

CIRCA Technology Overview

Applied Semantics *White Paper*



TABLE OF CONTENTS

A.	Introduction	3
B.	Bringing Order to Unstructured Information	3
C.	Enabling Computers to Understand Conceptual Information	4
D.	Other Approaches to Information Management	5
E.	An Ontology-Based Approach to Information Management	5
F.	Applying Conceptual Information to Solve Business Problems	7
G.	The Benefits of CIRCA Technology	8
H.	Conclusion	8

A. Introduction

Technological advancements in the last decade have enabled computer systems to record and put to work increasingly greater amounts of data. Systems designed for the creation, aggregation, storage, and communication of information have scaled accordingly. However, systems that organize this information have not scaled, resulting in disorganized data that users cannot locate and that cannot be put to effective use.

This problem is exacerbating the already existing information overload, creating a vicious circle for enterprises dependent upon information. To maximize the value of the unstructured information that computers are now collecting, enterprises need to focus on imposing order on their chaotic repositories. Adding this critical order can be accomplished in several ways: improving search technologies, automating document classification and categorization, and instantly generating summaries that simplify the navigation experience for users skimming through long lists of potentially relevant documents. In addition, these efforts must be precise and automated to effectively keep pace with the explosive growth in unstructured information in order to meet the ultimate goal of making all stored data locatable.

B. Bringing Order to Unstructured Information

Two schools of thought have emerged regarding how information can be organized to make it locatable. The first set of efforts focuses on creating open, common standards for imposing structure, such as "subject," "author," and other descriptive fields, on otherwise unstructured information. Current eXtensible Markup Language (XML) schemas are one such attempt to formalize and standardize methods for structuring information. However, even once standards are finalized, much of the problem relating to unstructured information will remain: individuals may agree on what to call the descriptive fields, but there will always be great variation in the contents of those fields. For example, an XML schema may call for a subject field, which would add structure to a document, but the person inputting the contents of that field could select any string of words to fill the field. This new descriptive information would be shorter in length than the full text of the document, but just as unstructured as the document it was intended to improve.

The second set of efforts aims to improve the ability of computer systems to automatically understand the concepts contained within a string of text, whether within an XML field or an entire document, and understand how one set of information relates to other associated, but not identical, strings of information. At present, computers successfully analyze the letters within a string of text and compare these signatures against other strings of letters. Computers do not comprehend the concepts expressed in information because they are not currently equipped to understand how data relate to other data, or how to perform complex operations on two sets of unstructured information.

The human mind understands how the ideas expressed in a document connect to other ideas outside of the document. For example, the concept of a "subject" field, as discussed above, is sometimes ambiguous, yet the human mind is able to interpret the ideas behind a piece of text, formulate thoughts about it, and select an appropriate

subject under which it should be classified. This ability is what makes it possible for a person to read a document entitled “The Volatility of Financial Markets,” determine that it is about rapidly changing equity markets, and that the document relates more to “investing” than to “volatile chemicals” or “farmers’ markets.” Humans understand information and search for ideas, not text strings. To date, computers have not been able to naturally understand the conceptual relationships in unstructured information. Humans using computers to automate their searching and organizing of documents have been forced to rely on technologies that process text strings only. To make information locatable, and to organize information effectively, computers need to understand and be able to process *ideas*.

C. Enabling Computers to Understand Conceptual Information

In the past, searching through data has forced users to understand and work within the organizational structures imposed upon information. For example, every American child learns at an early age to understand the Dewey Decimal System in order to locate information within a library. In other words, humans have been forced to become library-literate. Conversely, if a library could organize information conceptually, the library would become human-literate allowing for flexible searching that organized information based on the ideas relevant to a search rather than on a numbering system.

Imagine standing at the entrance to a library with one task at hand: to research and write a report on “technological improvements in the diamond mining industry over the last one hundred years.” The task would seem forbidding because the way that the information has been organized in the library does not at all match the demands of this assignment. Mining technology books have been segregated from those on geology; the material on the diamond business is in a different section from the biographies of those instrumental to the industry or the historical works detailing how the industry changed over the last hundred years. Yet, all of these broad disciplines likely contain pieces of the information necessary to complete this assignment.

Imagine being able to press a button that causes the library to automatically rearrange its shelves so that it brings together all the books linked to the technological improvements in the diamond mining industry. The volumes of material available within the library would be easily locatable through a technique that relates one item to another based on the concepts touched upon in these books. The broad categories under which books are normally filed would cease to be a hindrance to the multi-disciplinary searcher.

Interpreting unstructured information and determining what concepts lie behind this information make it possible for a computer, if not a library, to understand how ideas are related to other ideas. Giving a computer access to this conceptual information can add functionality to any knowledge discovery process, such as searching, indexing for later retrieval, categorizing, creating metadata, or generating summaries for business users to read.

A pre-requisite for equipping computers to process unstructured information as easily as they manipulate numerical information is access to an ontology. An ontology is compressed knowledge broken down into its core components. It often takes the form of

an extremely large database of words and phrases, their meanings, and their conceptual relationships to other concepts. Search, knowledge management, and other software equipped with an ontology improves a computer's ability to automatically relate one concept to another and begin to treat information as a composition of ideas that can be linked to ideas residing within different sources of information.

D. Other Approaches to Information Management

The most widely accepted approaches to information management, commonly referred to as text-based techniques, focus on allowing a computer to recognize rules and patterns in text. A rules-based approach enables an enterprise to create highly detailed guidelines for how documents are processed based on their content. Statistical pattern-matching approaches artificially construct simple relationships between words by analyzing a training set of documents. Each of these approaches is an attempt to make unstructured information more comprehensible to a computer rather than making a computer able to understand unstructured information.

In addition, these approaches are hindered by two main limitations. First, they are difficult to implement. Rules-based systems require extensive analysis of existing documents and the creation of very complex rules for future document analysis while pattern-matching approaches require the costly effort of creating a sizeable, clean training set of documents for the system to "learn" upon. The second limitation is that these approaches are difficult to maintain. With rules-based and pattern-matching systems, what you know today will dictate what information you are able to process in the future. Rules created today determine how all future documents, however the content of the documents change over time, will be processed.

Neither one of these approaches offers the ability to differentiate between various meanings behind every word within a document or allows continual development of the relationships between the meanings, rendering them highly inflexible. These inflexible technologies require enterprises to completely re-train their systems as their information needs change over time.

E. An Ontology-Based Approach to Information Management

Applied Semantics has decided to employ an ontological approach to address many of the limitations inherent in text-based information management technologies. Not a simple taxonomy, the Applied Semantics Ontology goes beyond parent and child relationships and captures the richer relationships between terms. These include lateral and equivalence relationships as well as the strength of relationships between concepts. Manually created and maintained ontological approaches to improving information management have been hampered, in the past, by scalability concerns. The Applied Semantics Ontology employs artificial intelligence algorithms that enable sophisticated, automated self-learning, maintenance, and growth. This self-learning ontology can grow organically, ensuring that concepts and relationships remain up to date as language usage changes over time.

The Applied Semantics Ontology can improve knowledge discovery processes because it works in conjunction with sophisticated interpretative modules. These modules leverage the millions of explicit connections between concepts contained in the Ontology to process the information that businesses need to be able to locate on demand. The primary module enables the processing of documents by analyzing each word and phrase in a document, breaking the phrases down into component parts (or tokens), matching the tokens to the terms contained in the ontology, determining which of several potential meanings each word or phrase represents, and determining the most representative overall themes expressed in the document.

Once this primary processing is completed, the Applied Semantics application modules use this conceptual information to create the resulting structural pieces, such as metadata, summaries, or suggested categories, that enhance the document. The user configuration module allows an enterprise to establish the rules that determine how information is processed and results are returned. A final module allows for the Applied Semantics Ontology to automatically determine the existence of concepts not presently in the ontology and to make a contextually appropriate determination of the meaning and placement of the term, subject to human oversight and approval. Our Ontology combined with the interpretative modules that allow the technology to process information is what we call CIRCA, our Conceptual Information Retrieval and Communication Architecture. With CIRCA Technology, we have created a communication platform that is scalable, language sensitive, intelligent, and highly accurate in making information locatable.

F. Applying Conceptual Information to Solve Business Problems

Applied Semantics has built its CIRCA Technology to bring the power of automatically understanding the conceptual information behind strings of text to improving mission-critical information management. Applied Semantics has created three applications that utilize CIRCA Technology to return the information that will best improve existing knowledge discovery systems by making information easily locatable. These applications are:

- **Metadata Creator:** An automated middleware application that determines the most important meanings on a page, suggests the metadata that express those meanings according to user-specified preferences, to improve the ability of search technologies to locate and return relevant material to a business user.
- **Auto-Categorizer:** A high-performance, automated application that understands the meanings of documents and assigns them to a user-defined taxonomy, with no need for a training period.
- **Page Summarizer:** A customizable application that automatically processes and determines the gist (or themes) of a document in order to extract the sentences most representative of the meaning of the entire document, speeding the ability of a user to understand what a document is about.

These applications, separately or in combination, can improve the ability of search technologies to find mission-critical information, provide the summaries that prevent a business user from having to open and read through irrelevant material, and rapidly and accurately categorize the information that will prove vital to creating enterprise-wide competitive advantage.

G. The Benefits of CIRCA Technology

Applied Semantics' CIRCA Technology improves knowledge discovery initiatives and provides the following benefits:

- **Language independence:** The language-independence of *meanings* has made it possible to support multiple languages; but more importantly, an ontology acts as a framework in which data in different languages can be linked together and centralized.
- **Scalability:** Our ontology itself scales through automated means and any of our applications can be deployed in a distributed environment and support extremely heavy volumes of usage.
- **Automation:** Applications of our core technology truly automate every aspect of metatagging, categorization, and summarization, particularly in the crucial deployment phase.
- **No training required:** Our technology already understands what it needs to know in order to serve the needs of your enterprise.
- **Speed of performance:** Applied Semantics has developed a proprietary database structure that allows us to search the ontology and return requests in a fraction of a second.
- **Ease of integration:** Any application or infrastructure standard that can send or receive XML can utilize these applications.

H. Conclusion

The key to maximizing the value of the information that computers are collecting is equipping computers with the ability to process ideas with the same ease that they process numbers. Recognition of the conceptual information in text strings gives computers an intelligence similar to what humans use when analyzing information—the ability to draw on information that is related, but not identical, to whatever is being processed.

An ontology and its related interpretation modules move beyond analyzing the surface text and use conceptual information to improve mission-critical knowledge systems through metadata creation, accurate categorization, and automated summarization. Equipping computers to process information by understanding how data relate to other data is an important first step down the path of making available information easily locatable to the enterprises that depend upon up to date, relevant information to remain competitive in challenging business environments.

About Applied Semantics

Applied Semantics (formerly known as Oingo) develops innovative software solutions that enable businesses to better organize, manage and retrieve digital information in Web-enabled, enterprise and e-commerce environments. Applied Semantics' solutions are based on the company's award-winning CIRCA Technology, which understands, organizes, and extracts knowledge from unstructured content in a way that mimics human thought and language, allowing for more effective information retrieval. Founded in 1998, Applied Semantics provides customized, stand-alone software applications for businesses, as well as tools and middleware solutions that can be integrated into existing systems.



10474 Santa Monica Blvd. Suite 200
Los Angeles, CA 90025
PH: 310.446.8162 x253
FX: 310.446.8172
www.appliedsemantics.com