conceptSearching

The Need for a Taxonomy 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

July, 2015

Abstract

At a fundamental level enterprises struggle with managing content assets which stems from the end user's inability to accurately and consistently tag content for search, re-use, storage, records identification, and archival purposes. Statistics show that over 90% of organizations rely on end user tagging, sustaining the use of inaccurate and erroneous metadata being applied to content. Typical taxonomy solutions often lack the ability to extract the concepts within content, so do not generate semantic metadata or provide auto-classification capabilities. The result is organizations are stymied, as the use of erroneous metadata impacts any application that requires metadata, and prohibits the development of workflow processes to improve the processing of content. A pragmatic, easy to deploy taxonomy approach should be evaluated that provides a rapid return on investment. This effectively overcomes the typical academic approach where complex ontologies are used and become difficult to deploy, manage, and maintain.

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 distilled from a variety of distinct content making available useable knowledge.

conceptTaxonomyManager has the capability to automatically group unstructured content together based on an understanding of the concepts and ideas that share mutual attributes while separating dissimilar concepts. This approach is instrumental in delivering relevant information via the taxonomy structure as well as using the semantic metadata in enterprise search to reduce time spent finding information, increase relevancy and accuracy of the search results, and enable the re-use and re-purposing of content. Using one or more taxonomies, unstructured content can be leveraged to improve any application that uses metadata. This flexibility extends to records management, information security, intelligent migration, text analytics, and collaboration.

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.

Organizing Content in a Universe of Complexity

The problem facing almost all enterprises is not the lack of information but the inability to connect, categorize, and analyze information to improve organizational performance. The overabundance of content is surpassing the frameworks and controls in most enterprises. With over 80% of business decisions being made using unstructured content, maximizing the use of information capital has become a key source of competitive advantage, business agility, and decision making.

Enterprises either ignore the problem, implement technologies they hope will solve the problem, or piece together a solution with the technologies they have. conceptTaxonomyManager is a radically different solution that delivers the ability to manage content and enable enterprises to maximize their information capital to deliver business results. Concept Searching technologies deliver automatic intelligent metadata generation, automated classification and taxonomy management. As opposed to traditional tools, the results are transformed into intelligent metadata enabled solutions that are rapidly deployed, easy-to-use, and deliver unique capabilities not available from any other technology. This flexibility extends to addressing business process failures in records management, information security, migration, text analytics, and collaboration.

How Do You Know What You Don't Know?

Enterprises commonly set boundaries and processes to control the flow of information to ensure best practices in a variety of applications such as records management, data protection, and compliance. This approach to control the flow of information is often cumbersome and impacts the ability to manage content throughout the information lifecycle to include: capture; storage; retrieval; archival; and disposal. The crux of the problem is the inability to capture accurate information that enables all the subsequent steps to be completed correctly.

"At least 80% of enterprise information is unstructured and growing at over 100% per year."

Gartner Group

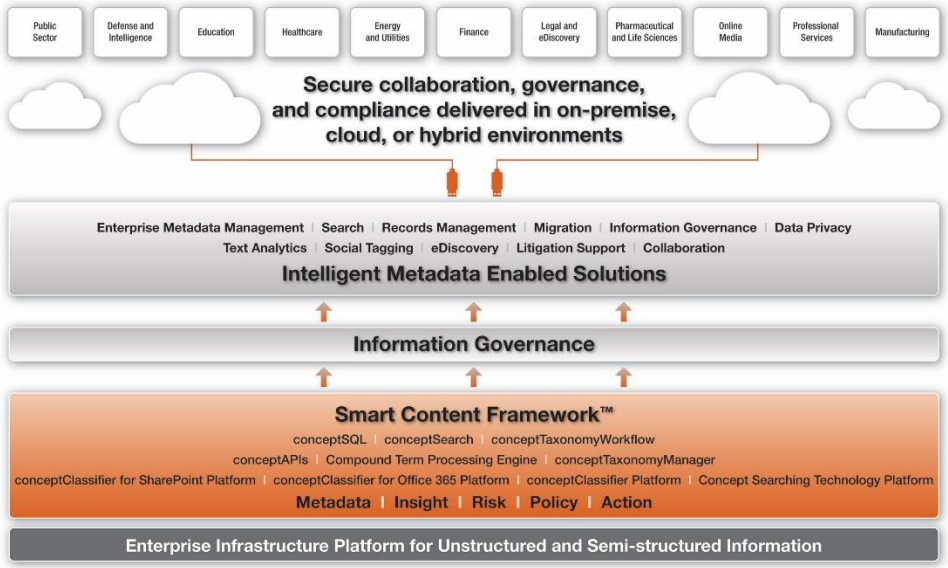
conceptTaxonomyManager has the capability to automatically group unstructured content together based on an understanding of the concepts and ideas that share mutual attributes while separating dissimilar concepts. This approach is instrumental in delivering relevant information via the taxonomy structure as well as using the automatically generated semantic metadata in enterprise search to reduce time spent finding information, increase relevancy and accuracy of the search results, and enable the re-use and re-purposing of content. Using one or more taxonomies, unstructured content can be leveraged to improve any application that uses metadata.

Making Sense Out of Chaos – The Smart Content Framework™

Concept Searching's proven approach incorporates its Smart Content Framework™, developed as a set of best practices that provides the enterprise framework to leverage unstructured content through a managed framework. The Smart Content Framework™ is a multi-disciplinary approach delivered through the Concept Searching technologies that encompasses the entire portfolio of information assets resulting in increased organizational performance and agility. The framework has proven to be a flexible solution to address recurring challenges in applications and processes, impacting organizations of any size or industry.

Underlying the Smart Content Framework™ is the ability to generate intelligent metadata, transparently tag content, and classify it to organizational taxonomies. The framework is being used to improve search, in records management, enterprise metadata management, identification of data privacy and confidential assets, eDiscovery, Legal, migration, collaboration, text analytics, to provide the organizational framework for compliance, content management, and information governance.

Providing a complete solution including intelligent metadata generation, automated classification, and taxonomy tools results in a flexible approach that significantly improves management and access to unstructured content.



Metadata Matters – Intelligent Metadata Generation

A taxonomy and metadata have a co-dependent relationship. The structure of the taxonomy and the metadata are reciprocal elements that work together to create the information architecture. Taxonomies provide the visual organization and structure for organizing content which metadata does not provide. At the same time, metadata provides more descriptive information about the content to improve access and use of the content. Intelligent metadata generation results in improving workflows and business applications that use metadata.

Intelligent metadata enabled solutions are delivered as outputs of the Smart Content Framework™. The intelligent metadata is generated as content is created or ingested and identifies how the data items are related as well as the meaning of the content, by Concept Searching's unique compound term processing engine. The intelligent metadata can be a single word term or multi-word patterns. These patterns can identify concepts based on one, two, or three words and occasionally four or five words.

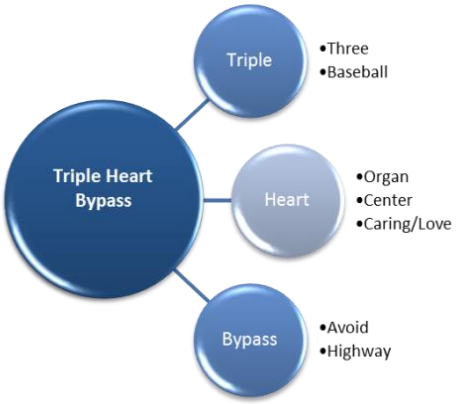
"The metadata infrastructure provides the critical glue that binds the information infrastructure to the underlying IT infrastructure. Sound information governance practices would take advantage of the metadata infrastructure, to ensure that content and data are managed consistently and adhere to written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

IDC
Digital Universe Study 2010

Compound Term Processing

Utilizing Concept Searching's compound term processing, the technologies deliver a set of outcomes that are not achieved by any other classification engine. Compound term processing means that Concept Searching's statistical engine can understand, out-of-the-box, the incremental value of keywords, multi-word fragments, and compound terms and as a result identify concepts resident within an organization's own information repositories that are highly correlated to particular topics. With the identification of these highly correlated topics in the form of keywords, multi-word fragments and compound terms the result is automatically generated intelligent metadata that is unique to that particular organization.

The example below illustrates how compound term processing impacts metadata generation. Many words have multiple meanings, and if they exist in a document the search engine does not differentiate the various meanings and metadata is generated as keywords. Below, the word triple, heart, and bypass all have different meanings.



Using compound term processing a search for "survival rates following a triple heart bypass" will locate documents about this topic even if this precise phrase is not contained in any document. A concept search using compound term processing can extract the key concepts, in this case "survival rates" and "triple heart bypass" and use these concepts to select the most relevant documents.

The technology enables the rapid creation of semantic metadata, which can be classified to organizationally defined taxonomies. The tagging and auto-classification of content can be aligned to business goals and the semantic metadata generated can be easily integrated with any third party application or platform that can interface via web services.

In addition to generating keywords, acronyms, compound term processing identifies multi-word terms that forms a complex entity and identifies them as a concept. By placing these compound terms in the search engine's index, or making them available to any application that requires metadata, the outcomes are highly accurate, because the ambiguity inherent in single words is no longer a problem.

"To get a better sense of just how much data are going unused, The Economist Intelligence Unit asked survey respondents to estimate their data efficiency. The results are surprising: 24% say that vast quantities of data go unused at their company, and 53% use only about half of the data that is of value. Only 22% of respondents say that they are putting nearly all their data that is of real value to good use."

Big Data – Harnessing a game-changing asset
Economist Intelligence Unit

Solving Business Problems

Most large and medium sized organizations have invested in content management, search, and portal technologies but end users still struggle with finding the right content at the right time and the right context. Typically associated with improving search, a taxonomy can also provide a consistent information infrastructure that can be shared across different applications and business divisions through an enterprise metadata repository.

conceptTaxonomyManager can be used for a wide range of applications. In records management the reason most cited for failure is end user acceptance to appropriately tag documents of record. Although an organization may have a robust retention schedule it is the end users who ultimately make it a success or failure by applying appropriate tags. Concept Searching technologies can automatically declare documents of record, generate intelligent metadata, and automatically classify the content as it is created, effectively enforcing governance at the desktop.

Corporate compliance initiatives cover a wide range of laws such as HIPAA, Sarbanes-Oxley, ITAR, and federal mandates. The processes to identify these potential non-compliant exposures need to protect the organization, reduce risk and legal ramifications. conceptTaxonomyManager can be used to automatically identify potential exposures within unstructured content. The taxonomy creates the standard for all content within the organization regardless if it is used for search, records management, compliance, data protection, text analytics, or eDiscovery/FOIA.

The following illustration shows how an enterprise metadata infrastructure delivered through the taxonomy can impact a variety of applications.

"Failure of ECM initiatives is being measured as an overall sense that we did not get what we paid for. We should not be surprised, considering most companies are going about ECM with the same approach that produced disappointing results before."

AIIM

conceptTaxonomyManager Features and Functions

Concept Searching technologies combine all the requisites to build a scalable enterprise metadata infrastructure to provide automatically generated intelligent metadata, auto-classification, and taxonomy management.

Concept Searching's, conceptTaxonomyManager is a simple yet powerful tool with an intuitive user interface designed for Subject Matter Experts (SMEs)without the need for IT or Information Scientists expertise to build, maintain and validate taxonomies for the enterprise. This facilitates the rapid creation of taxonomies and decreases the resources needed for ongoing management. The product was purposely designed for ease of use by SMEs. Interactive and still unique features not available in any other product include: automatic clue suggestion, document movement feedback, automated classification, and automatic generation of conceptual metadata. The features of conceptTaxonomyManager include the following:

- Compound Term Processing technology that identifies 'concepts in context' *(Unique to Concept Searching)*
- Automatic intelligent metadata generation as content is created or ingested *(Unique to Concept Searching)*
- Automated classification of content to one or more nodes in one or more taxonomies

- Rules based engine that eliminates the need for training sets and highly specialized human resources *(Unique to Concept Searching)*
- Easy to use by SMEs, providing the ability to build taxonomies and metadata models by knowledge workers not Information Scientists, or IT
- Aggregates unstructured and semi-structured content from multiple content sources
- Taxonomy management rapidly deployed and easily managed
- Controlled vocabularies
- Multiple taxonomy support
- Supports polyhierarchies and ontological relationships
- Automatic taxonomy node clue suggestion *(Unique to Concept Searching)*
- Dynamic screen updating to immediately see impact of changes in the taxonomy *(Unique to Concept Searching)*
- Document movement feedback to see cause and effect of changes without re-indexing *(Unique to Concept Searching)*
- Fully SOA compliant services for automatic classification and taxonomy management, delivered as web parts
- Supports 55 languages
- Adheres to industry standards such as OWL or RDF
- Import tool for industry standard taxonomies such as MeSH
- Provides working sets for each term enabling the taxonomy administrators to finely tune rules, by excluding false positives and including elite documents
- Calculations feedback showing why a document was classified or not classified and provides the scoring of the term that the system, or taxonomy administrator, assigned
- Boosting mechanism typically used for developing extremely complex taxonomies, the feature understands the importance of the taxonomy hierarchy in that each term does not exist in isolation, but is impacted by the hierarchy and classification results of its near neighbors. The boosting feature can be turned on or off for a specific term, or for the whole taxonomy
- Full security model enabling lock down of nodes, branches and complete taxonomies to particular users and/or groups of users
- Supports roll back to previous state
- Highly scalable, tested with taxonomies with 250,000 preferred terms and over 2,000,000 non preferred terms
- Data held in standard SQL/Oracle enabling BI tools to be layered over the data to build reports and dashboards
- Performs enterprise class Term Store Management when integrated with all versions of SharePoint, Office 365, and OneDrive for Business
- Available on-premises, hybrid, cloud environment
- Platform agnostic

Industry Unique Features

conceptTaxonomyManager remains unique in the industry in features that provide the ability to rapidly and easily change the taxonomy as the organizational needs and requirements change. This is important, as a taxonomy must remain fluid as opposed to static and must be managed in a way that easily facilitates change.

Auto-Classification

The value of classification spans a broad set of application uses. Classification fundamentally provides the organization improved decision making capabilities. Content is dynamic and the taxonomy should be flexible to change as business strategies and structures change. The classification process adapts to the organization as content is changed, moved, or deleted. The taxonomy coupled with automated classification form the foundation to realizing the benefits of ECM; in fact all content centric applications will realize business benefits by leveraging the capabilities of the taxonomy.

Concept Searching's automated classification process identifies during indexing categories that each document belongs to. Each category is identified by a unique descriptor and is associated with key descriptive words and/or phrases held in the database. This approach enables a rapid implementation of a corporate taxonomy with all documents classified to multiple nodes at index time. Ideally, the taxonomy can be used to browse the document collection or as a filter when running ad hoc searches.

Any document can be classified against one or more taxonomies as an automated background process or optionally with the user being given the option to review the suggested classification and make changes. conceptTaxonomyManager is a simple to use, intuitive user interface designed for Subject Matter Experts (SMEs) to build, maintain, and manage taxonomies. This easy-to-use taxonomy and automatic classification tool creates the framework to classify content based on concepts to one or more nodes in the taxonomy or multiple taxonomies.

Auto Clue Suggestion

Eliminating complex Boolean rules and the need for training sets, which typically limits scalability, the taxonomy nodes can be automatically generated from the compound terms found in the document corpus. The SME has full control of the terms to be used as well as the weighting of the term based on its relevancy. This enables a much more robust taxonomy as the terms are suggested based on the organization's own content and can offer the SME new terms from the relevant documents that may not have been identified.

The Clues can also be assigned a Score or weight, either positive or negative to improve the classification. Clues can also be assigned a Type. Types include standard, case-sensitive, metadata, phonetic, regular expression, required term, term boost, and language.

Document Movement Feedback

Automatic document movement feedback enables the SME to see the cause and effect on changing the clue weightings for a node in the taxonomy. The SME can also search within the refined node and bring back documents from the whole corpus now classified against the node. The system will indicate if the change has increased the score, reduced the score as well as identify documents that will no longer be classified and new documents that will be classified.

This feature is also used in working sets. Working sets enable the administrator, taxonomist, or SME to tune the rules within conceptTaxonomyManager to include or exclude certain documents.

This is to ensure that false positive are not getting classified and elite documents are getting classified. The option to add or remove a document from a working set is also provided. Each term can have a set of documents associated with it for testing purposes. Using a unique feature in conceptTaxonomyManager, 'Show Document Movement Feedback', in conjunction with the working sets visually shows the cause and effect of the changes without re-indexing.

Many clients have millions of documents and the ability to see the changes without re-indexing is a significant benefit, saving time, increasing productivity, and increasing the accuracy of the classification process. Within conceptTaxonomyManager, the administrators have access to a calculations link that shows why a document received the score it did, how many times the terms appeared in the document, and additional information to provide an understanding of the classification and provide the ability to change the score.

Distributed Taxonomy Development
This feature is a requirement for organizations that have many taxonomy operators, extremely large collections of documents, and where taxonomy management is a critical business process. This feature can be implemented on any number of servers and several taxonomy managers can be assigned to a server to ensure the level of throughput needed. Real time locking mechanisms are used to make nodes of the taxonomy inaccessible to other taxonomy managers while the node is being edited. The taxonomy managers can visually see when a node is locked and who has locked it as well as when it becomes available. The Distributed Taxonomy Management feature is totally transparent to the end user and all locking and unlocking of the nodes by the taxonomy managers are coordinated by the central server.

Security and Roll Back
The product provides a full security model enabling lock down of nodes, branches, and complete taxonomies to particular users and/or groups of users. Also supports roll back to the previous state.

Integration with SharePoint Suite of Products

Concept Searching's SharePoint Suite of products have been developed to run natively in SharePoint and are fully integrated with all versions of SharePoint, Office 365, and OneDrive for Business. The award winning platforms include the taxonomy manager component as well as the automatic semantic generation, and auto-classification engine. The versatility and integration of the technologies provides an enterprise with the ability to identify all content and make it available to any enterprise application that needs access to unstructured information.

With the Term Store functionality in SharePoint on-premises and SharePoint Online organizations can develop a metadata model using out-of-the-box SharePoint capabilities. Running natively and fully integrated with the Term Store, the technology can consistently apply conceptual metadata to content and auto-classify to the Term Store metadata model solving the challenge of applying the metadata to thousands of documents and eliminating the need to depend on the end user community to correctly tag content. The taxonomy manager component functions bi-directionally with the Term Store where changes can be made in the Term Store or in the taxonomy manager. This added functionality assists in expediting the development of the metadata models, offers sophisticated refinement capabilities, and significantly reduces on-going maintenance.

conceptTaxonomyWorkflow

conceptTaxonomyWorkflow is an add-on component that can be deployed in SharePoint and non-SharePoint environments. conceptTaxonomyWorkflow also serves as a strategic tool managing migration activities and content type application across multiple SharePoint and non-SharePoint farms, Office 365, and OneDrive for Business.

With conceptClassifier and the conceptTaxonomyWorkflow module organizations can automate the tagging of unstructured documents to deliver enterprise specific automated processes.

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

To address this issue, conceptTaxonomyWorkflow works with conceptClassifier to automatically apply correct content types when organizationally defined descriptors and vocabulary are identified within documents.

conceptTaxonomyWorkflow bypasses manual processes with the SharePoint Content Organizer and automatically applies the correct content types based on managed metadata properties. As a result, the combination of the technologies delivers a unique and powerful solution leveraging SharePoint content types.

conceptTaxonomyWorkflow deploys 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.

Content can be tagged and classified content to locations both within and outside SharePoint resulting in:

- Efficient automated migration of large volume projects
- Proven high performance architecture for throughput, multi taxonomy, multi-site requirements
- Accurate, consistent and automatic classification using conceptClassifier for SharePoint or conceptClassifier
- Comprehensive integration with SharePoint Managed Metadata Services, writing directly to the Term Store locations in real time
- Primarily used to identify and prevent data privacy/confidential exposures, to automatically declare documents of record for records management purposes, and for intelligent migration

Technology

The technologies are based on an open architecture with all APIs based on XML and Web Services. Transparent access to system internals including the statistical profile of terms is standard. Products include a Service Oriented Architecture (SOA) based search and classification technology, a browser based taxonomy management technology, and a tightly integrated feature set that operates with any search platform.

Tangible Return on Investment

Deployed at global organizations and organizations that place high value on content assets, conceptTaxonomyManager has been proven to deliver a highly scalable and flexible approach to effectively manage unstructured and semi-structured content. Industry unique concept identification enables the creation of organizationally defined taxonomies, reducing taxonomy development time by 80% as compared with competing products *(client source data)*.

The tagging and auto-classification of content can be aligned to business goals, and the semantic metadata generated can be easily integrated with any search engine or third party application that can interface via web services. The ease-of-use and interactive features assist the Subject Matter Expert (SME) in developing and managing the taxonomy results in rapid deployment and reduced costs.

The unique compound term processing capability, automatic generation of semantic metadata, and automated classification enables the organization to address a wide range of challenges and improve business processes. Delivering a quantifiable return on investment, enterprises are using the technologies to build and deploy an enterprise metadata repository that is consistent, scalable, and manageable; protect organizations where compliance is mandatory; reduce the costs associated with poor findability in search; ensure governance at the desktop through elimination of manual tagging, and facilitate the deployment of intelligent metadata enabled solutions.

Regardless if an organization only wants to improve search or needs to address multiple organizational challenges for applications that use metadata, conceptTaxonomyManager is a proven solution that is instrumental in achieving objectives.

About Concept Searching

Concept Searching is the industry leader in advanced semantic metadata generation, auto-classification, and taxonomy management. Its award winning products are the only statistical metadata generation and classification technologies that use compound term processing to generate intelligent metadata from unstructured and semi-structured data. Compound term processing, or identifying 'concepts in context', solves a variety of business challenges. Using the concept identification capabilities, organizations can transform content into business assets to improve performance.

Concept Searching's Smart Content Framework™ for information governance is a combination of best practices and underlying products that encompass the entire portfolio of unstructured information assets, resulting in increased organizational performance and agility. The output from the Smart Content Framework™ delivers intelligent metadata enabled solutions that are being used to enable concept based searching, automatic declaration of documents of record, identification and protection of privacy and confidential data, intelligent migration, content management, granular identification of content for text analytics, and improved delivery of social content. The solutions are deployed in diverse industries, Fortune 1000 companies, and smaller companies that need to meet strict compliance, data privacy, and information governance regulations.

Concept Searching has a Microsoft Gold Application Development competency and is a participant in the global Business-Critical SharePoint program. Although platform independent, the Concept Searching Microsoft suite of products uses a single code base, supporting all versions of SharePoint, SharePoint Online, and OneDrive for Business, providing clients with the choice of on-premises, cloud based, or hybrid environments to best meet their needs.

Headquartered in the US, with offices in the UK, Canada, and South Africa, Concept Searching solves the problem of finding, organizing, and managing information capital. For more information about Concept Searching's solutions and technologies please visit www.conceptsearching.com and follow on Twitter and LinkedIn.

Microsoft Partner
Gold Application Development

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)21 813 9633
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