. The FFIEC is charged with standardizing the federal scrutiny of financial institutions. Issues to be considered include response time for accessing the stored data, the vendor's responsibility to implement legal holds, and cooperate in responding to hold orders and regulatory request for information.

### **Synchronization of On-Premise and Cloud Applications**

There are applications available by a variety of cloud vendors for the data loading, synchronization, cleansing, and replication delivered as a cloud service. The goal of course is to develop an integration strategy and plan that can support transfer of data across on-premise and external services as well as ensure that data is always current, secure, and reliable.

The cloud provides several benefits to store and provide access to information that is static for example, providing price lists, product information, brochures etc. Vendors, partners, and clients can access the information, but storing the information in the cloud poses no risk to the organization. On the other hand, moving information from an active on-premise repository to a cloud based archive repository such as in records management requires regulatory and legal compliance requirements. In this scenario, information can be compromised.

The final consideration is when the content will be synchronized. If a cloud repository is used for various stakeholders in different geographic regions for example in project management, it becomes an imperative that the information is up-to-date, and changes



---

are reflected in real time, to avoid the potential of errors and result in poor decision making by knowledge workers, vendors, and potentially clients.

### **Global Considerations**

For organizations that cross global boundaries, additional considerations are also a factor, specifically in terms of security. For example, the Patriot Act allows the US government to subpoena all data stored within the country, the EU Data Protection Directive does not allow personal information to be transferred to any outside country, the Massachusetts Breach Law specifies that citizens' private information must be protected and has specified strict guidelines around storage, access, and transmission of personal information. A cloud environment by nature has no boundaries requiring careful thought on who might be accessing the content.

Deploying a global, or cross boundary, cloud application must also account for the fact that not all countries have high-speed, continuous access to the Internet, which in some cases can make real-time access to information a challenge.

### **Key Benefits of Cloud Adoption**

Cloud computing also offers a compelling ROI for many companies. Due to the economic turn down and continued uncertainty of the global market, many organizations can adopt Cloud service models to reduce costs and capital expenditures. In the cloud environment, organizations can also rapidly scale up or down depending on their own financial model and requirements. Finally, with a pay-as-you-go model they have the option to rapidly restructure at any time.

Cloud computing delivers quantifiable benefits and economic incentives. Selecting the appropriate delivery model will depend on the organization's specific requirements including, performance guarantees, security, and compliance. The appropriate due diligence and deployment can deliver significant savings, improved IT services, and a higher level of reliability going forward.

### **Reduce Costs**

There are several strong arguments for cost reduction in using cloud services. Operating expenses and capital expenditures can be reduced. The technology is also paid for incrementally, as needed, and can be scale up or down depending upon the organizations' needs. Internal IT resources can be used for more productive projects than server updates and installing and configuring hardware and software.

From an infrastructure point of view, organizations can store more data than on private systems. They can also reduce the costs associated with power, space and data center maintenance by taking non-critical services out of data centers. The cloud also removes the risk of hardware and software obsolescence. Typically when an organization purchases a new server, it is based on a future life of three to four years. This can significantly increase the cost as opposed to the cloud scenario, where you pay only for what you need when you need it.

### **On Demand Access to Information**

Depending on the organization, the ability to access information regardless of the location via a variety of devices - for example, netbooks, iPads, cell phones - increases the productivity of end users, no matter where they are physically located. The result is improved efficiency, productivity, and decision making.

“The financial turbulence of the last 18 months has meant every organization has been scrutinizing every expenditure. An IT solution that can deliver functionality less expensively and with more agility - remembering time is money - is hard to ignore against this backdrop.”

Ben Pring, VP  
Gartner Research

---

## Improved Collaboration and Communication

The Cloud is also an excellent tool for collaboration and sharing information. Information stored in the cloud ideally will be up-to-date and enables stake holders, both internal and external to share information, such as management of projects. When the project is completed, the cloud solution is no longer needed.

The ability to share information across global locations in real time reduces the possibility of error, reduces redundancy in emails, and locating the most current document. This enables better decision making, speed to market, and can increase innovation.

## Ability to Focus on Core Capabilities

Whether a large organization or a small one outsourcing non-critical applications to the Cloud provides the ability to focus on core capabilities. The options for cloud computing is not an all or nothing approach. Companies can select what goes into the cloud and what stays on-premise. This results in improved business operations and resource allocation to meet business objectives.

## The Advantages of the Smart Content Framework™

Organizations accumulate digital data at a rapid pace and it is growing exponentially. Using cloud services provides an attractive option to many organizations. The key issue is that content must be continuously managed and protected to remain secure and retain its value. This is not always readily accomplished in the cloud. Regardless of where the data is stored, on-premise or in the cloud, the demand for comprehensive information governance to manage and secure critical data remains constant.

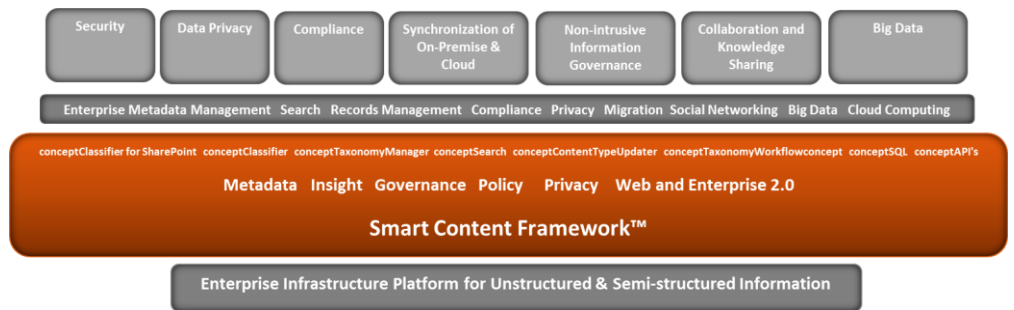
The **Smart Content Framework™** was developed by Concept Searching to address the need for information governance as it applies to unstructured content. One of the biggest problems tackling the implementation of an information governance strategy is that most companies still do not manage their unstructured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

The **Smart Content Framework™** provides the building blocks to not only manage unstructured content but to 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, security and compliance in the cloud environment.

The key building block in the **Smart Content Framework™** is an Enterprise Metadata Repository. Metadata is a necessity to provide the automated collection, storage, analysis, and presentation for on-premise data stores as well as cloud infrastructures. Concept Searching technologies automatically generate semantic metadata and auto-classifies it to organizational taxonomies. The taxonomies are then managed through an interactive, dynamic interface, providing industry unique taxonomy management tools. The metadata infrastructure auto-classifies information, not based on where it is stored, but on the concepts within the content. This can occur on-premise as well as in the cloud.

“In order to get the most benefit from cloud computing initiatives, companies need to develop a clear governance strategy and management plan that sets the direction and objective for cloud computing.”

“It Control Objectives for Cloud Computing: Controls and Assurance in the Cloud”  
ISACA



The building blocks provide flexibility to address the most critical challenges within an organization. For example, some of our clients may want to improve search, others may need to facilitated records management, while others face stringent regulatory and compliance guidelines.

## Technologies

Since 2002, Concept Searching technologies have remained unique in the marketplace. The suite of technologies includes horizontal classification and taxonomy management products that deliver the highest precision without the loss of recall when compared to any comparable technology available today.

Concept Searching’s software products deliver concept based metadata generation, auto-classification, and taxonomy management. Concept Searching is still the only statistical metadata generation and classification software company in the world that uses concept extraction to significantly improve access to unstructured information.

Although platform independent Concept Searching has developed a comprehensive Microsoft suite of products that provide native integration with all versions of SharePoint, automatic content updating, native integration with the Term Store, and integration with Office 365.

The core technologies that provide the framework for all Concept Searching technologies are described below.

*For more information about products, please see Appendix A.*

### Automatic Semantic Metadata Generation

The discovery, collection, and management of metadata are essential for the integration of content across disparate systems, both on-premise and cloud. 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 such as the cloud. Within a cloud as well as on-premise there is a great need to generate far richer metadata and manage it effectively to provide enhanced access to these resources by internal and potentially external stakeholders.

Concept Searching solutions automatically generate semantic metadata based on the concepts within unstructured information. The generation of semantic metadata enables organizations 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.

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

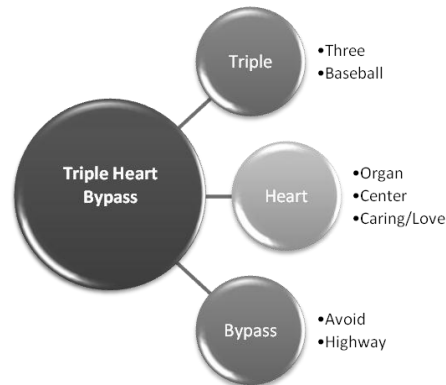
“Unstructured information - such as files, email and video - will account for 90% of all data created over the next decade.”

IDC  
Digital Universe Study 2011

---

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 organization.

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 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

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. The ability to overcome incompatible semantics and meanings provides a method of bridging vocabularies across global boundaries.

### Auto-classification

The manual process for classifying documents is both time consuming and labor intensive. Typically, a person associated with the application 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 organization. This process can be complicated, due to the sometimes complex conditions, which can lead to a classification decision. 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 decisions. 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, the cloud, and websites. All content can be classified on the fly, in real time and classified to one or more taxonomies.

---

## 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 business objectives. 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.

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 enterprises by up to 80%.

“To move critical data and processes into the cloud when there is very little visibility on access and ownership, traceability and data segregation. It is vital that organizations have data governance in order to provide secure collaboration and data protection for their customers, partners and employees. Without it, it will be virtually impossible to manage and protect digital information in the cloud or anywhere else.”

IDC  
Digital Universe Study 2010

## Turning Challenges into Solutions

Although the economic benefits of a cloud infrastructure are well understood many enterprises still do not manage unstructured content under the enterprise umbrella of information governance within their on-premise infrastructure. When cloud is added to the mix and unstructured information remains unaddressed content management, security, and compliance can be added to the new list of issues.

Concept Searching's technologies provide solutions for unstructured content management challenges across heterogeneous environments. The benefit is a comprehensive enterprise approach that protects, secures, and maximizes the value of knowledge assets regardless of where they reside.

Concept Searching technologies add value in the following:

- Non-intrusive Information Governance
- Security
- Data Privacy
- Knowledge Sharing and Collaboration
- Big Data
- Synchronization of On-premise and Cloud Infrastructures

Any cloud solution is based not only on the organizational requirements but the capabilities of the cloud vendor(s). The adaptability of the **Smart Content Framework™** and the technologies provide an extensible feature set that can be used in any environment and any type of cloud configuration.

## Non-Intrusive Information Governance

The automatic generation of semantic metadata removes the end user from the tagging process as well as the subjective ambiguity. 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 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, is reliable, and is correct. This ability to generate conceptual 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.

This capability is core to deploying an information governance strategy both internally and externally to the cloud. From this, the security, compliance, legal requirements, and data privacy can be more easily enforced before it becomes available on the cloud or classified within the cloud to ensure the security of the content. Content can be automatically classified to organizational taxonomies within or outside of the cloud making the content available and easily found by stakeholders depending on their security access, automatically removes and protects organizationally defined confidential content, and removes the subjectivity associated with erroneous meta tags and adds rich metadata to non-existent content.

## Security

There is a growing need for additional security measures to be applied to unstructured content. Although security is the number one concern for many businesses and government, studies show that less than one third of all stored data today has even minimal security or protection and only about one half of the information that should be protected is protected at all. Existing mechanisms to measure security are often subjective and in many cases vague, specifically in the cloud. This makes quantifiable measurement of security features in the cloud difficult.

Information Security should lie solely within the Information Technology department of an organization (McConnell, 2002) and, in the case of cloud computing, the associated vendors. 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. This issue becomes exacerbated in the cloud environment, since it is rarely addressed internally.

## Data Privacy

Concept Searching’s approach is to eliminate end user involvement in the process of metadata generation, unless specifically authorized. The solution augments traditional security products and compliance processes within an organization, by discovering where unknown privacy data PXX3 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 classified, the content can be managed in accordance with internal, regulatory, or government guidelines.

“Cloud computing should be approached carefully with due consideration to the sensitivity of data and guided by organizational cloud information governance strategy. The possibility that ineffective cloud security controls could lead to vulnerabilities affecting the confidentiality, integrity and availability of customers’ information.”

Network World  
“Inside the Cloud Security Risk”

---

*3 Any organizationally defined descriptors and/or semantic content that have been defined by the organization as confidential.*

---

This can be done before processing to the cloud or from the cloud itself, as well as in real-time as content is created or ingested. This is also extremely useful within the cloud to address national, regional, and local requirements dealing with the privacy of confidential information by setting up taxonomies that prohibit/allow access to specific information to unique user groups depending on their geographic location.

### Compliance

Compliance, legal, and government requirements include the integration of policies and processes to address the dependencies across both environments. The analysis of cloud based solutions must identify potential policy conflicts and the overall risk incurred from new technologies and the cloud environment itself. The ability to track dependencies across events, processes, and people becomes much more challenging.

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

Using concept [ContentTypeUpdater](#) in SharePoint or concept [TaxonomyWorkflow](#) in a non-SharePoint environment enable organizations to take advantage of 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 to the appropriate repository (i.e. taxonomy). For organizations with large content repositories this is no trivial matter. Taxonomies that mirror the Records Management file plan can automatically identify content that should be declared a record, archived, or processed differently providing consistency across on-premise and cloud environments. This application can also be used for archived information that may be stored on the cloud to reduce storage costs internally.

“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

### Collaboration and Knowledge Sharing

Enterprise and Web 2.0 can cause chaos as well as deliver organizational benefits. 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.

Any type of sensitive information needs to be protected; security issues must take into account the end user as well as the content asset. Social networking must deliver results and not become a waste of time for end users who will eventually abandon the application specifically when deployed in a cloud environment.

Knowledge is a corporate asset. Managing it within an Enterprise or Web 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 environment neither encourages relevant knowledge sharing nor does it drive a return on investment.

Tools that encourage collaboration, can link employees, partners, suppliers, and customers to share information, and are becoming more useful 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.

For example, organizations are increasingly recognizing the inefficiencies of using email as a collaboration tool, given how poorly it performs in situations requiring collaborative work on single documents in situations such as project management, acquisitions, innovation, and general administration. In a cloud environment, relevant content can be accessed by authorized users regardless of their global location.

The objective is to provide structure when implementing Enterprise and Web 2.0 technologies. The Concept Searching technologies provide improved search outcomes by providing insight into content from a common framework where similar users, concepts, and content are grouped together; identify people with expertise, knowledge or interest in a topic; and protects and secures confidential information from unauthorized participants. The end result is a consistent understanding of the value and context of information. It also provides confident cross-organizational decision support capability, shared knowledge, and enterprise availability of metadata knowledge to increase organizational performance.

### **Big Data in the Cloud**

It is predicted that within the next ten years the cloud environment as a repository for Big Data content and analysis will become prevalent. Big Data requires inexpensive storage, high-velocity capabilities, a proliferation of data capture from a variety of technology sources, and the ability to analyze results and predictive reporting. The cloud infrastructure is an excellent repository for Big Data storage.

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 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.

Placed in a secure cloud environment, the Concept Searching capabilities become an enabler for organizations to begin to make sense of Big Data for unstructured content and how it can be used to drive business objectives, regardless of what they might be.

### **Synchronization of On-premise and Cloud Infrastructures**

Ensuring content on-premise and in the cloud is current and accurate is a necessity to eliminate multiple versions of the same content existing in multiple repositories and that the information being used enables accurate decision making to only authorized users. The cloud solution must address:

- Migration
- Replication
- Synchronization

“The use of big data will become a key basis of competition and growth for individual firms. From the standpoint of competitiveness and the potential capture of value, all companies need to take big data seriously. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value from deep and up to real-time information.”

McKinsey & Company  
Big data: The next frontier for  
innovation, competition, and  
productivity  
McKinsey Global Institute



---

## Migration

Companies preparing to move from legacy IT infrastructures to cloud-based systems will need to take ownership of the transition process instead of pushing this responsibility to their service providers to prevent the migration from going awry.

Migration of applications to the cloud carries with it the same issues as on-premise migration of content. Migrations to the cloud are an extension of the operational perimeter of the business and should be subject to the same access controls, policies, and security regardless of where they reside. Once in the cloud the same rules and policies must exist and be applied to all content.

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 may 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 use concept [TaxonomyWorkflow](#) or concept [ContentTypeUpdater](#) in conjunction with concept [Classifier](#), in either the SharePoint on-premise or cloud environment. The solution includes combined workflow capabilities and enables 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.

## Replication

In regards to replication, rules based on the semantic metadata tags provides the ability to replicate content as determined by the needs of the business, for example, backup, disaster recovery, archiving, publishing, etc. In certain instances, organizations may want data to exist in multiple repositories. Of course, this also entails making sure that the content existing in multiple repositories, is current and the same version, otherwise it still remains unmanaged and essentially unusable.

As content is created or ingested, workflow rules can be applied that when content contains organizationally defined descriptors and/or concepts it is automatically routed to one or more repositories, either on-site or in the cloud or in both environments.

## Synchronization

In terms of this white paper, synchronization refers to the synchronizing of unstructured content from an on-premise to a cloud environment in real-time, or vice versa. There are sophisticated data synchronization applications that are based on the premise of keeping multiple copies of a database in coherence with one another or to maintain data integrity that organizations may require but are beyond the scope of this paper.

Concept Searching aids in the synchronization of unstructured content as it becomes updated or created in diverse environments, in that unstructured content remains current regardless of where it resides. Synchronization as opposed to replication must ensure that the content is up-to-date regardless of where it resides. In a co-dependent environment, if some portion of knowledge work is completed on an internal device that information must also co-exist in the cloud. If not, the end result is the same problem as replication where it remains unmanaged and essentially unusable. It must also work the same way in reverse, where changes to content that are made in the cloud, the internal enterprise metadata repository is also updated accordingly, and in real-time.

---

## Microsoft Office 365 Integration

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 concept **Classifier** and concept **TaxonomyManager** 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 stores, 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.

## Summary

The cloud offers great benefits and just as the Internet invaded our lives, so will cloud computing from a business perspective. In the future, organization will create and transmit more information that was previously created and transmitted. Not only will the cloud meet the demands for expandable infrastructures, but also bring cost benefits, and line-of-business improvements, whatever they may be. The importance of an information governance strategy for the cloud organizations must identify strategies, risks, and processes to govern these new initiatives.

Cloud computing is somewhat still in its infancy. As the industry as a whole grows and matures, organizations will be able to take advantage of the cost, availability, and flexibility these technologies promise to deliver. For right now, organizations must still exercise due-diligence in the protection and use of content assets regardless of where they reside. The organizations that have achieved the ability to harness metadata to support their information governance practices will have a far greater chance of succeeding the cloud.

---

## 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 API's based on XML and Web Services. Transparent access to system internals including the statistical profile of terms is standard.
- Easily customized for your requirements.

### 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.

concept**Classifier** 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

### concept**TaxonomyManager**

concept**TaxonomyManager** 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

### concept**Classifier for SharePoint**

concept**Classifier** 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. concept**Classifier** 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.

---

## Optional Components

### **concept**ContentTypeUpdater

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.

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.

*This add-on component works only in SharePoint 2010 and SharePoint 2013.*

### **concept**TaxonomyWorkflow

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.

*This add-on component is platform independent.*

---

## 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 365.

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>.

**Microsoft** Partner  
Gold Independent Software Vendor (ISV)

© 2012 Concept Searching



**Americas**  
+1 703 531 8567  
[info-usa@conceptsearching.com](mailto:info-usa@conceptsearching.com)

**Europe**  
+44 (0)1438 213545  
[info-uk@conceptsearching.com](mailto:info-uk@conceptsearching.com)


**Canada**  
+1 703 531 8567  
[info-canada@conceptsearching.com](mailto:info-canada@conceptsearching.com)

**Australia**  
+61 (0)2 8006 2611  
[info-australia@conceptsearching.com](mailto:info-australia@conceptsearching.com)

**New Zealand**  
+64 (0)4 889 2867  
[info-nz@conceptsearching.com](mailto:info-nz@conceptsearching.com)

**Africa**  
+27 (0)21 712 5179  
[info-sa@conceptsearching.com](mailto:info-sa@conceptsearching.com)

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

Follow us on Twitter  
 [@conceptsearch](https://twitter.com/conceptsearch)

[www.conceptsearching.com](http://www.conceptsearching.com)