conceptSearching

# Unstructured Content and Big Data

## The Dilemma of Unstructured Content in the Big Data Puzzle

For most organizations, data is abundant and overflowing but not exploited to derive the most value to improve decision making, drive profitability and gain competitive advantage. There are several definitions of Big Data, such as Volume, Velocity and Variety; Variable Attributed Subjects, People or Time (VAST); and Algorithms, Machines and People (AMP).

Regardless of the definition, Big Data deals with structured data, semi-structured or unstructured data, and unstructured content. The first two items are the primary focus of the term Big Data and unstructured content, although included, is more often than not vaguely alluded to.

One of the fundamental problems is the view that unstructured content must be managed in databases for analysis, in the same way as structured and semi-structured data, which is not the right approach. Data is machine driven, whereas unstructured content is driven by people, which makes the nuances, insights, relationships of disparate content, sentiment, and knowledge capital much more difficult to extract. Unstructured content is continually in a state of flux and changes rapidly.

Organizations that can capitalize in real time on unstructured content can simplify their business processes, drive positive business outcomes, and transform unstructured content into business assets.

## Struggling with the Basics

Many organizations still struggle with the most basic aspects of managing unstructured content, which include free-form language, emails, documents, and social networking applications. The perceived lack of need, or seemingly overwhelming challenges, for managing unstructured content has caused not only the inability to manage content but has also led to poor information governance practices. This has far more immediate and serious implications in terms of compliance and data privacy issues, which can lead to fines, sanctions, and loss of business.

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

This enables not only the effective management of content but also the use of semantic metadata and enterprise taxonomies to improve search, records management, compliance, data privacy, Enterprise 2.0, and migration.

## The First Step

Many organizations, if not most, do not have a plan for, nor do they proactively manage, their unstructured content. Not only that, they are not using their unstructured content at a most basic level to improve business processes. For example:

- 80% of Enterprise Data is unstructured
- 60% of documents are obsolete
- 50% of documents are duplicates
- A typical knowledge worker will spend 2.5 hours per day searching for information
- 85% of relevant documents are never retrieved in search
- The average cost of manually tagging one item runs from $4 to $7 per document, and does not factor in the accuracy of the meta tags nor the repercussions from mistagged content
- 67% of data loss in records management is due to end user error
- 70% of data breaches are due to a mistake or malicious intent by end users

To manage unstructured content an *Enterprise Metadata Repository*, the first building block in the **Smart Content Framework™**, needs to be in place. This component is required to extend the use of capturing an organization's unique nomenclature and vocabulary, which will be used to broaden the use of content assets to drive business processes.

There are two issues that are consistently stumbling blocks that organizations typically do not know how to solve. The first is the end user's inability to correctly tag content for re-use and the organization's inability to enforce policy. The second issue is the resources, time, and money to build and manage taxonomies.

Concept Searching has eliminated both of these obstacles through automatic semantic metadata generation and easy to use, yet powerful, taxonomy management tools. Both of these technologies are still unique in the industry.

## The Technology

Concept Searching's technologies provide intelligent automated classification and taxonomy management to develop a consistent structure to more effectively manage content assets. Unlike traditional tools, where taxonomy planning and deployment take considerable time and resources, or fully automated approaches, where rich human knowledge capital and input is eliminated, the technologies combine the strengths of both approaches.

The products expedite the process, by providing rich multi-term metadata to rapidly build taxonomies through innovative, real time taxonomy management features.

## Compound Term Processing

Unlike traditional metadata generators, Concept Searching uses *compound term processing* technology, which is unique in the industry. Instead of identifying single keywords, compound term processing identifies multi-word terms that form a complex entity and identifies them as a concept.

## Automatic Semantic Metadata Generation

Since the technology is not restricted to keyword identification, compound term metadata - *concepts in context* - can be automatically generated either when the content is created or ingested. The generation of metadata based on concepts extracts compound terms and keywords from a document, or a corpus of documents, highly correlated to a particular concept.

By identifying the most significant patterns in any text, these compound terms can then be used to generate non-subjective metadata based on an understanding of conceptual meaning. The compound terms, keywords, and acronyms identified are then used by the taxonomy developer to rapidly deploy enterprise taxonomies.

## Powerful Taxonomy Management

Industry statistics indicate the average cost of building a taxonomy for an organization is approximately $75k to $100k, and takes an average of 2 to 6 months to build. The taxonomy component is easy to use and designed for Subject Matter Experts. There is no need for extensive training or highly experienced taxonomists. The tool is extremely powerful and includes features that are not available in any commercial products.

These features have been proven to reduce taxonomy development time by up to 80%. This represents a substantial reduction in the cost and manpower needed to develop and manage enterprise taxonomies, for which the total cost of ownership is typically very high.

## Tying It All Together

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

From the Big Data view, turning everything into structured data is an option, but the current maturity of text analysis tools rate the certainty of the information at less than 70%. This is just data extraction, not concepts or ideas contained in the unstructured content. Unreliable information ultimately produces random garbage. A new and more accurate approach is needed.

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 identify competitive advantages, customer perception, regional trends, and, perhaps more importantly, identify the internal knowledge capital that exists but is rarely used because it cannot be found.

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

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

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

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

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

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

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

Follow us on Twitter
@conceptsearch

**www.conceptsearching.com**