



SOLVING GOVERNMENT CHALLENGES IN INFORMATION SEARCH & CLASSIFICATION

CONCEPT SEARCHING

This document discusses the challenges and obstacles facing government agencies in the organization and effective retrieval of information and the tools available to reduce costs and meet knowledge management challenges.



Introduction

Although technologies can improve government Information and Communication Technology (ICT) initiatives the real goal is to deliver value, whether internally or externally to a variety of constituents. The challenge has been to identify technologies that can be deployed to create a transformation in the culture and structure of government in order to create sustainable and measurable benefits. Having made great strides, the public sector still falls behind the private sector in focusing on knowledge management initiatives that improve organizational performance through increased efficiency and innovation. Government entities are realizing the importance of the management of information capital to address the growing challenges created by the explosion of digital content. The essence of knowledge management is to provide strategies to get the right knowledge to the right people at the right time and in the right format¹.

eGovernment

The U.S. e-Government Act of 2002 intent is to improve the management and promotion of electronic government services and establish a framework of measures that utilize the Internet to improve citizen access to government information and services. Section 207 of this Act states that the requirements for federal web sites must address “the speed of retrieval of search results, the relevance of the results, and tools to aggregate and disaggregate data”. The U.S. Federal Enterprise Architecture (FEA) and Data and Information Reference Model (DRM) defines the structure that facilitates the development and effective sharing of government data across communities of practice and lines of business. This framework consists of a model for the discovery of information, a model for the exchange of information, and a model for the representation of information. Although these models serve as the ultimate goal for government agencies, identifying the optimal tools to deliver the results has remained a challenge.

The primary delivery models in e-Government are Government-to-Citizen or Government-to-Customer (G2C), Government to Business (G2B) and Government-to-Government (G2G) & Government-to-Employees (G2E). Within each of these delivery models, information must be available via the Internet, enable two-way communication, provide transaction capabilities, and governance. Delivering these initiatives can derive benefits of improved efficiencies, convenience, and better accessibility of public services. However, there are many considerations and issues required in the implementation of an e-Government solution that effectively serves the diverse audiences, meets the objectives of the e-Government Act of 2002 as well as the agency’s objectives.

e-Government challenges are not only technical. e-Government initiatives must be organized and delivered based on the needs and preferences of the customer not according to government administrations. To be successful, public administrations must be willing to transform the culture of government and focus on using ICT to change the culture of government. Providing seamless integrated online services enables users to interact with the government as a single organization. This approach can provide higher value to customers than on-line access to disparate government sites. This customer focused approach requires collaboration and co-operation between agencies and as services become more complex, this collaboration and exchange of information can increase the efficiency of government as a whole.



Connected Governance

Connected Governance is a systematic approach to collection, reuse and sharing of data and information². In the framework of Connected Governance, intergovernmental processes and access to information can be integrated vertically between various government agencies and/or horizontally between agencies at the same level and/or with the inclusion of private sector and other stakeholders. The end result is better organized and integrated access to information and knowledge reduces costs and improves the effectiveness and efficiency of government. To achieve this, agencies must identify content and data spread across multiple systems and effectively organize the information to enable the reuse and sharing of that information to meet the needs of multiple constituents. Although taxonomy tools can assist the agency in the organization of the content more effective tools to further define the content and identify relationships between information are needed.

Federated Data

Sharing data across physical boundaries and managing and accessing that information, data, and metadata in a secure manner is a critical component to delivering effective and efficient government performance. Regardless of the type of boundary, federated data management provides a method to deliver data in a uniform and consistent manner so that the content can be shared among the government entities. This type of environment allows users to seamlessly access and find relevant information regardless of where the content is physically stored. The ability to automatically integrate content from diverse sources and deliver that content to the end user where it can be searched, accessed and integrated into one interface is a challenge. In all cases, although content may be visible to a wide range of users, the content maintenance and management of these assets need to be owned by the responsible agencies.

FOIA

The U.S. Freedom of Information Act (FOIA) is a law ensuring public access to U.S. government records. FOIA carries a presumption of disclosure; the burden is on the government, not the public, to substantiate why information may not be released. Upon written request, agencies of the United States government are required to disclose those records, unless they can be lawfully withheld from disclosure under one of nine specific exemptions in the FOIA. Recently reformed, it is still unclear how the Office of Government Information Services (OGIS) will address the persistent problems in the FOIA system including excessive delay, lack of responsiveness, and litigation by federal agencies.

There are two aspects of the FOIA that impacts federal agencies. The first is to make appropriate information available to reduce the number of FOIA requests without publishing information that should remain classified. Secondly, to provide the agency staff with the ability to find the information to fulfill the FOIA request in the time-frames required. These requests need to be reviewed line by line and may include thousands of pages of documents.

The US government currently creates thousands of classified documents, in fact over 20 million classification decisions were made in 2006. In addition, there are currently classified documents that will become declassified according to the law which allows release after a pre-determined time period. The present process for classifying documents is both time consuming and labor intensive. Document review and searching through the document to identify material called out in the classification guidelines is arduous and sometimes complex. Proper document marking of the security classification may take a few hours or several weeks, depending upon the document length and complexity of the classification guidelines.

The Costs

The challenge of managing, finding and purposing content to meet the needs of diverse groups within a government entity impacts their ability to successfully meet their objectives. Some of the inherent issues include:

The inability to share or exchange data efficiently costs government agencies time and money. These inefficiencies result in increased time spent on finding and accessing relevant information, if it can be found at all.

The inability to satisfy citizen and stakeholder requirements places additional burdens on agency staff to meet the ever increasing information access requests of diverse constituents. Further impacting the agency is the inability to identify and/or contact the resource for the right data.

As more and more content is created and published the agency faces increasing costs to manage and integrate content so it is available within the entity or potentially across entities.

Responding to FOIA requests in a timely fashion and ensuring government compliance is laborious and costly to agencies. Furthermore, the classification and declassification process is often inefficient and requires considerable human resources and costs to manage and perform the process.

In agencies where access to current information is mission critical and must be aggregated from internal and external sources the impact of not having access to these information resources carries far greater implications that can affect the safety and security of the country.

The Concept Searching Approach

Concept Searching is the only statistical semantic 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 tool set provides advanced search, semantic metadata generation, automatic classification and taxonomy management.

The technologies provide the framework to enable agencies to build robust taxonomies that leverages content from both internal and external sources as well as provides multiple levels of granularity to enable the reuse of data from multiple stakeholder views. The software has the ability to automatically identify new unstructured content and generate semantic metadata (concepts) within the content and classify the documents to the taxonomy. Providing a variety of methods to search and retrieve content based on concepts dramatically improves access to relevant content for diverse users.



Semantic Metadata Generation

The discovery, collection, and management of metadata is essential for the integration of content across disparate systems. The primary issues are the lack of metadata associated with the content and the relating of content in one system to similar or equivalent content in a different system. There is a growing need within the public and private sector to generate far richer metadata and manage it effectively to provide enhanced access to these resources by individuals. The lack of a common and consistent way to describe or define unstructured content contributes to the inability to share information and results in duplication rather than reuse. This stove-pipe approach has created boundaries around the information making it a challenge to share information or to make it available to internal users who require the information to complete their tasks.

Concept Searching can automatically generate semantic metadata based on the concepts within unstructured information. The generation of semantic metadata enables the agency to extract compound terms, acronyms, and keywords from a document or corpus of documents that are highly correlated to a particular concept or metatag. 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. When the same concepts are prevalent within a particular document that document is automatically meta-tagged, eliminating the requirement for an individual to read that document and subjectively apply metadata to the properties of the document. This ability to identify 'concepts in context' eliminates inconsistent or non-existent tagging processes and overcomes different publishing conventions that may exist within the agency.

The Need for Taxonomies

Due to the explosion of electronic content, government agencies need to develop a comprehensive approach to identify, organize and retrieve content assets. Internal communities need to find and reuse content rather than recreate it or make do without it. Implementing a framework that enables identification of content for reuse provides a return on investment for the time and effort spent to originally produce the material.

Government agencies may have a wide variety of constituents that need access to content to meet different needs. Internal vocabularies are often specific to that agency and may not be easily translated by personnel outside of a particular community rendering the content unusable. Further complicating matters is that within an agency there may be varying solutions for identifying and storing electronic documents. The inconsistency of these systems hampers the ability of users to find relevant information, specifically when searching across multiple silos of content within an agency. Although knowledge workers need unified and universal access to information, at a more granular level they need to be able to find exactly and only the content they need.

Concept Searching's robust automatic classification and taxonomy management tools are designed to provide as much depth or hierarchical granularity that is often needed in government agencies. Since the automatic semantic metadata generation has the ability to identify concepts as opposed to keywords, documents with the same concept can be classified against multiple nodes within a taxonomy or multiple taxonomies. From an end user perspective, knowledge workers can locate pertinent information from his or her own individual viewpoint without knowing the exact search terms to use. The easy-to-use taxonomy and automatic classification features function as a labeling mechanism to quickly create the foundation that can be altered to suit the requirements of the agency.



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 can 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 (SME's) can utilize rich features such as node weighting, ability to see the 'concepts in context', ability to search the corpus, auto-clue suggestion for categorization, and instant feedback on the impact of changes. Traditional taxonomy tools often require significant investments in time, expertise, and money to develop and maintain. Concept Searching's taxonomy management tool has been proven to reduce the time to build and subsequently maintain taxonomies in government entities by up to 80%.

conceptSearch

Search engines utilize complicated algorithms for the frequency, location, and proximity of words and phrases which do not necessarily retrieve the information the user is seeking. Retrieving content based on a user entered search string often retrieves irrelevant content based on the keywords (metadata) associated with the document. Ensuring all content is properly and consistently tagged still does not guarantee that the information retrieved is relevant and is an even greater challenge to organizations that typically do not enforce governance of tagging documents as they are created.

Knowledge workers need to identify content in the context of what they are seeking. The search engine must look beyond the ambiguity of natural language and identify the content fragments they require to solve the problem they are facing at that moment.

The ability to search on concepts as opposed to keywords for relevant information within an agency can be greatly improved. Concept Searching has developed the only software that delivers both high precision and high recall. This is done by weighting compound terms (multi-word phrases) instead of single words used in isolation. A key feature to assist the end user is dynamic summarization. In traditional search, when a document is retrieved, the end user is shown a static summary that is the same regardless of the user's query. conceptSearch provides dynamic summarization, where an extract of the document is displayed as an aid to the user. These extracts will comprise whole sentences or short paragraphs and can be configured to meet the needs of the agency. This is particularly useful as the users can visually see the corresponding results of their query within the document saving considerable time in determining relevancy.

Presenting relevant information to the various constituents through effective search is enabled via taxonomy based navigation or through faceted navigation. Taxonomy based navigation dramatically improves the search experience⁶. Faceted navigation is a logical extension of the taxonomy. The end user controls the search experience and the search results present 'facets' of documents grouped together based on the concepts identified. These facets extend the search process as documents will be grouped by concepts and assists the end user by offering content that may not have been found. This unified view and access to relevant information across disperse silos of information can increase productivity and enable knowledge workers to effectively query, use and re-use agency wide content.

Governance at the Desktop

Additional functionality is provided to enable knowledge workers to automatically classify content from within the traditional Microsoft Office interface. This can be done automatically, or optionally the knowledge worker can add manual adjustments to the classification to provide further refinement if required. This ensures that the organizational content is consistently classified and available to others as needed.

Summary

Concept Searching technologies address many of the challenges facing government entities. The tools can assist agencies in reducing costs and increasing productivity. Allowing knowledge workers to effectively query, use and re-use agency wide content improves the speed and efficiency of operations and information can be shared and leveraged throughout the business cycle. The elimination of inconsistent tagging and different publishing conventions across multiple content stores provides access to content from internal and external sources. For the agency significant benefits can be achieved by removing the ambiguity in content through the identification of concepts within a large corpus of information. Concept Searching's solutions can be the catalyst to improve access to unstructured information, encourage innovation, and deliver real benefits to government entities, their constituents, and stakeholders.

About Concept Searching

Founded in 2002, Concept Searching's software products deliver advanced search, auto-classification, taxonomy management and advanced metadata tagging solutions from the desktop to the enterprise. Concept Searching is the only 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.

Headquartered in the U.K. with offices in the U.S. 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.



References

¹ Milton, N. Shadbolt, N. Cottman, H. and Hammersley, M. (1999), "Towards a knowledge technology for knowledge management", International Journal of Human-Computer Studies, Vol. 51, pp 615-641.

² United Nations E-Government Survey, 2008, "From E-Government to Connected Governance".

³ Hao Chen & Susan Dumais, 'Optimizing Search by Showing Results in Context'.

Europe
9 Shephall Lane
Stevenage
Herts SG2 8DH, UK
P: 44 1438 213545
info-uk@conceptSearching.com

Americas
8300 Greensboro Drive
Suite 800
McLean, Virginia 22102 USA
P: 1 703 531 8567
info-usa@conceptSearching.com

South Africa
15 Conifer Road
Tokai, 7945
Cape Town, South Africa
P: 27 21 7125179
info-sa@conceptSearching.com