GILBANE GROUP
A DIVISION OF OUTSELL, INC.

Semantic Software Technologies: Landscape of High Value Applications for the Enterprise

July 27th 2010

by Lynda Moulton,
Senior Analyst and Consultant

Outsell's Gilbane Group: Multi-client Study

Gilbane Group gratefully acknowledges the support of the sponsors of the research informing this report. This work would not have been possible without them. Please see the *Vendor Deep Dives* section of the report for detailed descriptions of these search suppliers and their offers.

COGNITION GIVING TECHNOLOGIES NEW MEANING™

EXPERT SYSTEM
SEMANTIC INTELLIGENCE

Linguamatics

Gilbane Group Inc. a division of Outsell, Inc.
763 Massachusetts Avenue Cambridge, MA 02139 USA
Tel: 617.497.9443 Fax: 617.497.5256
info@gilbane.com http://gilbane.com

Table of Contents

Page #

Table & Figure Titles

Page #

Introduction to Semantic Software Technologies

The phrase *semantic search* has entered the lexicon of the workplace with good reason. We are all searching for meaning in the flood of information that surrounds us inside our work space and from outside media sources. We need better tools for filtering, organizing, and efficiently focusing in on specific content resources that are required and wholly sufficient for doing our jobs well. "Cutting to the chase" is probably an apt expression for how we want to "sort the wheat from the chaff" of the information overloading our spaces.

In the previous paragraph two hackneyed expressions (or idioms) are used to illustrate the issue of semantics. Because they are over-used and well known to most readers, we have a good sense of their implied meaning in any context. The human brain understands how to recognize and interpret a metaphor, distinguishing it from its literal word-meanings. This is a problem with which computational search must grapple. The complexities of language and conceptual understanding by computers are at the core of our subject.

Note that the deliberately chosen title of this study refers to *semantic software technologies* not *semantic search technologies*. We want to be clear that the end point of any semantic software is to improve finding and interpreting content, a search activity. But there is so much more computational processing that is possible to improve search than simply creating an index of text for searching. It is these surrounding software tools in the context of better retrieval that are the focus of this report.

Over thirty companies that embody semantic technologies are routinely featured in surveys of the enterprise search landscape. But dozens more contribute semantic software solutions in the broader information marketplace and they are largely unknown to the average knowledge worker. Researchers and subject and functional specialists, who need to improve their own ingestion and digestion of vast quantities of information, are the experts who seek and use semantic solutions. Because these semantic tools are not familiar to IT and business managers, they are underutilized where opportunities for major enterprise semantic search improvements could be made. In this study, offerings that are complementary to search will be examined and highlighted for their genuine business benefits.

History and Context

As early as the 1980s significant research appeared in information science literature about the development of expert systems for improving search results. Hundreds of universities, start-up companies, and major corporations have published research and filed patents on various algorithmic techniques for machine-aided searching over three decades (and earlier when much of this work was classified as artificial intelligence). By the late 1990s and early 2000s, these technologies began to be described as semantic search components. In 2001 Tim Berners-Lee published an article in *Scientific American* proposing a semantic web evolving out of the expanding worldwide web.

Quite simply, the vision of semantic search is the availability of software algorithms that would improve retrieval for the average person by interpreting their native inquiry and returning semantically relevant results. The idea is that something as mundane as typing, "where can I find a gas station in Bolton, Mass?" could be answered as accurately by a search engine as by a human being. On the internet, this would be a "semantic web" query. As web search engines continue to improve, good results to such a query have become a reality. This type of Q & A often makes use of a semantic technology called "natural language processing," one of many related technologies that comprise the semantic software technology landscape.

However, in the enterprise, expectations for relevant search results are much higher than for finding content already optimized for e-commerce on the web. Each business unit in an organization has specialized requirements for finding information needed to do its work more efficiently. This is where other types of semantic processing can give organizations a competitive edge by getting workers to answers more quickly, with more conceptual relevance, and even with pinpoint accuracy. The idea is to get only the right information (only relevant) and all the right information (everything that is relevant).

In the enterprise, semantic content technologies and "intelligent indexing" improve retrieval in many vertical domains and for numerous functions. This covers a spectrum from finding a single critical engineering drawing for completing manufacturing plans, to discovering an e-mail thread that might absolve a client in litigation, retrieving all available research in the past year on a particular gene biomarker, or collecting all invoices submitted to a delinquent customer. The semantic underpinnings that enable each type of search are the technologies this study will describe, define, and illustrate by providing examples of how and where they are now deployed.

Finally, this study is for the business reader to gain a high level of understanding of the scope and depth of software technologies that comprise the semantic tool landscape. Experts and specialists in the field may find it useful for explaining what it is that they need to procure, why it will improve their work results, and how that can impact the "bottom line." The focus is better business outcomes for the enterprise by identifying, selecting, and implementing the appropriate software for a particular work function in an enterprise. It is not exhaustive but representative of the range of semantic software technologies in use today.

Using This Report

This report is structured for readers with a business function that might benefit from semantic tools *and* those on the front lines for selecting, implementing, or depending on (as direct or indirect users) semantic software. Three major themes govern the body of the report: definitions and background, application information aligning players in the current software landscape with use cases, and guidance for buyers and sellers of semantic software.

It is probably most appropriate for anyone new to the subject to digest the information up to the *Applying Semantics to Business Challenges* section before moving on. The information in the first third of the document is to prepare a reader for a new type of business software and illuminate what it is.

As the decision is made to plan for adopting semantic tools, the second two-thirds of the report sets the stage for specific application usage, and describes what products are in the market today. Then it provides information needed to actually select and implement a product for a solution.

Market Landscape

On January 28, 2010, a member of the LinkedIn Enterprise Search Engine Professionals Group began a discussion with this question:

> *Semantic search technology – does it actually exist?*
> *We've all seen promises that semantic search will be the next big thing. However I'd love to know if it actually exists in a workable form for the enterprise, or whether it's still just a marketing department's dream. Comments? Examples? Thoughts?*
> Note: Only LinkedIn members of the Enterprise Search Engine Professionals Group will be able to see this discussion thread.

The discussion thread was active for several months, although most of the comments revealed very narrow ideas of what the total landscape includes. Although the question was about "search," the broader focus belongs on multiple semantic technologies that *impact* search, beginning with content enhancement and modification software that improves the chances for successful, meaning-based content retrieval.

IT and business leaders generally lack awareness about how widespread the use of semantic tools really is. Their knowledge of semantic technologies is sketchy and usually confused with the semantic web, an evolving kaleidoscope of search engine techniques to make internet search results more relevant.

Technology landscapes are interesting and, like environmental settings, ever changing in an evolutionary mode. It is rare to have transformative, revolutionary changes appear. The landscape of semantic technologies is particularly interesting for two reasons.

First, it was the internet that spawned a collective awareness of "semantic technologies" not too long after the internet accelerated online retrieval of content by everyone who had access to "the Net." Tim Berners-Lee, who initiated talk of the semantic web, may have created an expectation that it was imminent, just around the corner. But it wasn't, in spite of the visionary who stimulated the discussion. So, instead of something revolutionary like the internet, we have something evolutionary and rooted in technologies that preceded the internet by decades.

Second, the technologies whose landscape we are mapping are based on language and meaning. The study of these domains is ages old, dating to a time when technologies of those eras would hardly have been connected to the study of meaning and linguistic expression.

In this context, readers should note that the very problems semantic technologies address are linguistic human communication. Problems of a complex semantic nature are not easily solved, requiring thoughtful approaches to identify, select, and apply appropriate technologies. There is extreme variability of human expression; all the forms, nuances, and surrounding context that accompany any piece of content (written, illustrated, and spoken) require extreme computational modeling to achieve accurate results.

Couple these linguistic complexities with the pace and amount of content humans are required to ingest and digest in their work and private lives, and we have pressures that demand technological solutions. In enterprises we also encounter impatience by management with any requirement for humans to be part of the solutions. It is true that humans are way too slow to conceptually organize the full domain of information that flows into the work environment, but they are essential to building, evaluating, implementing, deploying, and using technology solutions. In the *Team Composition* discussion in the *Guidance for the Team* section later in this report, we describe the need for and competencies of necessary human resources.

What Are the Semantic Software Technologies?

Consortia, standards bodies, and notable software technology conferences informed this study. While many companies tack the word "semantic" onto their offerings, there is enough evidence that only about eight software technology classes are truly based on semantic and linguistic processing. We also looked for evidence that a class of products is accepted in the software industry, albeit often self-defined, as having some type of semantic analysis processing. This study is focused on the following broad categories, each of which embodies both computational processing and linguistic interpretation of the meaning of content:

- *Text mining and text analytics* – Processing that gathers text from unstructured files or structured applications (e.g., databases, CAD systems, and e-mail systems) and manipulates it to reveal or create new models of the information contained in the text.
- *Concept and entity extraction* – Processing that analyzes (usually unstructured) mined text for conceptual topics and distinct noun entities (e.g., names, organizations, places, and phone numbers) and orders them for further application.
- *Concept analysis* – Processing extracted concepts for relationships to other concepts within a particular context to solidify the precise meaning within the originating content.
- *Natural language processing (NLP)* – Automated application of the results of concept analysis to determine the meaning of human articulated assertions or queries using computational linguistics.
- *Content data normalizing* – Processing semi-structured content to a standard form of expression, format, or structure.
- *Federating and de-duplicating* – Technically a process applied to content from multiple sources that has been indexed for retrieval to present each item in a uniform format, eliminating completely identical results but often deployed to reveal and organize similar results in a more easily understand framework for simpler evaluation and analysis.
- *Sentiment analysis* – Processing that applies rules of linguistics and grammar to detect the tone of content on a pre-determined scale to graphically express the judgments or tone of a particular population.
- *Auto-categorization* – Processing that applies concept analysis and pre-defined vocabularies to a specific corpus of content for the purpose of organizing the content by topics and/or entities.

The reason for the order of this list might be debated but the first four categories are intended to reflect some processing that needs to happen with content before user-facing applications can emerge from them. They are core processing functions without which the other processing activities would not be considered semantically enabled.

In conjunction with most of these computational technologies are dozens of complementary technologies that support one or more of the above. As examples:

- Turning Word documents, PDFs, or scanned files into ASCII text using OCR (optical character recognition);
- "Scraping" web pages to support text mining;
- Harvesting database content and applying XML tags;
- Applications with embedded tools for building up ontologies (or taxonomies and thesauri) to be applied to concept analysis and automated tagging.

The study focuses on the eight categories in the previous list because they are the most widely available in commercial applications and adoption has become widespread. These categories of technologies are highly interdependent. Most commercial semantic software application offerings make use of several of these processing technologies.

Product Underpinnings: Linguistics and Computation

Semantic software is rooted in processing content for its meaning. Algorithmically eliciting contextual meanings from content produces additional complexities for processing it in order for it to work in advanced applications.

Semantic technologies are distinct from traditional software applications that process and index content strictly as data. From native data sources, computerized indexes are built directly using the strings of characters, as they exist in the original source, together with associated identifiers that link index entries back to the source. Search from native indexed content is relatively simple, requiring only that strings being searched are matched to strings indexed. Over time, search engines that employ only string searching have been paired with layers of additional processing; this processing applies rules to make better guesses or broaden the scope of what should be considered a qualifying retrieval result.

One example of processing layered on string searching is "stemming," a technique that returns, as relevant, documents with not only the string being searched but also documents containing other forms of the word string (e.g., a search for "floor" might also return documents containing "floors," "flooring," and "floored"). The chances are that content with "floored" is not semantically relevant to the search. To overcome this relevancy disconnect, semantic software is designed and developed to make better semantic selections based on very rich linguistic analysis that takes on the whole body of a language and contextual relationships to get at the true intent of the query, returning only relevant results.

The task of defining semantic software technologies is risky because this is a new field in which no one technology method or model has reached a position of clear market leadership. Expert developers and theoreticians abound but it is difficult to find consensus around the computational algorithmic approaches that are "best."

We position the issue a little differently, from the point of view of those readers trying to solve a problem for which semantic technologies might offer a solution. Such a problem, once defined, might have many technological solutions, and computational approaches. Without mathematical modeling expertise, or a linguistics degree, it is pointless to agonize over the mechanisms under the covers. Instead the reader needs to know that:

- Teasing concepts from any unstructured documents is among the most difficult of computational problems.
- There are relatively few experts in the field who have reached "rock star" status yet.
- There are no companies or products in this field that have eclipsed all others, as offering universal semantic processing or semantic search, yet.
- Success with any semantic software option will depend on how it is implemented, supported by a team of people who understand the tool, and how much computing resource is devoted to the software.
- If you begin now, you will be in an early adopter category.
- Better results may well be achieved by using a number of software tools that complement each other, either in an integrated fashion or in tandem.

In the appendix to this study is a glossary of terminology used throughout the document, providing a synthesized definition of each term as it applies to the fields of computational linguistics and semantic processing. These are the two disciplines that repeatedly surface as germane to every product we have included in the vendor directory, also in the appendix. In addition, articles with more background, organized by some of the more prominent topics mentioned in this study, are in a bibliography. The intent is to nourish greater understanding of the difficult business and technical issues that make up the landscape.

Computational linguistics is a topic that requires an expert's explanation. We discovered an excellent two-page summary from Hans Uszkoreit of Germany, who has provided an English language version of *What is Computational Linguistics?* , which gives boundaries to the subject matter in our report.

The second topic, *semantic processing*, is best summed up as moving from merely keyword indexing through steps that begin with "morphological analysis," then parsing content, executing sentence analysis (the logic of a sentence) to arrive at semantic analysis. Done well, this requires advanced expertise in understanding word forms, parts of speech, how words relate to each other, and contextual relationships. The expertise for this field is linguistics, which when combined with computer science makes up the foundation of computational linguistics.

This background on the technology underpinnings, while interesting and intended to help understand the difficulty of "doing semantics" well, does not supply any judgment on what is the best solution for a particular problem. As with any technology we strongly recommend running realistic tests and proofs-of-concept (POCs). More recommendations on evaluation and selection are in the *Guidance for the Team* section.

Linguistic Challenges

The named semantic applications described previously share several challenges posed by the way human beings express themselves, whether speaking or writing. Some brief descriptions serve to illustrate requirements for computational algorithms to make sense of what humans write or say.

Concept Discovery

This is complicated by poor expression, lack of context, muddled grammar, or confusing terminology. When presented with this excerpt of a press release:

> *In his opening address, Henning Nielsen (Novo Nordisk), President of the P-D-R-, commented on the continuing consolidation within the information industry, with primary as well as secondary publishers involved in....*

A dictionary look-up to discover what the "P-D-R-" is or performing a text string search on the web probably will not easily provide an answer. A Google search produced, as top results, references to the *Physicians' Desk Reference*.

A human, seeing the name of a pharmaceutical company (Novo Nordisk), and the name of a person in leadership might try adding some known context to the search (e.g., "pharmaceutical" or "Hennig Nielsen"). The search results would immediately top out with the organization name: Pharma Documentation Ring, which is what the acronym stands for. The mental process that engages an individual who is trying to discover the meaning or significance of any content, when simulated in automated semantic processing, is highly complex. The next four computational linguistic processes are among those that help deliver discovery of uncommon ideas or little documented names and facts.

Meaning Understanding

Consider the statement,

> *Beaver is attempting to correct some cutting defects.*

This might require clarification to the listener. The context is the production of surgical instruments developed by Beaver Surgical Instruments, a company, but its founder's name was Beaver, and then it was owned by his son. Humans, with this context, would ask, "Is someone in the company doing the trouble-shooting or is it being done by Mr. Beaver himself or the son?" Automated semantic systems that are designed to interpret queries (natural language processing) must include methods for resolving just such ambiguities, usually by prompting a human engaged in the process for the correct interpretation.

Entity Extraction

This is algorithmic processing to detect information that can be clearly defined as an explicit distinct string from the content in which it is found. Examples would be names of people, places, products, organizations, or explicit dates. Using existing dictionaries, glossaries, tables, or ontologies, entity extraction software successfully finds and re-purposes entities for metadata. Like other semantic software it makes use of additional linguistic rules and relationships to disambiguate identical entities names.

Context

When computer software does semantic linguistic processing for the purpose of precise or complete retrieval, it leverages rules provided by supporting tables or ontologies of terms and term relationships in concert with surrounding information from the content being searched. In our example in which there was a reference to "P-D-R," surrounding context was needed to make a connection to the right entity. In simple search examples, distance between words or phrases is used to provide measures of relevance. In semantic searching, these simple measures are made richer by more complex rules of grammar and semantic nets containing all known word meanings mapped to all known relationships with other words. Leveraging semantic nets plus rules applied to use all the surrounding contextual information is what differentiates rule or statistically enhanced string searching from semantic search. How well context is "understood" by the technology is the semantic qualifier.

Term Disambiguation

Figure 3 illustrates one of thousands of words that have multiple meanings that are completely different. When rules apply meaning to content and then perform a function like stemming, weird results appear for searches if disambiguation of a term contained in the query is not performed. An example would be a search for the name of a person whose last name is "Rising" and having search results appear that are about "roses," the flowers. Apparently, the search engine applied stemming rules to "rising," and interpreted it as a form of the verb "to rise" then looked for all the variations on that verb and came up with "rose." What followed was such an illogical consequence of badly applied stemming rules that it was silly. Algorithms that detect the potential for ambiguity must resolve the terminology algorithmically using context or other means, or prompt the searcher for his own disambiguation option.

Sentiment and Tone Analysis

Among the earliest and most widespread commercial applications of semantic text analysis was the measurement of the tone of language in a context. Again, this is linguistically difficult to automate because it relies on a variety of techniques. To detect whether essays on a concert performance, political conduct, or speech are conveying a positive or negative message, the net tone or sentiment has to be judged on the total context to be accurate or meaningful. Typically, content of a judgmental nature is a mixture of positive and negative and the results, whether evaluated by a human being or software, will reflect the balance on a scale of tonality.

Recently, there has been widespread adoption of this semantic technology to provide analysis of social media commentary, aggregating across blogs or Twitter. We note that the shorter each piece, the more likely it is heavily balanced in one direction or another, therefore, more easily rated on a sentiment scale.

Perspective and Authority

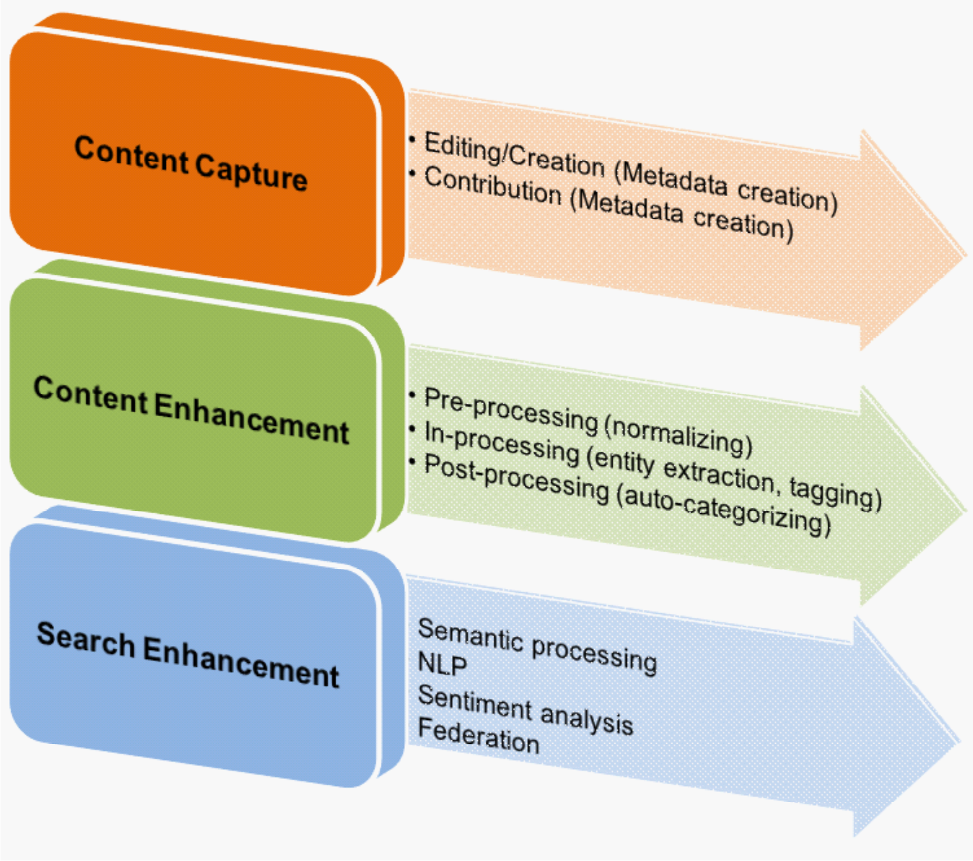
Using automation to ascertain the authority or perspective of the content contributor or author is probably the most complex challenge. Without the context of other sources, outside any one document, applying rules and ontologies would require a level of sophistication that includes factual assessment. With the context of access to vast information resources for checking authoritativeness or testing for perspective of a source, it is fair to say that the computational resources needed to run these evaluations strictly through automation makes this type of semantic processing impractical to apply routinely. However, it would be desirable to have semantic tools assessing content for perspective, and to be able to gauge its authoritativeness. This could be put to use in any number of business and governmental intelligence applications.

In the next sections we'll look at how semantic software technologies are currently in use: for generalized business challenges, in commercial product segments, as well as technology uses in vertical markets and technology uses in functional (horizontal) markets.

Semantics in the Life Cycle of Information

With these simple definitions established, the landscape of semantic technologies can be viewed from several perspectives. A simple high-level view would include where they might be positioned in the life cycle of content.

Figure 1. Semantics Applied in the Life Cycle of Content


Content Capture
• Editing/Creation (Metadata creation)
• Contribution (Metadata creation)
Content Enhancement
• Pre-processing (normalizing)
• In-processing (entity extraction, tagging)
• Post-processing (auto-categorizing)
Search Enhancement
Semantic processing
NLP
Sentiment analysis
Federation

Source: Outsell, Inc.
© 2010 Outsell, Inc. Reproduction strictly prohibited.

Content Capture

Metadata is automatically part of all electronic content, whether it is explicitly contributed by the person creating a document or not. When sending an e-mail, basic information describing the message as document type "e-mail," sender, and recipient are automatically established, usually supplemented by the subject line. Most desktops are set up with default information that is attributed to documents that are created on them: the username of the desktop owner, a date created and last modified, etc. Content authoring systems may be configured to automatically assign default metadata based on access control group settings or sign-on information. Document management and content management systems can be configured to require basic metadata (e.g., title and category) when a document is created and saved, or contributed. All of these pieces of automatic or assigned descriptors are metadata. In a library system it is called "bibliographic data," and in a database it is the content in fields that make up a table.

Ideally, good metadata would be created by the author of every e-mail, memo, report, CAD drawing, document, policy statement, marketing piece, and so on. We know this happens rarely, if ever, in current business. The days of having administrative support for producing written communication are long gone, and with that any attempt to enforce a standard of document preparation on employees. Anecdotal evidence supports the view that even when writers do tag their own content, the results are poor and inconsistent. It is not a model to ensure reliable metadata and excellent retrieval based on that metadata.

However, there are process improvements that can be automated to contribute quality metadata based on semantic technologies. For each stage of processing content, consideration must be made for differences in functional areas of an organization. Different groups produce and use different types of content. Because the ways that groups use content types are highly variable, enterprises need to accept building different models and processes for each, increasing the probability that different software will be required for different purposes.

Semantic technologies that can contribute to improved topical metadata during the document creation and capture process include those that:

- Detect concepts within the content (its "aboutness");
- Detect entities (authors, organization, or groups) and entity attributes, or relationships to other content;
- Automatically categorize.

Applying technologies for text mining, entity extraction, and concept analysis requires interfaces that engage a document author directly or operate behind the scenes to tag newly created or contributed content. Interfaces for human *curation* to complement automated processing can help establish semantically correct and contextually relevant metadata.

Organizations may seek packaged (out-of-the box) products for performing entity extraction, concept analysis, and auto-categorization but the evidence is that baking these processes into their existing document management or content management systems using component technologies works best. Legal departments and law firms, professional services firms, and some manufacturing operations that have extraordinary documentation requirements are the industries that have realized cost benefits by building solutions that address very specific requirements. Metatomix is a company that supported an aircraft manufacturer in building an ontology of parts, processes, and design elements for standardizing language and building common understanding across all systems for a single aircraft manufacturing operation.

This article by Amit Sheth is a useful summary of metadata considerations for thinking about the nature of enterprise metadata in general terms. The following paragraphs present areas for applying the suggestions in Sheth's article. They suggest where semantic technology is being embedded to assist metadata creation and control, or could be through customized programming:

- *Platforms for content production* – In publishing or documentation production, coordination among the components of large complex content entities is critical. This means establishing linkages among existing data repositories, glossaries, and controlled vocabularies that need to be coordinated across the entire editorial process. A platform that can manage all content elements is an area where semantic components can be applied. One such component would be to detect and extract entities and concepts in the content for the purpose of building and applying controlled vocabulary with synonym equivalents. Process integration would deliver metadata to defined fields associated with documents from the controlled vocabulary lists, as new content is created or delivered to the DMS, an operation supported by Infolution. Mondeca specializes in building platforms in collaboration with Temis, a partner that contributes text mining and categorization software.
- *Vocabulary framework and maintenance* – Establishing a simple taxonomy for navigating an intranet is often where enterprises begin to build controlled vocabularies. Initially, this may be a manual process but as content and usage increases, the need for automated tools becomes apparent. When a high level or "top term" approach is already in play, various text mining, entity extraction and analytics tools can contribute to building up vocabulary structures for metadata application. These then must be integrated with content production or content management system workflow. A major consideration without many good solutions is the use of synonyms; factoring them into any controlled vocabulary is essential to ultimately improving a retrieval system. Keyword searches that position a searcher into a navigation scheme should be able to detect non-approved synonyms and get the user to the approved terminology in the taxonomy. Semaphore is an example of a company that understands building taxonomies, ontologies, and thesauri with the stated purpose of establishing a semantic model for search and navigation.

  "Synonymy" is an aspect of semantics that can improve even the simplest taxonomy management. True deep semantic analysis of content operates at a level of granularity that is from the bottom-up in a vocabulary structure. This is as opposed to going from the broad (top) down through a topical taxonomic structure, which is common use in a human applied tagging system. To move from simple top-down methodologies for simple semantic metadata selection to a bottom-up semantic complex analysis approach will generally require a shift in technology software type. The simple approach, in which human curation is involved, can only scale to contain domains of a certain size before it becomes too expensive and impractical to maintain. Cognition has devoted a couple of decades to building up an ontology of the English language with synonyms and relationships to recognize them in all their contextual relationships.

- *Publishing repositories and governance* – One of the failure points of content production and content distribution is lack of policies and procedures that are easy to understand and that fit smoothly with enterprise work flow. Like most technology implementations of high value, this takes some special retrofitting of existing software to make an impact. Custom programming is best done in a test mode to get the kinks out but prototypes must be designed to be scalable.

  Any content type, from small files and e-mails to major documents, can be pushed to a permanent repository that is "baked into" a governed enterprise framework. The idea is to make a pre-emptive strike at the proliferation of redundant files and lost knowledge assets, to control categorization at the time of document production.

  Without explicit directives from top management and the development resources to get it done, governance is a hard sell. However, it is ideal to have a commit process when a document is "saved," e-mail "sent," or "received and moved" to a desktop location. An archived copy would be placed into a location governed by a single classification (e.g., author, project, department) complete with required metadata. System profiling would determine what metadata is required and prompt for missing data before the transaction can be completed.

  Much of the metadata can be assigned automatically based on parameters established for everyone's content being entered into the system. Some metadata might be established when a new piece of content is started and some contributed at the end. While most off-the-shelf systems don't support automatic topical tagging based on an enterprise defined thesaurus or taxonomy, consider that possibility as a future desirable. In the publishing and media industries, Nstein, an OpenText company, has established a reputation for excellent content support for metadata governance. Mentioned earlier, Temis is another company that has a good reputation for managing vocabularies associated with metadata excellence.

Content Enhancement

When metadata creation has not been or cannot be established during document creation, content semantic enrichment is a good alternative solution, after creation, but before it is pushed out to be indexed by a search engine. This is probably the most reliable approach for enterprises that want high quality retrieval for internal content that is unique and valuable for its employees, for customer service operations or client self-service, and for marketing related content.

Enhancing metadata with semantic software technologies before publication to an internal or highly valued customer/partner/prospect audience will ensure higher satisfaction with search results. It will make findability closer in reliability to an e-commerce site in which pinpointing all the products that meet specifications is easier. To create a similar search experience behind the firewall, getting metadata in excellent shape is essential.

There are basically three types of semantic processing that will bring the most immediate benefit to improving retrievability of existing content:

- *Normalization* – Pre-processing documents, which have similar attributes but are labeled with different nomenclature, into a uniform mapping will be dramatically improved using underlying semantic processing. When applied to disparate repositories of similar documents, all containing roughly equivalent data types or having some data elements in common, semantic parsing and re-assembly is recommended. It will improve content exchange for business purposes in which further analysis is desirable. The semantic processing applied to this function includes parsing, concept and entity extraction, and transformations.

  Semantic normalization is being applied across systems that must work together using or sharing the same data. This would be institutions within law enforcement networks, health care systems, global organizations made up of numerous widely distributed units, military units, or other governmental agencies with overlapping responsibilities. Wherever vast domains of content are collected through forms or spreadsheets and that content needs to be analyzed quickly to achieve institutional or inter-institutional goals, new computational methods are needed. These semantic tools are built to understand the relationships among all the data sets and are the best option for gaining efficient rationalization of what exists. Once normalized in a common labeling network, data can be sliced and diced visually or in easily understood reports. A company that is tackling the coordination and consolidation of data from spreadsheets from different parts of an organization is Cambridge Semantics.

- *Entity extraction and concept extraction* – Used in document processing to discover the entities for metadata (e.g., names, dates, and projects) and concepts that exist in the content. When delivered using a semantic net or rules for extraction, transformers can contribute metadata where none existed before. The better a semantic net of language relevant to the domain, and the rules for relevant entity detection, the better the metadata that will be built.

  For enterprises with years of accumulation of random unstructured content, data mining and parsing content is the first step to finding entities and concepts. Once processed, the resulting entities and concept mappings are applied to documents as metadata for improving its findability. It is also a first step to auto-categorizing the content for any number of search models. Clarabridge and Attensity have out-of-the-box applications for these processes while Cognition offers a semantic richness that is better suited to very large scale projects of concept extraction.

- *Auto-categorization* – Once content has been supplied with semantically built or enriched metadata, auto-categorizing is a simple post-processing process, regardless of what stage in document creation or use that metadata was established. However, newer technologies are available that can ingest un-tagged, unstructured content and then perform auto-categorization after data mining and linguistic processing steps.

  Applications are in demand for processing millions of documents in the form of e-mails, scientific papers and articles, business memos, office documents and reports, and court documents that need to be ingested and processed rapidly for litigation. Obviously, missing any relevant material or conversely including many irrelevant documents is a burden for discovery. Narrowing the corpus to just the right documents in each category is a dramatic overhead savings. Recommind and Clearwell Systems are two companies focused on the legal market, each with very different approaches to content auto-categorization in that space. Another example that illustrates a hybrid solution is the use of Connotate to scrape and parse web content, embedded in an automated metadata creation and categorization solution that Cormine developed for WorldTech.

So, summarizing the applications of software to content that lacks sufficient metadata or context, we can use these types of semantic processing (some of it packaged as semantic middleware and other as components of larger solutions) to bridge the gap between content at its point of capture and content at the time when it will be searched. For both web and enterprise content, thinking about content intent, audience, scope, and depth of content preparation, and how it needs to be found are essential to determining sufficiency of metadata.

Search Enhancement

When cost, time, and resources are scarce for performing concept and entity analysis on content prior to indexing, it falls to semantic processing components bundled with search engines to perform these operations *in situ*. Indexing full-text directly by search engines typically results in indexes that support keyword searching. And, as already noted, various algorithmic processes may be layered on these engines to improve findability. The results are often pretty good, especially when pinpoint precision is not required. In cases where search is needed to gather background information, find well tagged documents, or products for purchase, generalized search engines will usually retrieve sufficient content. These engines also work sufficiently well as site search engines when the amount of content is relatively small, in a unique topical domain.

Web content is significantly more diverse and diffuse than enterprise content, which is denser and more complex but narrow in focus. It needs a different approach to semantic interpretation because the language tends to have unique vocabulary that is best understood by subject matter experts in the organization. Here are the differences in how we typically apply search to public domain content on the web and content that we find only behind the firewall:

Table 1. Comparison of Reasons to Search for Web Content and Enterprise Content

Web Content	Enterprise Content
Product searching: Product name, type, or application or purpose	Product searching: Product name, type, focus on *application or purpose + business impact*
Answering a question: Personal or professional topic of interest to gather data, gain understanding or background, learning	Understanding: How to operate in a work environment (e.g., policies, practices, tool support for our work environment)
Researching a topic: Personal or professional topic of interest to gather data, gain deeper understanding or background, achieve learning	Discovering and understanding: Organization behaviours and expertise
Solving day-to-day living problems: Personal health, safety, travel, diet, household maintenance, vehicle, social	Recovering work results: Subdomain specific (e.g., R&D, manufacturing, quality control)
Solving day-to-day technology problems: Personal or professional electronic equipment, software, tools, etc.	Discovering and learning: The organization's products and services, marketing and customer related issues
	Accruing expertise: Proficiency in industry legal and regulatory, environmental issues, contracts, governmental agency compliance and requirements
	Solving problems: Test results, specifications, differentiating new issue from solved problem, analyze results

Note: There is a substantive amount of content accessible via the internet accessible only through controlled access paths, usually for a fee or through membership. In general, deep-web content uses align with enterprise content and professional purposes.

Source: Outsell, Inc.

©2010 Outsell, Inc. Reproduction strictly prohibited.

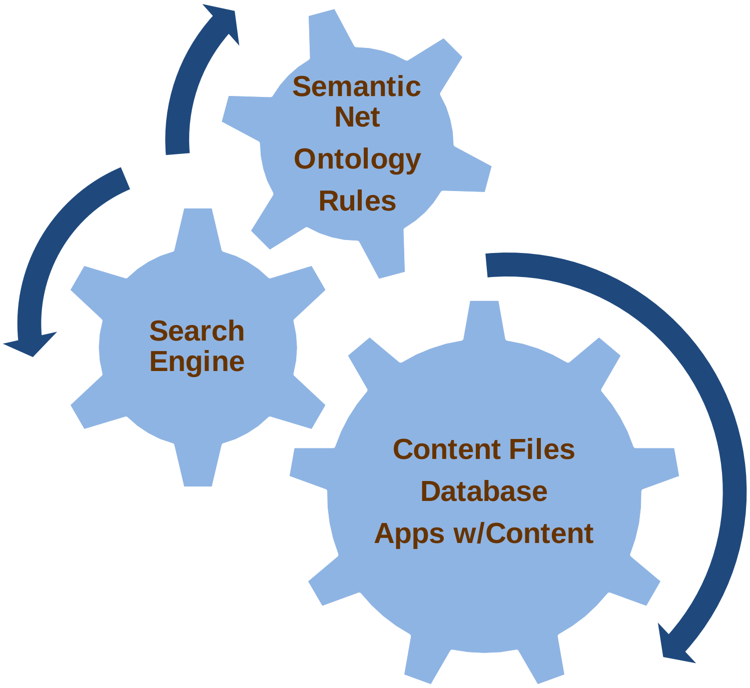
Semantic processing makes sense in any domain for improving retrieval when content is:

- Voluminous corpus (millions of documents);
- Complex in scope and depth;
- High-value to audiences seeking only small portions out of the entire corpus;
- Needed by experts for use in their areas of expertise;
- Otherwise undifferentiated for purposes of research or e-discovery interest;
- Likely to impact that bottom-line, directly or indirectly, when discovered.

These are the bodies of content whose value is fully realized when processed in such a way that they can be aggregated, federated, pinpointed, or analyzed to reveal concepts or meanings that otherwise would not have been recovered. Semantic technologies are enablers to bring these corpuses into focus in ways that human beings simply could not logistically process for lack of time.

Semantic software technologies can help bridge the gaps between and among enterprise repositories. Here is one very simplified model of how they contribute in the context of search and retrieval.

Figure 2. Simple Model of a Semantic Platform



- **Terminology with meanings are defined in one location**
- **Target data and content are elsewhere**
- **Search engine applies meanings to find appropriate content**

Source: Outsell, Inc.
© 2010 Outsell, Inc. Reproduction strictly prohibited.

In this model the meanings, rules, dictionaries, etc. are established and stored in one location, while data and content may be located anywhere. The search engine, when it crawls content repositories, associates the rules or meanings to individual pieces of content to make it findable in a more semantically relevant way. Then it indexes the content components with linkages back to the source documents.

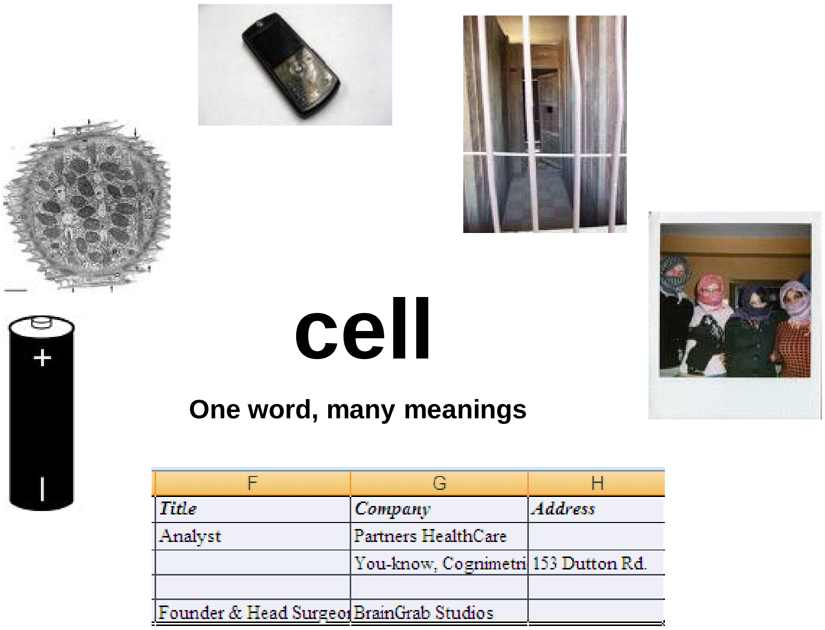
In more sophisticated environments, a wide spectrum of semantic middleware may have already been applied to content, during creation, post creation but pre-published, post publication, or any combination of these, to add topical metadata, categories and other attributes to improve indexing and findability. These were described in the *Semantics in the Life Cycle of Information* section.

Whether the model is simple or has more layers of software, the interlocking of content and its management with terminology maps (ontologies) or rules with the search engine indexing and retrieval process is the essence of semantic searching.

Human language technologies (HLTs) are the foundation of all semantic processing; the computational linguistics components are: morphological analysis, content parsing, sentence analysis, all in consultation with the rules and dictionaries to arrive at semantic analysis. Before natural language processing, sentiment analysis, and federation can be applied at retrieval time, these HLT processes, whether preprocessing or embedded, prepare the content for semantic findability.

Natural language processing (NLP) is applied when a fairly complex query with levels of granularity and linguistic qualifiers is posed. The semantic search engine will be able to understand the question, by applying linguistic understanding and then return semantically relevant results. For example, if I ask, *What are the probable causes of corrosion on the casing of this lithium cell?*, the semantic search engine will disambiguate the meaning of "cell" to understand that I am referring to an "electric battery," "electrochemical cell," or "electrochemical device" and discard from the results anything that relates to the other types of "cells."

Figure 3. Word Disambiguation, a Function of Natural Language Processing



F	G	H
Title	*Company*	*Address*
Analyst	Partners HealthCare	
	You-know, Cognimetri	153 Dutton Rd.
Founder & Head Surgeo	BrainGrab Studios	

Source: Outsell, Inc.
© 2010 Outsell, Inc. Reproduction strictly prohibited.

This is but one of several aspects of NLP, and it is done by assessing context and the relationships among the words in the query and the target corpus. Any truly semantic search engine makes use of natural language processing. Among semantic software technology companies with built-in NLP are those that directly market semantic search engines (Sinequa, ZyLAB, and Imbenta), those that are recognized for their semantic solutions in particular markets (Linguamatics, Concept Searching, and Expert System), and those with strong offerings as middleware or embedding with applications (Cognition, Basis Technology, and Smartlogic).

Sentiment analysis is another application of NLP that processes content for its tone, again by applying linguistic rules to understand a range of judgments about the focus of the inquiry. For example, someone might inquire about the reputation of Helene Curtis. Through a combination of entity extraction and auto-classification, which would have made a distinction between persons named "Helene Curtis" and the company, a search interface that is tuned for a sentiment analysis query would need to prompt the searcher for clarification, not knowing which option was intended.

Once this intervening operation is complete, the search engine would look for tonal content to answer the "reputation" question. Lexalytics, Expert System, and Attensity have all positioned themselves with reputations for performing sentiment analysis.

Federation, in its classic definition related to search, refers to federators mining the results returned in response to a query and presenting them in a context that is easily deciphered. In essence, federation has evolved to become an engine for unifying a single query targeting multiple disparate sources and a post-search semantic processing activity for interpreting results. It integrates the results returned from any number of search engines, normalizes the content, organizes logically identical citations or records into a unified structure, and provides contextual information about results that facilitates understanding for the searcher. MuseGlobal is a federation vendor with connectors to thousands of formats and applications; they are embedded in hundreds of search applications.

Federation is chosen for situations where searching will be executed across numerous very large repositories, all of which contain content resources (structured and unstructured) in many formats. The federator brings a library of connectors to the search operation to normalize the attributes of documents in each repository so that each is searched for similar attributes in a normalized query. For examples, a database of records might contain an attribute "description," while documents in a CMS have "titles," or in a records management system "document names." A federator will disambiguate document attributes based on the intelligence built into the connectors, de-duplicate, or conflate results, presenting them to the searcher in an assembly that is easy to interpret.

Applying Semantics to Business Challenges: Why Now?

Most of us begin searching the internet through a generalized search engine that indexes all the content free to the public on the web (e.g., Google, Bing, and Yahoo!) or using an aggregating, meta-search search engine (e.g., USA.gov) or a single website search engine (e.g., PicoSearch, SurfRay, and Endeca). Once there is a business reason to find content for its direct or indirect relevance, semantic software technologies become interesting for their impact.

Significant use of semantic technologies is increasing in two major domains: publishing and life sciences. This is not to say that there is not widespread use across other market spaces; those will be highlighted in the *More Applications and Comments* section. But first it is important to understand the business drivers for the publishing industry and life sciences.

Publishing, particularly non-fiction works that target specific industries, has always been at the forefront for indirectly leveraging search technologies. In the early 1970s, publishers benefited from two shifts in information access.

The first shift was the availability of online search engines that gave librarians the tools to research specialized literature previously only available through print index sources (e.g., *Chemical Abstracts*, *Index Medicus*, and *Psychological Abstracts*). This early automated search technology required a high level of expertise using command languages and best done by librarians with good working subject knowledge of the domains they searched. Furthermore, online indices retrieved only citations and abstracts pointing to print sources for the full text.

With the internet, expectations have arisen among the working population, knowledge workers, that they can easily and efficiently access content that previously required arcane skills, and much practice for successful execution. Web technologies with hyperlinking, coupled with a huge ramp up in the availability of electronic full text, was the game changer. From electronic indices novice researchers could instantly link to full text. This prompted the second major shift, the demand and technology for converting legacy content, previously only available in print. Volumes of electronic images of the older material have become accessible in recent years.

Users' expectations, as typically happens with new innovations, soon outstripped search aids, particularly in the absence of high quality metadata. In most cases, publishers were relying on metadata contributed by producers of print indices and those producers have been scholarly enterprises or government agencies. The amount of content being published in special fields plus the cost to manually tag it in reliable and controlled formats, reached a critical break point toward the end of the 1990s. The quality of human indexing declined as costs drove index publishers to scale back human subject specialist indexing staff.

In the 2000s, publishers began to understand the potential for revenue gains by selling their content in pieces. Previously, they made their money selling subscriptions and by gathering royalties on individual article reproduction, a demand driven by irregular or spotty discovery through online searching. It has been clear to many in the information industry that providing easy access with highly reliable retrieval mechanisms to researchers could drive more business toward individual documents of high interest. The alternative would be to increase the number of subscribers to buy a journal subscription, subscribers who would bet on finding something of interest in each issue. Rising publishing costs and decreasing library and professional budgets made the latter scenario unlikely. This was not lost on publishers, hence their increased interest in semantic tools to improve metadata production and enhancement, facilitate auto-categorization for website navigation by topic and faceted entities, and improve semantic search of full text with natural language querying. Semantic software technology use in publishing is growing, driven by the positive economics of publishers' e-commerce business, selling articles and individual documents.

Life sciences industries, including pharmaceutical, biotechnology, and healthcare management, are looking inward to leverage semantic software tools. For these companies, semantic tools provide a competitive edge to make discoveries and get to market faster, or to engage in efficiencies leading to cost containment.

There is also a significant relationship between publishers and life sciences. Among the most innovative publishers employing semantic software tools are those whose largest clients are in life sciences. Scientific and technical publications are as important to life sciences research as the bench scientists; the linkage between the literature and people doing research is essential to discovery.

Use of semantic software in life sciences shapes the way scientists and business analysts do their work in four ways:

- Parsing millions of published and unpublished document in order to find facts or data; applying NLP to expose answers to hypothetical questions;
- Mining across disparate and seemingly unrelated corpuses to uncover content relationships and to stimulate innovation;
- Exposing professional expertise by working from content to discover experts for collaboration or further investigation;
- Exploiting published content for competitive intelligence about industry and research trends.

All of these purposes might apply to any research-based vertical industry, but life sciences leads for good reasons. There is a vast domain of legacy content funded by government research that has reached the public domain (some free and some for a fee). There is also a vocabulary (ontology) with a base previously established in MeSH (Medical Subject Headings) from the National Institutes of Health. The life sciences ontology (UMLS) has continued to be built up through work of the government, and repurposed by academic institutions and entrepreneurs in the semantic marketplace. This vocabulary is a building block for some important semantic nets and continues to be a model used for natural language processing experimentation, product development, and testing.

With high value content and rich life science vocabularies available for experimentation to an industry with the deep pockets and incentive to use them, we have seen the opportunity for semantic software technology taken into the commercial realm. As in publishing, life science industries' use of semantic software targets impact on the bottom line. But rather than leveraging linguistic acumen to sell more products directly as publishers do, they use it to improve internal operations or to bring more products to market faster.

This brings us to a consideration that underscores a difference between web searching and searching in the enterprise.

On the web there is no consequential balance between seekers and providers. If providers of content, technologies, and applications do poorly in assigning quality linguistic enhancers to improve findability, or interpreting what the user seeks, they bear the burden.

If a content provider is trying to satisfy the searcher's need or interest, to make money or provide a service, or to gain recognition for itself or the company, and content is not found, the result is that the seller has lost more than the searcher in the transaction.

Conversely, in an enterprise, the reasons for needing enterprise content are work enhancing and enabling. Dependency is huge and interdependency is also a significant factor.

When enterprises operate in environments that do not support cross-organizational understanding, perspectives get skewed and the potential value of compartmentalized content is compromised. For example, mechanical engineers will put a project at risk when they are working on components of complex equipment and do not label parts in the same way as design engineers who will be specifying the use and assembly of the parts. This is because all the data, documentation, and specs for parts may not be found in time to account for critical points of failure when everything is assembled – stress and strength tolerances, size information, thread data, melting points, etc.

Within enterprises that are aware of these vital interdependency issues, employees share perspectives and think about ways of normalizing language, labeling, codifying, cross-referencing, etc. to make sure that everyone is using the right data for the right components needed for a common project. It is truly remarkable to see the failure of many organizations to assign human oversight to manage this terminology control issue. Just as there is recognition of the need for software code control in software companies, foresighted business planning calls for consistency in vocabulary and content practices. Managers need to be paying attention to the semantics of their domain and governance of content repositories.

We have already observed that semantic processing for a particular domain makes business sense when content is sizable (millions of documents), complex in scope and depth, high-value to narrowly focused audiences seeking only small portions out of the entire corpus, or needed by experts for use in their areas of expertise. We also know that obscure content with importance to niche audiences can be brought to light for a creative research by using semantic processing. We see that one or more of these content conditions exist for publishers and knowledge workers in the life sciences. For both there is willingness to invest in ensuring the right answers, the best information, and efficiency of processing (minimal overload). Other industries have those interests as well and are following the early adopters, as ontologies are built and content grows to proportions unmanageable by humans in any specific domain.

What Is Real and What Is Experimental?

Perception of the role of semantic technologies on the internet and behind the firewall is extremely fragmented. This is based on a review of eight years of discussions, presentations, and articles, and numerous more recent interviews of professionals in information technology fields. Those who have been using or developing semantic technologies may be somewhat insulated from market realities, because they are already adopters. The next two sections come from our research and describe what people told us, and then the evidence we found of actual case implementations of semantic technologies.

Research and Surveys: What People Say and Think

The focus of our report turns to products that have emerged over the past decade, finding customers, commercial viability, or recognition for technical quality. A few dozen entities will be called out for individual mention, and others are listed in the vendor directory. Research across numerous information industry listings uncovered dozens more that have recently emerged or appeared briefly and then were never mentioned again, or their websites showed no recent updates. There are undoubtedly hundreds of projects, programs, and start-ups seeking to find a niche in this very nascent marketplace. Our report is bounded by our knowledge of current commercial viability or technical uniqueness that has been recognized in the industry.

As already noted, the field of computational linguistics is very much at the heart of semantic software technologies. In previous decades artificial intelligence (AI) and neural network research contributed significant ideas to how computers could be employed to solve problems previously assigned to humans. We found that vendors of semantic software often refer to having AI and neural net technology expertise. While synergies and integration among these computational domains exist, we have tried to establish clarity around specific product applications in which some type of computational linguistic methods exist.

There were many individuals eager to speak about either their interest in semantic technologies, as eventual users, or just enthusiastic bystanders. Others are contemplating developing products, working for companies that are developing products or trying to assess what types of semantic technologies are on the market that would benefit their work. Many were willing to talk about something they are exploring, testing, or developing. From these discussions we learned that need is great, understanding is diffuse, aversion to risk is high, and skeptics abound.

It was more difficult to find a critical mass of individuals available to talk about use of any one commercial product. Some reasons are obvious, the first being the small number of customers who have enough experience to feel confident talking authoritatively about a very complex set of technologies. The few who are willing to be interviewed are in demand and speak at many conferences or write about their work. Second, many are in early stages of implementation and testing, and may be reluctant to make judgments about a product that has not yet proven its value. Third, in order to evaluate business impact or value, time and professional understanding are required; some may not feel comfortable sharing early impressions.

Over the past few years, we have interviewed a few dozen semantic technology adopters, those involved in the selection, implementation, ongoing administration, and tuning, or expert users. These are summaries of the more recent experiences and insights about this market and the technology challenges, some of which explain slow growth.

Table 2. Our Take on What People Are Saying About Semantic Technologies

The marketplace is trying to figure out what problems really can be solved by semantic software technologies, and return value in the enterprise relatively quickly.
Every technology seems to find champions depending on the use cases.
Enthusiasts who are on the sidelines are inclined to seek ways to solve semantic problems with open source software.
Adopters of semantic software have high expectations for very precise and relevant retrieval.
The projects that received the highest rating for semantic queries made heavy use of ontologies.
Those who have adopted the types of semantic tools described in the following sections (vertical and horizontal) have done so with significant constraints: human resources, infrastructure, IT support, and budgets.
Those who have adopted semantic technologies, even with constraints, are already certain of the return on investment.
Benefits cited as being realized most often were: • Professional time savings; • Improved precision in information retrieval with the indirect benefit of enhanced research, results and risk mitigation (by reducing time wasted on already proven bench science); • Significant improvements in market awareness (customer opinions) and competitive intelligence (signals and changes in the wind); • Improved quality of processes and products.
There is a lot of debate among semantic technology champions about statistical versus semantic net methodologies, but some consensus that using a hybrid approach or finding a place for each might be advisable, depending on the problem.
There is a lot of misunderstanding about the semantic web, what it really means or will impact, and not much thinking about semantic technologies as they might apply to the enterprise. Perhaps because semantics has been talked about in the context of "search" for so long, finding people outside of software development who can project what the benefits of semantic middleware might be for their organization was difficult.
Good tools for synonym expansion in the ontology or vocabulary rule base receive rave reviews and are in demand. Users recognize the importance of this when they see relevant search results that do not contain their query terminology in the retrieved text but synonyms are there instead.
Several users and experts who have deployed multiple search systems commented on the importance of being able to find out why they received the results they are seeing from a search. Without such a mechanism, it is very difficult to tune a solution to improve results for future searches.
Science researchers cautioned about expecting quick results and the need for constructing carefully thought-out queries, particularly with the NLP tools. They stressed the importance of having clear research strategies in mind for both their bench science plan and literature inquiry, taking a systematic and iterative approach as they learned the semantic platform, and taking time to study results and understand the logic of the output.
Among the positive comments for the newer semantic software technologies was surprise at the tremendous improvements in processing speeds. These comments were tempered a bit by notes that the amount of computing resources needed for indexing millions of documents could be significant and working with a vendor to understand compute platform requirements is needed.
Highlighting the precise context in a full-text retrieved document where the most relevant content is located is one of the most valued features of newer tools. Several observers noted that reading through full documents to discover relevant content is too time-consuming. Researchers want to be lead straight to the answers.
Correlating content from disparate information sources and discovering new concepts and relationships between seemingly unrelated documents is a major advance in information processing that is attributed to semantic technologies. Being able to bring all relevant documents related to a project, case, or product together from across enterprise domains of millions of documents has the largest business impact.
Pharmaceutical companies, being the most advanced users of the HLT tools, are aggressive about trying many products and committed to find those that return measurable results. Some focus on leveraging one or two tools for multiple semantic challenges, working with the vendors to fine tune for each solution. Others take a completely hybrid approach, seeking out the best-of-breed for each problem. The latter recognize the time commitment to become expert with every tool.
There is consensus that with out-of-the-box systems, there may not be as much flexibility for precise tuning. High value semantic problems cannot afford to struggle with "black-box" products.
Finally, several commentators made the observation that terminologies are hard to manage, hard to collaborate on and grow effectively. Smaller teams with more authority on controlling the growth seem to work better.

Source: Outsell, Inc.

©2010 Outsell, Inc. Reproduction strictly prohibited.

Scope of Market Demands and Market Realities

Of the semantic software types rooted in human language technologies (HLTs) there are 40 to 50 packages with significant market presence and enough out-of-the-box capabilities to support evaluation or testing and POCs in a matter of days to a couple of months. An equal number are in beta testing or contribute to semantic platform products and projects.

Market demand originates from within expert functional groups that recognize the potential for leveraging data and unstructured text to improve content findability and analysis. They are seeking ways to influence the bottom line by operating more efficiently, faster, with more accuracy and confidence in the content they uncover. *Time savings for valuable professionals are important, but adopters need to help management understand direct business impacts of semantic tools.* The most frequent impacts gathered from surveys and our research are noted:

- Discovery of new facts that redirect bench science in a more productive direction leading to opportunities for new product development, confirmation of methods and practices, and identification of experiments that have been executed with positive or negative outcomes that would impact the direction of bench science toward more productive methods.
- Discovery of business developments indicating competitive shifts, new opportunities for growth, and newly exposed areas of risk.
- Learning about and understanding market trends that present opportunities or improve enterprise market and customer focus.
- Rapid access to unique facts and experts that might go undetected or require major human research effort to uncover.
- Detection of patterns of irregularity in finance and banking indicating potential illegal activities.
- Ability to sift through and qualify unstructured content rapidly, eliminating truly irrelevant results, reducing human review time.
- Improving navigation and search through better metadata and categorization.

The reason for major business impact in these areas is that it has become impossible for humans to adequately survey the vast electronic information domains accessible to them in a reliable and timely manner. The quantity of structured and unstructured content, plus its lack of meaningful linguistic uniformity, has overwhelmed any human process previously used to extract value.

Because buyers are usually information seekers with expert requirements, they approach their search for tools through professional meetings, software industry events, and colleague referrals. In the semantic software market SEMTECH, now in its sixth year, has been the leader for attracting speakers and exhibitors to its events. The Text Analytics Conference and Search Engine Meeting each have a focus on niches in the semantic space. Exhibitors of semantic software technologies are present in smaller numbers at information technology industry meetings such as Enterprise Search Summit, AIIM, KMWorld, and LegalTech. Professional associations for law, scientists, engineers, business, finance, and librarians have increased the number of presentations related to search and some semantic technologies in recent years; their exhibits reflect products for these functional areas.

Educating professionals in any new and technically advanced market is a first stage to positioning and is where we see the most focus at meetings. As might be expected, many of the presentations are being done by software developers and the complexity of the topic does not lend itself to discussions at the level of a business audience, even when the meeting is attended by many non-technical professionals.

We view this as one reason for slow adoption; it is not uncommon for a business manager to learn about new software applications that "sound promising" at such a meeting. What follows is usually a report back to the enterprise, which in turn sends technical (IT) personnel to the next or a similar meeting to learn more. This sets up a disconnected process of having different professionals getting information independently; they should be seeking, evaluating, and planning for new technologies as a team. The *Guidance for the Team* section discusses selecting, evaluating, and implementing software in the semantic realm later, but it is important to note that exploring technologies in an immature market requires a disciplined approach for buyers because vendors themselves are trying to find their "sweet spot" in the marketplace.

In 2010 the industry is still evolving for early adopters; so too are standards and packaging, pricing, and support models are being tested and tweaked. Those who educate themselves and have clear vision about the business impact they seek will benefit from being in on the ground floor of a new industry. They will influence product development and establish closer (and valued) relationships with suppliers who need to have their customers succeed. The synergies are paying off in publishing, life sciences, and government (intelligence and defense). Finance, legal, and electronics are coming along to the marketplace, as well.

As already noted, we worked to find enough experts who actually had a business use that they could talk about from first-hand experience. It is not surprising that the most experienced developers and implementers have worked at a number of companies. Those that we did find are an enthusiastic and garrulous community and they communicate among themselves through social networking tools and at meetings. There was surprising uniformity around the challenges of implementing semantic software, even when particular techniques and standards are still being heartily debated. It requires great focus and attention to detail by smart people with a passion for understanding of complex problems. It requires discipline.

Because this study is focused on technologies that are state-of-the-art in the semantic space, the vendors we follow in earnest and highlight are those whose principal endeavor is this field. Major software companies are all in the game, too, but they are not featured in the next sections for a few reasons. Most of their offerings in the semantic space are embedded, either as an add-on to one of their product suites or as a new complementary package to be integrated with other products. They may or may not have developed the technology themselves; often the most innovative products on the market are from smaller companies. These firms are assigned to their strongest positioning in Figures 5 and 6, showing vertical and functional markets.

Of course, the larger software firms are always looking to innovators for opportunities to acquire software IP (intellectual property), bypassing a long development cycle themselves. We have tried to focus on the companies dedicated to only semantic software development. For this reason, we do not see Autonomy, Google, IBM, Microsoft, OpenText, Oracle, SAP, or Yahoo! on this list although most appear in the vendor directory in relevant segments because they have components, modules, or tools for solving semantic processing challenges that fit with the landscape.

The next two sections position the companies in various vertical and functional categories.

Table 3. Companies with a Multi-Year Market Presence in Semantic Software

ai-one	Collexis (Elsevier)	MuseGlobal
Ariadne	Collibra	Netbreeze
Attensity	Concept Searching	Nstein (OpenText)
Attivio	Connotate	Ontoprise
Basis Technology	Endeca	Ontos
Bitext	EntropySoft	Recommind
Brainware	Exalead (Dassault Systems)	RiverGlass
Cambridge Semantics	Expert System	Sandpiper
Cerebra Inc.	Inbenta	SAS (Teragram)
ChartSearch	ISYS	Semantra
Clarabridge	Lexalytics	Sinequa
ClearForest (Reuters)	Linguamatics	Smartlogic
Clearwell Systems	Metatomix	Temis
Cognition	Mondeca	ZyLAB

Source: Outsell, Inc.

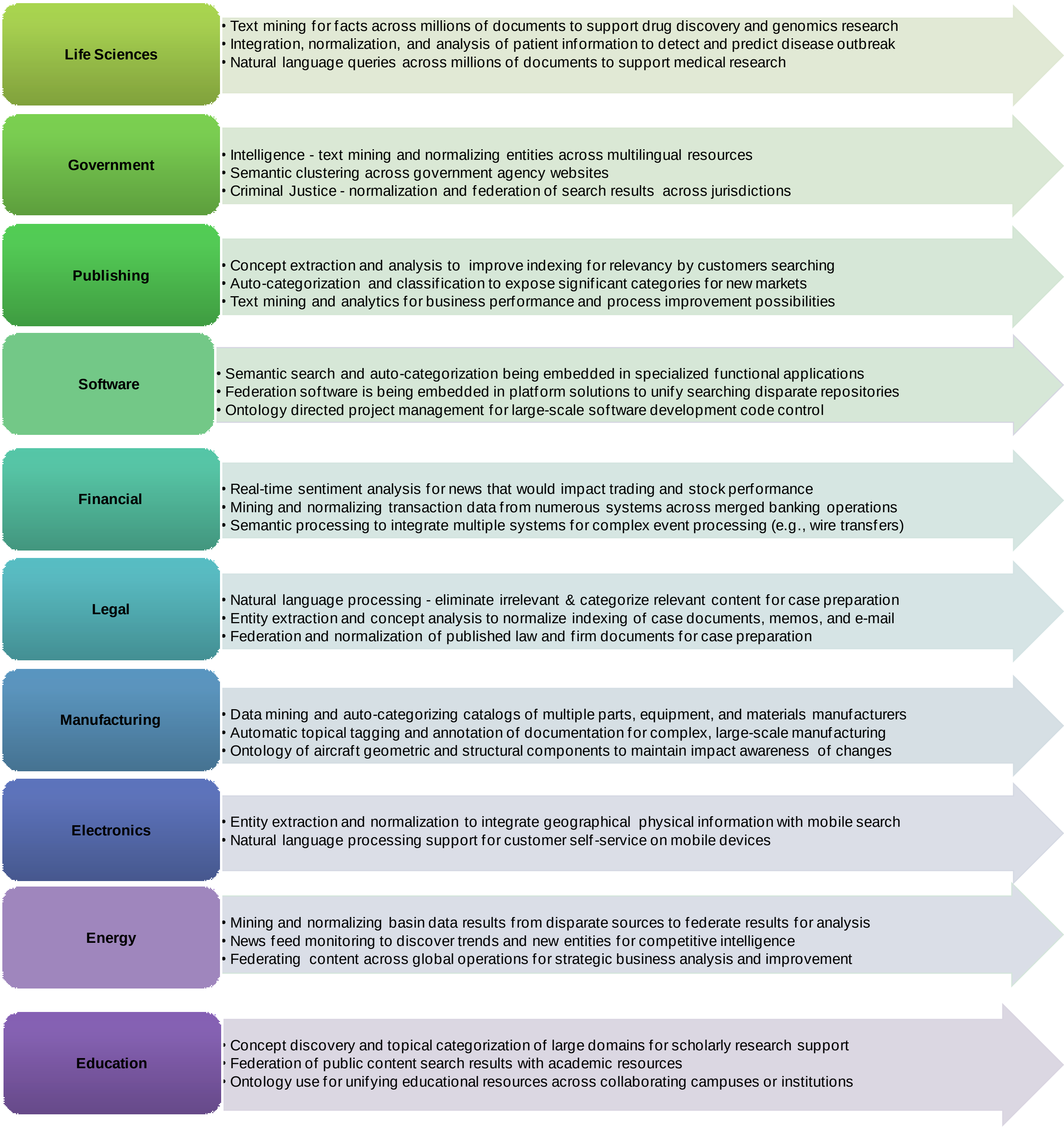
©2010 Outsell, Inc. Reproduction strictly prohibited.

Use of Semantic Software Technologies in Vertical Markets

Readers are always interested to know what technologies other companies in their industry are using to solve information management problems, even early adopters. It is certainly an easier sell for middle managers or team leaders in a functional area to point to how a competitor is gaining an edge through the use of new technology. The key for any company adopting an adolescent technology is to employ these tools better and smarter than a competitor. Following closely behind a lead adopter is not a bad strategy.

The next two illustrations show the semantic problems that are being solved by applications on the market today in specific vertical industries. These are solutions that target information challenges unique to an industry. Other products in the directory provide semantic solutions across many vertical markets. Figure 4 is organized in descending order with the strongest solutions designed for a particular industry in the top positions. Life sciences is a field clearly in a leadership position with adopters because of its very strong linguistic foundation in the form of controlled vocabularies, which have evolved into highly advanced ontologies. Other industries are making use of more generalized semantic nets, linguistic rules, and smaller domain specific ontologies.

Figure 4. Synopsis of Semantic Software Technologies Applied to Verticals



Life Sciences

- Text mining for facts across millions of documents to support drug discovery and genomics research
- Integration, normalization, and analysis of patient information to detect and predict disease outbreak
- Natural language queries across millions of documents to support medical research

Government

- Intelligence - text mining and normalizing entities across multilingual resources
- Semantic clustering across government agency websites
- Criminal Justice - normalization and federation of search results across jurisdictions

Publishing

- Concept extraction and analysis to improve indexing for relevancy by customers searching
- Auto-categorization and classification to expose significant categories for new markets
- Text mining and analytics for business performance and process improvement possibilities

Software

- Semantic search and auto-categorization being embedded in specialized functional applications
- Federation software is being embedded in platform solutions to unify searching disparate repositories
- Ontology directed project management for large-scale software development code control

Financial

- Real-time sentiment analysis for news that would impact trading and stock performance
- Mining and normalizing transaction data from numerous systems across merged banking operations
- Semantic processing to integrate multiple systems for complex event processing (e.g., wire transfers)

Legal

- Natural language processing - eliminate irrelevant & categorize relevant content for case preparation
- Entity extraction and concept analysis to normalize indexing of case documents, memos, and e-mail
- Federation and normalization of published law and firm documents for case preparation

Manufacturing

- Data mining and auto-categorizing catalogs of multiple parts, equipment, and materials manufacturers
- Automatic topical tagging and annotation of documentation for complex, large-scale manufacturing
- Ontology of aircraft geometric and structural components to maintain impact awareness of changes

Electronics

- Entity extraction and normalization to integrate geographical physical information with mobile search
- Natural language processing support for customer self-service on mobile devices

Energy

- Mining and normalizing basin data results from disparate sources to federate results for analysis
- News feed monitoring to discover trends and new entities for competitive intelligence
- Federating content across global operations for strategic business analysis and improvement

Education

- Concept discovery and topical categorization of large domains for scholarly research support
- Federation of public content search results with academic resources
- Ontology use for unifying educational resources across collaborating campuses or institutions

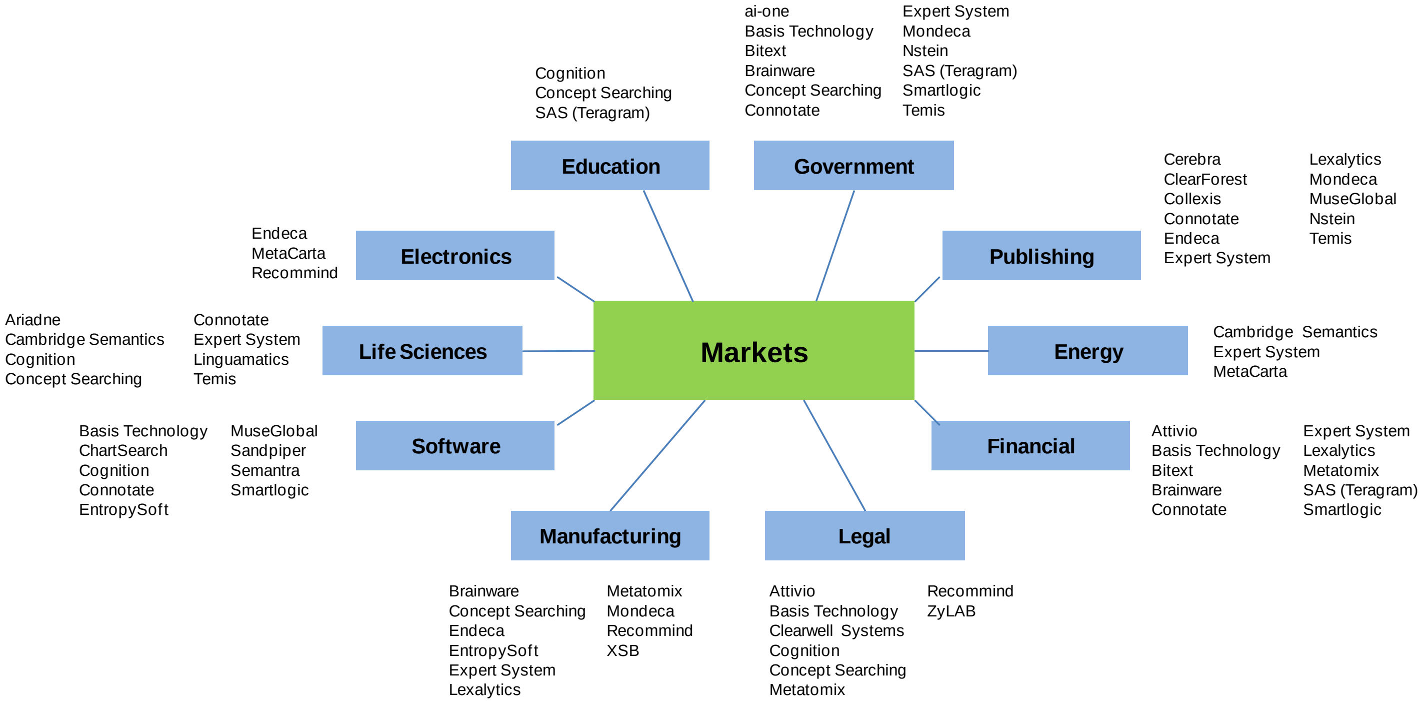
Source: Outsell, Inc.
© 2010 Outsell, Inc. Reproduction strictly prohibited.

Government, especially agencies related to defense, homeland security, and justice, are making use of all text mining and linguistic processing tools currently available, particularly those with multi-lingual capabilities. Publishing (including all media types) is text focused and relies on search that is accurate and relevant for exposure to its market. Some publishers are acquiring companies with semantic technology because it is that important to them. The latest are the Collexis acquisition by Elsevier and previously the ClearForest acquisition by Thomson Reuters. Software companies are embedding semantics into their applications so they are also a strong emerging market; the financial and legal industries need semantics to deal with the volume of text in their businesses.

Figure 5 was assembled from information provided by vendors in announcements about new customers and from the customer listings on their websites. Some companies make a point of defining the industry specific semantic solutions they provide and have case studies to illustrate those applications. Many companies offer semantic content processing or semantic search solutions that are appropriate across all vertical markets and those are included in the next section.

Companies in Figure 5 are listed in alphabetic order and are grouped simply to reflect broad positioning in those industries that are at the forefront of leveraging semantic software. Undoubtedly, there are case studies not yet posted or easily found but the trends are clear; some vertical markets have been penetrated by a significant number of vendors. Often one adopting company will have solutions from more than one vendor, each addressing a different semantic business problem in different parts of the enterprise. Early adopters tend to experiment with a lot of versions of new technologies before weeding out those that do not perform for them. A year or two from now, this landscape will look very different.

Figure 5. Landscape of Semantic Applications and Tools Across Vertical Markets



Source: Outsell, Inc.
© 2010 Outsell, Inc. Reproduction strictly prohibited.

It should be noted that many of these vendors have partnerships with other software companies doing business in particular vertical markets, not just those shown here. Vendors whose principal business has evolved into partner relationships with other software companies are highlighted with "software"; some of those also do direct marketing to specific vertical industries.

Use of Semantic Software Technologies in Horizontals (Functional Groups)

As companies with semantic software technologies reach market readiness, they must make choices about positioning. Some begin development with a clear vision of exactly what type of business is a strong candidate for the product they have developed. This is especially appropriate when a semantic net of domain specific terminology is embedded. Linguamatics has distinguished itself by focusing on life sciences, particularly pharmaceutical and biotech companies. Being able to leverage and extend the public domain UMLS ontology and other life sciences thesauri is a clear-cut point of definition.

Other companies such as Expert System, Cognition, Attivio, and Cambridge Semantics, for example, have more generalized approaches for broader reach. The first two have built out semantic nets over many years, Expert System for multiple languages and Cognition focusing on the English language. They are able to bring these highly evolved computational linguistic and NLP tools to any number of vertical markets that seek natural language processing. There are differences in packaging and service approaches, which require direct comparisons for serious buyers to discover what is most suitable.

Attivio and Cambridge Semantics have taken the approach that they would build on evolving semantic web standards, using open source tools to create components for broad reach into many of the functional areas of any vertical market where integration of data and unstructured content across repositories is much sought after. Having observed a decade or more of enterprise search drawbacks, often due to the lack of consistent vocabularies or metadata, they are seizing the opportunity to leverage semantic methods for improving the enterprise content construction and retrieval experience. They are using semantic building blocks to deliver better total content solutions.

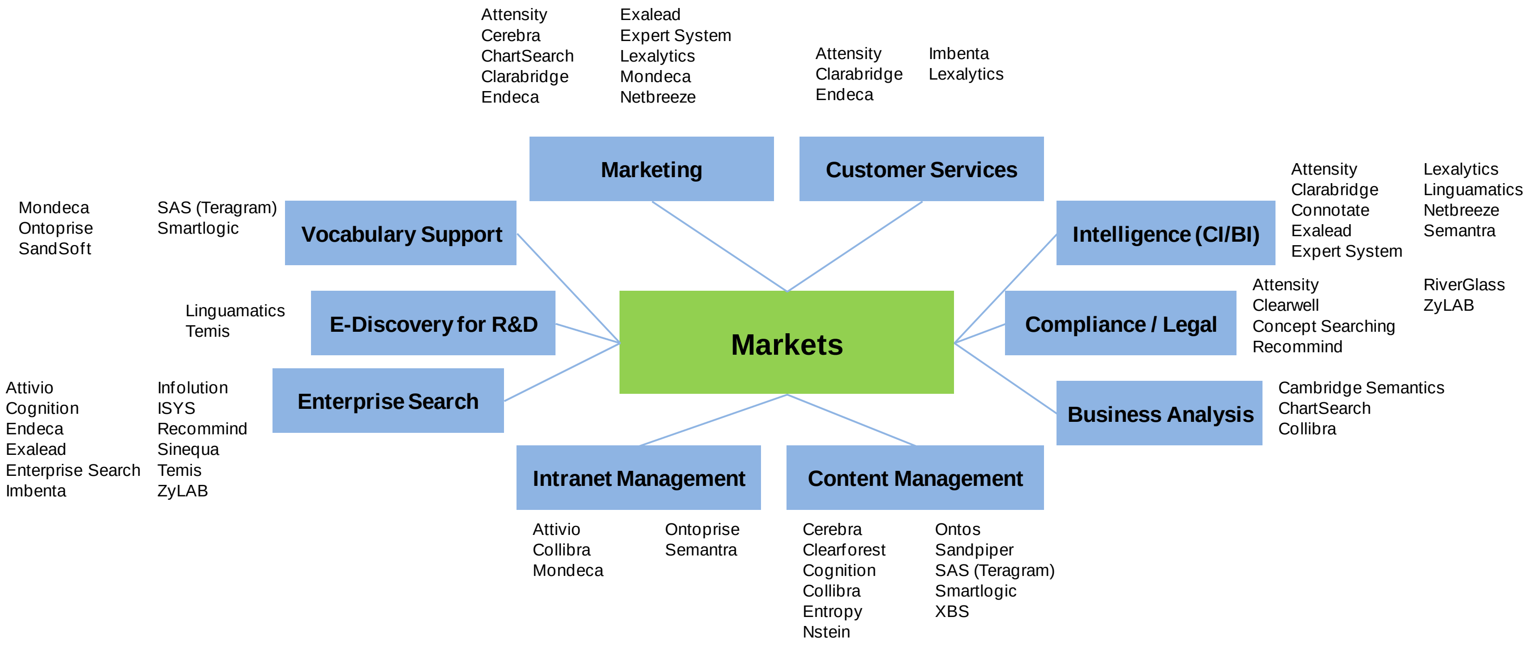
From the world of text analytics with semantic techniques for mining, extraction, transformation, and computational linguistics, Attensity, Clarabridge, and Lexalytics have found business in niches like sentiment analysis that reaches into all verticals.

While every functional business area of an enterprise is a candidate for semantic software technologies, the areas that are now catching on are illustrated in Figure 6. They are defined as follows:

- *Marketing* – Positioning, customer care (listening to the voice of the customer), social media tracking, sentiment analysis, and sales opportunity analysis;
- *Customer services* – Product support via web self-service, on mobile devices, and employing NLP to interpret and answer questions by support experts;
- *Intelligence (CI and BI)* – Sentiment analysis for business and competitive intelligence;

- *Compliance and legal* – Regulatory and risk analysis, e-discovery for case support;
- *Business analysis* – Strategic planning, data federation, financial modeling, and analysis;
- *Content management* – All the activities of content capture and enhancement that contribute to improved retrieval: metadata creation, ETL, normalizing content, and creating auto-categorizing rules or vocabularies;
- *Content platform development and management* – Where companies offer a range of semantic components that are building blocks for complete end-to-end solutions (e.g., intranet staging and enterprise integration of content and document management solutions);
- *Enterprise search* – Semantic search engines for intranets or portals embodying one or more semantic components (e.g., NLP, ontologies for auto-categorization, and federation);
- *E-discovery for R&D* – Application of scientific and technical semantic nets or ontologies for sifting millions of documents using natural queries to pinpoint answers to difficult or unique questions that may never have been asked before;
- *Vocabulary support* – Taxonomy, thesaurus, or ontology building and maintenance platforms.

Figure 6. Landscape of Semantic Applications and Tools Across Business Functions



Source: Outsell, Inc.
© 2010 Outsell, Inc. Reproduction strictly prohibited.

More Applications and Comments

We have taken a high-level view of what various semantic software products do for enterprises and described how the markets are reacting to these tools. It is time to look at a few of the cases that we think are most illustrative of impact. Descriptions come from those who have been using and trying to use these semantic software solutions; they are dogged and tenacious in their work. Also, through their experiences, they are very realistic about the state of the industry. None were expecting major breakthroughs, just incremental improvements as suppliers ebb and flow in the face of tough business climates and demanding buyers.

Commentary and experiences have a positive influence on the market; everyone is learning and gaining understanding. A pervasive attitude is that buyers need to be sensible about their software choices and realistic about what they can get out of each product. They need to choose the right type of product for each semantic challenge, and everyone seems to recognize that there are numerous kinds of semantic challenges and business problems to be tackled. We did not speak with anyone who was advocating a single solution for every type of business application.

When investigating vendors and products, readers should visit their websites and seek out customer success stories and case studies. Some of the following information from interviews is published and other was obtained with the promise of anonymity. To avoid bias toward any one product, the stories are presented without product names. The point is to create scenarios that are commonly experienced in this industry.

Life Science Applications

For both a major pharmaceutical research company, which is a heavy user of subscription content, and a content provider of life sciences publications, identifying all the key concepts in publications is a priority. In one case the direct business impact is to ensure complete retrieval of every piece of content that answers a science question to further research efforts. In the second, it is to ensure that customers find everything that is relevant to their searches, but not irrelevant content.

In both cases the content is the same but the reason for employing text mining, concept and entity extraction, and analysis is different. The net outcome for the first is to speed up scientific inquiry to bring better products to market faster by finding factual information already published and proceed with bench science more efficiently. The net outcome for the content provider is to create a more reliable search experience for their customers, delivering quality that will bring repeat and new business.

The immediate benefit for both is cost savings. The two different software products they are using mine content for entities and concepts, and create a type of categorization and tagging that dramatically improves accuracy and delivers the option for highly precise natural language queries or excellent topical navigation. This is done largely through automated processes, backed up by human curation when the systems detect ambiguity or conceptual confusion. The human subject matter experts respond to these content "exceptions" with the correct interpretation, thus improving the underlying ontologies of terms and term relationships.

In the case of the pharmaceutical company, the complex queries are done in real-time against the mined and indexed full-text, usually millions of documents, while the content provider uses tools to create metadata that will typically remain permanently with the documents. The first situation is one that calls for enormous flexibility in uncovering new facts (using new query terms) that have not been unearthed previously (more precision) while the second is designed to be more generically categorized.

In both cases the users commented on these benefits from the products they have selected:

- Minimize human curatorial indexing effort (by factors of 10 or more);
- Reduces time to process accurately complex, high-value content for precise retrieval;
- Entity tagging of the databases against published ontologies combined with proprietary thesauri and synonyms, enhancing the value of the results;
- Provision for efficient feedback loops designed to work smoothly with subject matter experts;
- Better retrieval that brings with it opportunities for uncovering new, unexpected and research stimulating information.

Publishing and Media Applications

Readers can judge for themselves how well semantic technologies work for the publishing industry when they use these publishing products or websites:

Expertise discovery is a major interest in any enterprise, often a strong mandate for the knowledge management team. Elsevier, a major publisher of scientific and technical journals, has adopted semantic technology to create expertise linkages across current and legacy content. A semantic map of entities and content relationships can be established by leveraging entity extraction to identify individuals and institutions associated with concepts, mined from content.

The Huffington Post has acquired semantic NLP technology to manage comments, particularly to moderate for offensive material. It is used to build up language rules for detecting content that should not be published, and provide rapid detection to quickly resolve the "Comments" queue.

Knovel has been a hit among corporate librarians with a heavy focus on scientific and technical reference content since it began "semantic-izing" standard references books from all major publishers about ten years ago. It illustrates the power of linking, combined with HLTs, and federation to deliver precise facts and the ability to manipulate table information in retrieved full text. Knovel has brought over 2,000 texts into their library of content, available by subscription. Scientists and engineers find the application of the technology outstanding, eliminating the need for individual desk copies of reference books and dramatically reducing the amount of time to look up and verify facts (like property data) across multiple sources. It provides direct answers to questions, straight from the most authoritative sources.

Traditional print news media is fighting to sustain its legacy reputation for quality journalism while entering the digital era, trying to find ways to monetize their content while making it readily accessible and priced to sell over the internet. This requires excellent metadata and a semantic search platform that ensures success for even the most naively worded queries. Many news organizations are working with semantic technologies for establishing excellent taxonomies, auto-categorization, and digital asset management (DAM) Financial Times, Bonnier, D.C. Thomson, and Hearst Newspapers are among the prestigious publishers who have invested in semantic content management tools from one vendor.

We have already noted the application of sentiment analysis to detect what customers are saying about a product or company. One news content service provider, BurrellesLuce, is benefiting from embedding these semantic tonal measuring applications into its service offerings. Another frequently cited reason to use sentiment analysis is for the visual clues to content tone that some applications provide in the form of graphs and charts that give an instant view of what is happening in the media.

Multinational Corporation Intranets

Multinational organizations, particularly after mergers and acquisitions that bring extremely different content repositories to the marriage, are seeking out semantic platforms to perform multiple functions. Earlier in the report we discussed all the possibilities for semantically enhancing existing content and also for using semantic search to bring many repositories into a unified view. One of our interviewees described a time pressured need to replace existing, multiple embedded search applications with federated search to improve relevancy, ranking, and performance in a few months. The platform needed to support hundreds of international organizations that would be collaborating on an international event.

The goal was simply to have single-point intranet search across all applications used by the sponsoring organization. They were able to deploy rapidly, due to the out-of-the-box components for installing and adding applications (about two weeks per repository). With no tuning they achieved improved search results, which became even better with tuning (adding terminology and synonyms).

For this application three additional benefits were the ease of adding new connectors for new applications, fast indexing and retrieval, and very impressive, intuitive contextual navigation that required no training.

Guidance for the Team

We have established that you can use multiple software approaches to bridge the gap between content at its point of capture and content at the time when it will be searched. Whether enriching content that lacks sufficient metadata or context, or performing deep concept analysis to discover the true meaning of content, software can bring significant improvements to your enterprise search and retrieval operations. *Guidance for the Team* is a digest of the best advice Gilbane heard from our interview subjects. We are redirecting their comments to you who are managing, championing, or funding a semantic software technology initiative.

For both web and enterprise, semantic content management means choosing a team to select and apply the appropriate software and addressing the questions in this checklist:

- *Intent – What* is the content for?
- *Packaging and marketing – How* will the content be delivered and how are you going to establish user expectations?
- *Audience – Who* is the population that will use the application and how do they look for content in their work (behaviors)?
- *Scope – What* is the depth and breadth of the content, and how much is there?
- *Infrastructure – What* is the mechanism and *who* is involved in selecting, implementing, deploying, and maintaining the content?
- *Planning and schedules – When* is it reasonable to build the infrastructure and do the software tuning, enhancements, and implementation that will fulfill the mission?

Having thought about content, the team that is available is as important as the software it will select and use. The preceding list implies a lot of decision-making and expertise. Consider all these points that depend on human interactions with semantic technologies and you will have a sense of just how important the team composition will be. People have been and will be part of the solution at every stage:

- Building applications;
- Building conceptual support frameworks (vocabulary and topical domains);
- Integrating content with applications that produce or use the content;
- Implementing technologies;
- Deploying, supporting, and maintaining infrastructure;
- Using applications.

To assume that technology is the only component in the solution is misguided. In the following example of content "scraped" from a web news item quoting a notable politician, we see an example of a semantic challenge that automation is unlikely to solve without significant human intervention.

> *You know, there are man's activities that can be contributed to the issues that we're dealing with now, with these impacts.*

A semantic processor would look at the surrounding context and try to answer these questions: *Who said it? Why is it being said? What does it mean?* A lot of context (e.g., time, place, and circumstance) is needed to even begin to extract the possible meaning of this statement using NLP. Extracting concepts when language is complex or foggy is a huge semantic challenge that technology alone cannot solve easily. There are still plenty of areas in which humans are needed to untangle complicated language and this is one of them. It serves as an example for those who believe that installing a software application is the solution to anything.

Team Composition

We asked people who have several years of experience as team leaders with semantic software where to find people like them and what competencies they like to have on their teams. Because the field of informatics and the use of semantic software in academia are relatively new, most team members bring little direct experience with these technologies to their work. However, having a multi-disciplinary background, being a conceptual thinker with an orientation to systemic modeling and problem solving are advantages. This coupled with having some familiarity with the subject discipline provides an edge with query formulation and tuning aspects.

One expert commented on a generational attribute; people with a lot of experience with electronic gadgets, social networking tools, and games seem to adapt well to the semantic tools. Being socially garrulous and intellectually curious was also mentioned. Some experts commented that librarians and scientists do not come to the technologies with a special facility for formulating queries. Asked if this might be because structured command language Boolean searching is so different than NLP, the experts acknowledged that this might be the case. However, everyone stressed that with time and practice, their team members do become more fluent.

Among the core competencies that are valued on teams were backgrounds in computational linguistics, informatics, subject matter experts in the domains being targeted, taxonomists, and use experts with enthusiasm and motivation to get better results from search. In our research, a blog entry last year carried an interesting concept, that of having a "Metator" on the team. For more about the role of an individual charged with vocabulary and metadata governance you might want to read Tony Byrne's comments in *Let Us Now Praise Metators*. Also, check out the books by Brandy King (about ontologists and semantics) and Heather Hedden (about taxonomists) in the bibliography.

Finally, it is important to comment about the role of IT professionals in this mix. We found a recurring theme that is not new to discussions about the adoption of complex software applications designed to be used by expert professionals in the course of their work. Of course IT must be involved in planning and installation because they will be responsible for the computer platforms and networks that support the software. Unit managers must take into account the lead time for getting infrastructure ready because semantic software does require additions and upgrades to existing systems. Crunching millions of full-text documents is computer intensive, whether doing text mining, building large complex indices, or direct retrieval on very large indices.

The tension usually occurs in the selection and implementation phases when IT asserts control over evaluation criteria, and fails to respect the need to give support for POCs. This is compounded by often weak communication from team leaders and the experts who will be implementing and using the software. The effort must be collaborative and well communicated at every stage. In enterprise software planning, there must be a differentiation between point solutions that can only be truly evaluated by experts using them, and general business platforms that everyone in the organization will use. The roles of IT for each class of software are very different; for the first IT must be supporting actors and for the second they will play a leading role. Having computer science experts with database and scripting language experience to support the primary team is necessary for most implementations. We encourage enterprises to sort out the issues of ownership and leadership before making any moves toward acquiring semantic software technologies. It is complex enough without getting into turf battles over its selection and implementation.

The following steps leading up to product usage is a checklist for planning and deciding the best people to have on the team for their competencies, at each stage.

Selection and Procurement

One of the people interviewed commented on the overly laborious process that many organizations engage in to plan for and select a product. The pace of change for software technologies and the internet is so rapid that the web from one instant to the next is completely altered. Likewise, any problem you are trying to solve right now will change in some, perhaps many, aspects before an enterprise team can even begin to deploy its product choice. Numerous products are referred to in this report, and more appear in the vendor directory. Narrowing down possible solutions quickly is the first step because looking at all of them will take more time than is practical or necessary. Here are some guidelines to get to a short list faster:

- Applications that semantically improve query interpretation and retrieval on the internet are driven largely by commercial interests and may not work well or be appropriate for the enterprise. With some adaptation and re-packaging they may be very useful for enterprise use. Eliminate any that cannot be easily adapted to your business purpose.
- In either case (web or enterprise) constant content enhancing, adding, and subtracting, plus updating terminology is challenging and requires good tools for making changes. If what you want is a single solution that gets plugged in and forgotten, you will find semantic relevance will degrade over time. Governance and tuning will always be required and have a human component, an expense that must be considered.
- Management must be sold on the imperative for human engagement in the process of keeping up with a changing domain. Make sure there is a budget for software, ongoing software maintenance, and permanent human support for governance and maintenance. Know what the budget numbers are to keep your selection in a range and talk to customers of the solutions you are considering to understand the human overhead requirements, both internal and external support.

- Semantic software technologies can chew up significant computer resources. Be prepared to give vendors good parameters for the problem you are trying to solve, including the amount of content you expect to work with in the short and long term. Request an evaluation and general plan for your computer infrastructure requirements.
- Identify every absolute requirement, which if not present will eliminate a product. This includes political considerations and biases against certain classes of products, companies, and operating systems. It is a waste of time if you fail to overcome objections early and continue to look at products that will not be funded by your management.
- Get a plan and schedule with milestones in place. Include:
  - Preparation of requirements;
  - Time to go through a "request for information" process from vendors, restricting it to unique requirements that will differentiate vendor products;
  - Coordination with IT for proof-of-concept and determining final selection implications for computing resources;
  - Plan for a POC cycle for two or three solutions;
  - Use cases that represent at least the primary reason for selecting a product;
  - Team assignments for selecting candidate products, evaluating vendor responses, and conducting the POC;
  - Conducting and evaluating the POCs;
  - Procurement and contracting process;
  - Implementation;
  - Testing;
  - Deployment and training;
  - Ongoing maintenance and governance.

A note about proofs-of-concept: A few years ago when interviewing an IT person leading the investigation for a search product, we received a disturbing response about a POC being a "waste of time." The solution would be selected and if did not work out, another would be brought in. Given all the planning, implementation, and execution needed for selecting one product, it was difficult to understand this person's willingness to engage in the process multiple times.

Besides not being able to understand how a product would work or behave in a particular business environment without a POC, there is the issue of what is happening to the target user community while products are being tried and failing. They are impatiently waiting or moving on with their own solutions, an approach that leads to information chaos, a situation impossible to govern. POCs are highly recommended and the process is worth the time spent. You will learn more in this stage that will better prepare you for the end game of product implementation. More about POCs is described in *Product Administration and User Interfaces.*

After establishing the business need, type of semantic solution, and preparing a list of most likely vendors to consider, you need to establish a few fundamental criteria for narrowing your choices, even before engaging in a proof of concept. Two areas stand out that might be considered in haste after settling in on a single solution. These are important to think about before you engage in conversations with vendors and their customers: product packaging configurations and product administration options. Be prepared to ask lots of questions about the options available and ask existing customers their choices and how those choices have worked out for them.

Product Packaging Models

Software comes in many forms and most vendors offer multiple options. Here are some choices that may be available and the questions to ask:

- Does the software license provide for in-house installation and what configurations are there: installed on the desktop for single user, on a shared server or over a network?
- Is software-as-a-service (SaaS) an option? If there is a choice, the cost of "renting" use on external servers can be a good testing situation but may be more expensive if longer term use is already determined.
- Will the product being considered perform what you need as a standalone product or does it need to be integrated with other software to give value? Are there components that are embedded in other software and pre-configured for a special use already available in-house? Is the product of interest one of a suite of products and will you be planning for adoption of the entire suite eventually?
- How does any content enhancement tool relate to other systems already in use (e.g., search engines)?
- Are there tools for developers to extend the application and do you have the expertise in-house to use them, or will services be required and how easy are they to obtain?
- Open source is ubiquitous, and many of the standards for semantic technologies have open source options available for creating applications that conform to those standards. However, system integrators and services are almost always required from outside; knowing from whom and how available those services are for a given product is a question that must be answered.

The answers to these questions get to the issues of evaluating total-cost-of-ownership, human resource requirements, and infrastructure required for products to perform the functions you need.

Product Administration and User Interfaces

Proofs-of-concept are strongly recommended for reasons already touched on. Knowing, with confidence, how a product works in any situation, given a specific enterprise infrastructure, and an existing team for implementation and maintenance, requires that it be tried *in situ*. It takes time, planning, and commitment of human resources for thoughtful POC implementation and testing of use cases.

During a POC it is common to have a third party do the installation of all products being tested to ensure a uniform baseline installation. We recommend that an internal resource be present during set-up, observing and documenting decisions and choices made. Also, ensure a sufficient amount of content to give the products a complete work-out; this will vary for the type of application. Once installed here are some areas to examine during the POC:

- Test all the "turnkey components" without any tuning to understand the product baseline and how it behaves with whatever content problem you are addressing.
- Determine what is "black box," inflexible or only tunable by the vendor.
- For any product that claims to support inclusion of enterprise unique vocabularies, work through how the curation of terminology is done, where it impacts the nominal functionality, and if it does what is needed.
- If searching is reason for using a product, test and practice tuning for relevancy.
- Test and practice query formulations that target content you know exists and answers questions you know are in the content.
- Study query results and understand why every retrieved item appears; scrutinize the form of results displays and order for suitability to your audience.
- When security is an issue and select content requires permissions, test those permissions for any holes in the system that might allow improper access.

Implementation and Deployment – Miscellaneous Comments and Caveats

Our discussions with customers did not always provide consensus on how to proceed but the strongest sentiment was about vendor relationships. It is not surprising that early adopters would instinctively understand how much their success depends on having the vendor succeed. Those who successfully forge collaborative and mutually respectful relationships with their software suppliers have much greater opportunities for improving and influencing software development. When a company determines that a solutions provider is an honest and receptive partner, and establishes that there is essential expertise that can contribute value, it is in each partner's interest to make the relationship work.

There are two other things you can do to further sustainability of a genuinely valuable technology. The first is the willingness to share valuable experiences with the professional communities in which the software plays, either by writing papers, blogging about successes, or a willingness to talk with other prospects directly. The second is having realistic expectations and making sound judgments about schedules and performance. This means that communication must be continuous and good questions must be asked before jumping to conclusions or making changes to vendor recommended practices for implementation.

Along these lines, here are things to think through before make hasty decisions:

- This type of software requires time by implementers to become proficient. Time and training must be extended to target audiences.
- Expectations and full discussions of the purpose of the software and how it works must be shared. It seems obvious, but many times end-users are not told what content they can expect to find or the types of questions they should be able to answer. Boundaries must be articulated.
- When a large body of content in multiple domains is part of the semantic landscape, team assignments of subject matter experts need to be considered. One firm in the publishing industry recommended spreading the subject nodes across editorial specialties. Managers needed to understand that some subjects are fast-track, while some take more effort and are more manual. Over time they reached equilibrium among expert assignments and workloads but there will always need to be extra editorial oversight for new areas of research
- For enterprise intranet projects, federation receives a lot of discussion in print and among experts. This threaded discussion, *Federated and/or Universal Enterprise Search - Real-life Experiences?*, in the LinkedIn Enterprise Search Engine Professionals Group has interesting comments for those who are members. Elsewhere, we received some cautions about federating across multiple enterprise repositories, noting that you need to "watch out for federator applications and …security authentication/authorization pass-thru – that can get tricky."
- Integration with infrastructure and other applications needs to be reviewed periodically. Any place in the software application landscape that consolidation of shared content can be achieved is a cost savings. With experience and increased team expertise using the tools, more opportunities can and should be found.
- Enterprises should not overlook planning for other technologies or next generation technologies. Over the past three decades there have been software revolutions, the most intense in the past ten years. We all have enough experience with legacy systems to understand the risk of getting stuck with unsupported tools that have no trained experts to use. Plan for and know when it is time to move on.

Semantic Technology Standards

The short story on semantic technology standards is that developers talk about them and engage in standards communities all the time. Customers do not. Still, buyers need to be aware of discussions and evolving standards. The vendor directory of companies and organizations contains links and brief comments about the nature of the emerging groups directly involved in semantic technologies. The W3C is the most prominent and covers most aspects but their work is evolutionary and is under continuous review. The most commonly mentioned by developers are OWL, RDF, and SPARQL.

Controlled vocabularies have the longest history of standards development and are established as ANSI standards available through ISO. ANSI/NISO Z39.19 – *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* is universally accepted as the standard for thesaurus development and many of its tenants apply to developing taxonomies, as well.

Getting into the Customer's DNA: Vendor Guidance

We have probably learned more about semantic software technologies from vendors than their customers over the past few years. There are, however, common themes from customers that can benefit vendor growth and improve adoption. These are the most significant.

Earlier in this report is guidance to customers about relationships with their vendors. Vendors, the ones we have spoken with, also recognize and encourage strong relationships with their customers and welcome each new customer as a partner in extending their technologies.

Whether the customer is a large, well-known organization or a small one, this is particularly important. We have observed a couple of early adopters of search technologies, small professional firms, in which the search engine champions have been substantial differentiators for the start-up company. They really got into the trenches with the vendor and worked on areas for technical, packaging, and deployment improvements, and were listened to. They also presented their case studies at numerous conferences and talked about their collaborative relationships. Most importantly, the vendors were excited about what their customers were doing with the product, and listened to any issue comments; clearly the customers sensed the commitment to their enterprise success. The vendor is doing extremely well now and customers have had internal success stories to tell, as well. Nurturing every client relationship is crucial.

These are the top areas where customers need your assistance, directly through services, good account management, and surrounding collateral materials:

Product packaging that makes sense for prospects in their industry or functional area. It has to be clear what they are licensing or subscribing to, what it does, and how it is intended to be used. Point solutions are sought after, recognizing that vendors need a critical market mass to package for a niche audience. The simpler the product is to understand and procure, the better your chances for success. This includes good labeling, clear infrastructure requirements, documentation, and installation support.

Proofs-of-concept (POCs) are recommended to buyers. Too many software products have been procured as platforms that required significant development to actually work. Unless you are in the tools, services, or systems integration business, products need to be used "hands-on" by your prospects. It is not sufficient to evaluate any product of the complexity of semantic software applications by seeing demonstrations. Prospects need to use content with which they are familiar, in their own environment, with experts testing real use cases that are meaningful to them. Your own business success may hinge on how well you deliver the product for this evaluation and support the effort.

Pricing models need to be simple and make sense within a prospect's budgetary constraints. It is probably well understood by now in the software industry that very expensive licenses with heavy ongoing support and service charges do not often achieve the highest revenues. Finding ways to package and price for more sales is the way to operate with the expectation that good product plus good service will lead to recurring business, add-on sales, and new opportunities in the organization.

Staffing for superior service and support during new customer adoption phase is critical. Our experiences in the software industry demonstrate that getting a new customer off on the right path, with all components fully operational, and team members up-to-speed is worth a lot over the long haul. They cannot be left hanging for answers. For the early weeks or months after installation customers want and need a lot of "hand-holding" and have numerous questions. The payoff for you is that once these well-supported installations stabilize, your support overhead for them will decrease dramatically. The goal is to leave them in a position to be experts in your software and the confidence that you will respond when needed.

Champions and evangelists are individuals you need on your team. Seek out the influencers in new customer organizations who went to bat for your product's selection and stay in touch. If there are issues or special circumstances that need sorting out, involve them in discussions (if they are not already) and seek their guidance to resolve problems. Ask them lots of questions and be open to what they can tell you about how to operate in that enterprise. They may also have good insights into the larger marketplace and, if your relationship remains healthy, referrals will come.

Partnering with complementary technology vendors. Because your success as a vendor is dependent on many variables, including those over which you can exert little control, it is important to find and solidify relationships with other software vendors whose products need to integrate with your offerings. By having good lines of communication and providing plug 'n play capabilities with other software that is frequently installed at customer sites, you will reinforce the perception of your company's significance in the software technology arena. It just makes good business sense to do what is in your customers' best interest.

Summary and Status of the Semantic Software Industry

Skeptics appear at the advent of any new technology market as evidenced by one search consultant who made this comment to our original question, which was "Semantic search technology – does it actually exist?"

I'm a convert to entity extraction, but other than that, am skeptical.

We know that fads come and go and labeling can get buyers and sellers into a state of confusion. We did not try to boil the entire "semantic sea" of technology but focused on where the most growth appears to be in 2010, computational linguistics-based and natural language processing enabled technologies. Speech recognition, translation, image processing are related technologies that are on the upswing, as well, and arguably in the "semantic family." We are watching closely how they will impact content retrieval and improve semantic understanding.

In the meantime, we see a class of semantic software technologies that have gained important market share in knowledge-based industries and very strong interest that will continue to grow. A critical mass of adopters is years away because buyers need to be educated about how software understands: textual meaning and contextual relevance. Industry growth requires a new type of expert and these professionals are in short supply: computational linguists, subject matter experts, search technology developers, and ontologists.

The current state as we canvas the landscape shows that technological gaps are closing but there are human resource gaps, both in terms of expertise but also staffing constraints. Temporal issues will continue to be a challenge because implementation and fine tuning take time to learn, but once expertise begins to take hold benefits accrue. Becoming proficient at linguistic mapping and query formulation requires intelligence and artful competencies that are not innate.

We cannot predict technology breakthroughs that might enable some leapfrog improvements to semantic software. Several entrepreneurs were interviewed who are certain that they have game-changing technology soon to be released. That would add to the mix and interest in semantics.

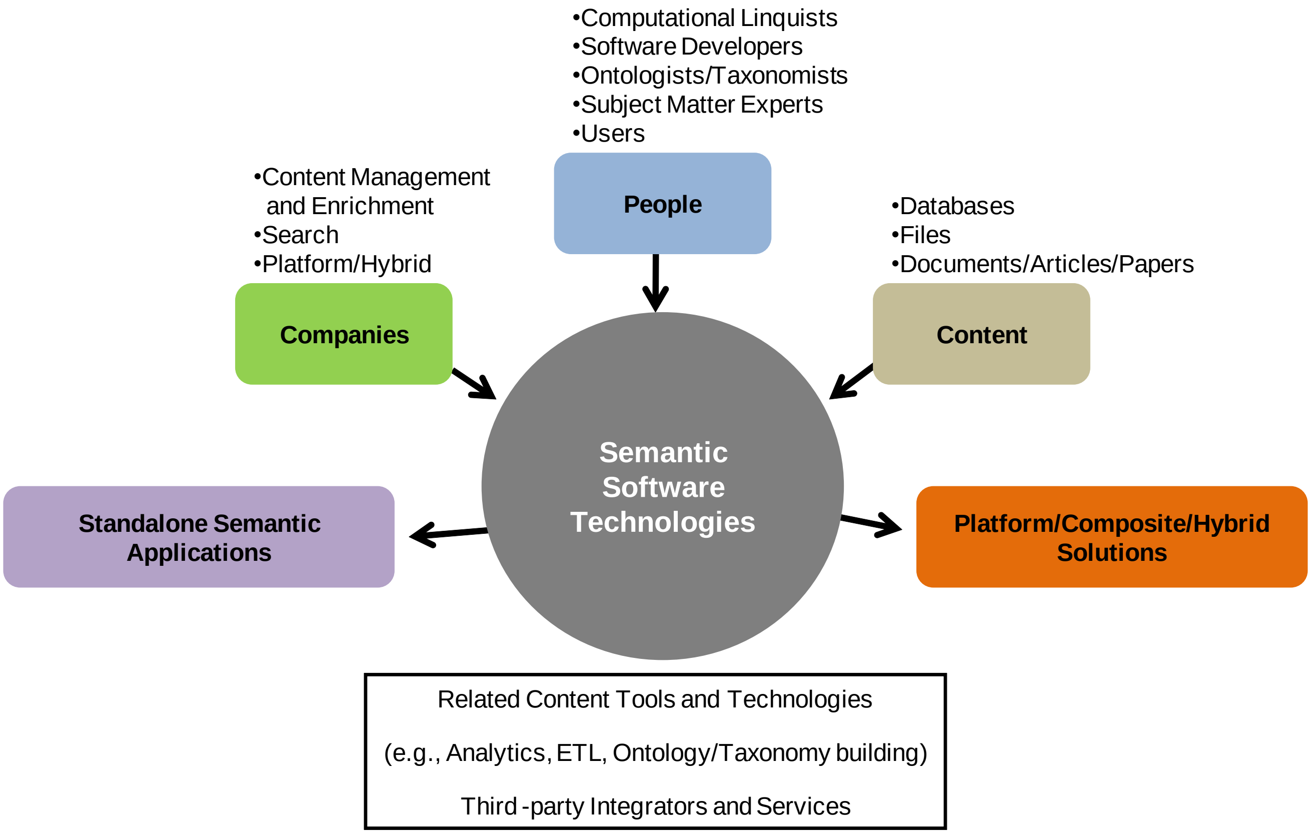
Risks are being assumed by technology developers in increasing numbers, which is a strong indicator of how large a role information overload is playing in our professional lives.

Beware the skeptics, especially computer scientists who worked hard on AI development projects in the 1970s through 1990s and never saw a lot of commercial success from these efforts. Judging from the many applications that have and are accruing customers and genuine success stories, newer applications are solid. We believe that packaging is part of that success, making software that fits efficiently into a particular workflow paradigm while meeting a special business need. When coupling product workability with the conversion of vast corpuses of content to electronic formats, the opportunity for making an impact by leveraging that content using semantic software are substantial.

This sums up the way Gilbane views the landscape in 2010:

- No *technology* is a solution for *all* the semantic challenges an enterprise faces.
- No single *software product* is a complete solution.
- Quality *people* to take care of language management and ongoing content curation are required and hard to find. They are essential to the successful implementation of any semantic technology.
- The industry is not well formed and still sorting out it sub-domains.
- Consolidation, partner collaborations, and integration of tools will be the norm for several years.
- Adoption will drive understanding; user insights will ramp the dialogue between buyers and sellers improving product packaging and integration, and clarifying definitions.
- Standalone solutions for improving content accessibility and retrievability will dominate early adoption; as companies gain more experience with semantic technologies and their practical applications they will migrate to more cohesive implementations of composite, hybrid-integrated or platform solutions.

Figure 7. State of the Semantic Software Technology Industry, 2010


•Computational Linquists
•Software Developers
•Ontologists/Taxonomists
•Subject Matter Experts
•Users
People
•Content Management and Enrichment
•Search
•Platform/Hybrid
Companies
•Databases
•Files
•Documents/Articles/Papers
Content
Semantic Software Technologies
Standalone Semantic Applications
Platform/Composite/Hybrid Solutions
Related Content Tools and Technologies
(e.g., Analytics, ETL, Ontology/Taxonomy building)
Third -party Integrators and Services

Source: Outsell, Inc.
© 2010 Outsell, Inc. Reproduction strictly prohibited.

Appendix

Glossary of Terms Related to Semantic Software Technologies

The terms in this glossary are given explanations that relate to their use in the body of the report and sponsor deep-dives. Some of the terms have other definitions in other contexts. There was no attempt to cover other uses.

Term	Description
AI	See: Artificial intelligence
Analytics	See: Text analytics
Application programming interface	Vendor supplied add-on software tools to facilitate programming new features or functional enhancements to integrate a software product with other software. Also referred to as API
Artificial intelligence	Software to instruct a computer to perform operations or activities that are normally thought to require human conceptual reasoning or thought.
Attribute	An entity that defines a property of electronic content, content object, element, or file. (e.g., Date of modification)
Auto-categorization	See: Categorization
Boolean searching	Use of explicit commands to define the relationships between terms. The commands limit or narrow the scope of a search (AND), expand its scope (OR), or exclude explicit content (NOT). (e.g., search for content limited to containing both "energy" AND "solar" where AND is the command.)
Categorization	A computational or human activity assigning labels to sets of content to explicitly organize or sort according to labeling.
Category	A human defined "grouping" label to systemize how pieces of electronic content will be organized in a particular domain. See also: Class See also: Taxonomy
Citation	Information that accurately defines and describes a publication or data file; structured bibliographic metadata. See also: Results
Class	A collection of content sharing a common attribute or property.
CMS	See: Content management system
Computational linguistics	An interdisciplinary field that involves both linguistics and computer science. It is concerned with automating the analysis of human language and applying that analysis in software programs.
Concept	An idea or thought that corresponds to a word or set of words (term) in linguistic expressions and thus plays a part in the understanding of a larger piece of content. The context in which terms exist provide additional meaning to help interpret meaning.
Connectors (Federation)	In federated search, software modules that link and exploit content across disparate data sources. They make it possible for each source's content to be handled by search software in a unified operation (e. g., searched concurrently).
Content	The target of search regardless of format or medium. Everything included in a database, collection of files, or application repository.
Content management system	Software application that supports document creation, modification, or importation in a systematic and governed environment, usually with multiple users collaborating.
Context	Surrounding content that elucidates and clarified a linguistic unit of content and helps to determine its interpretation or meaning.

Term	Description
Controlled vocabulary	Terminology from approved lists used for tagging content. See also: Taxonomy See also: Thesaurus See also: Ontology
Crawl	Software process, usually part of a search engine, that traverses a specified domain or set of domains for the purpose of indexing all content encountered. Sometimes referred to as *spidering*. See also: Indexing
Curation	Human oversight process with functions related to editing, monitoring, reconciling, and governing operations related to managing semantic content software applications. Vetting software processing outcomes for anomalies and incorrect results.
DAM	See: Digital asset management
Data federation	Organized content state formed by merging and normalizing a collection of similar electronic information objects. See also: Federation
Database	Repository of data organized by explicit records and fields, or tables, rows and attributes.
De-duplicating	Identifying and eliminating data redundancies, usually operating on discrete content resources.
Digital asset management	A type of content management that automates the application of rigorous governance rules for how the content is created, modified, and maintained with access controls.
Disambiguation	Establishing and identifying a single grammatical or semantic meaning (sense) of a word or phrase in a given context.
Domain	A corpus of content bounded by system architecture definitions.
Dublin core	A standard 15-element metadata element set maintained at http://dublincore.org/ as a baseline for content. See also: Metadata
Embedded search	Retrieval algorithms delivered as a part of a software application for searching the content within the application.
Enterprise search	Software used to index and retrieve content that exists within or for an organization, ideally optimized for specific enterprise business requirements.
Entity	Unit of content that is meaningful for purposes of indexing and categorizing for a particular audience.
Entity extraction	A process of content analysis by which the software identifies and classifies data by type or attribute for the purpose of creating metadata for unstructured content from which it is extracted.
ETL	Extract, transform, and load suite of algorithms or programs. See also: Extractors See also: Transformers See also: Loaders
Extractors	Software programs that harvest data content from databases, files or other applications, usually for the purpose of then manipulating the data for eventual exposure to other applications or search engines.
Federated search	Process of retrieving content either serially or concurrently from multiple targeted sources that are indexed separately, then presenting results in a unified display.
Federation	Collection of automated processes for a multi-domain environment (internal sites or a mix of internal and external) to facilitate searching all domains simultaneously with a single operation. Across domains it supports at least four distinct functions: Integration of the results from a number of targeted searchable domains, each with its own search engine. Disambiguation of content results when similar but non-identical pieces of content might be included. Normalization of search results so that content from different domains is presented similarly. Consolidation of the search operation (standardizing a query to each of the target search engines) and standardizing the results so they appear to be coming from a single search operation.
Filtering	Applying additional search criteria to narrow or alter the results of an existing search or stored search strategy. See also: Boolean searching

Term	Description
Full text	In searching it refers to the entire document in contrast to citations only. Often intended to correspond to unstructured content.
Fuzzy	Describes words and phrases that require some form of disambiguation using software algorithms.
HLT	See: Human language technologies
Grammar	A system of rules and principles for speaking and writing a language. The study of structural relationships among words and phrases in sentences.
Human language technologies	Software or programs that function to identify linguistic properties in content for the purpose of "understanding" meaning and intent, usually applied to enriching retrieval experiences.
Index	Systematically arranged list; in computerized systems it is a representation of content to speed retrieval by the governing algorithms.
Indexing	A human intellectual process for organizing content to optimize retrieval. A computerized process for organizing content to optimize retrieval.
Interface	The architecture controlling the methods and design through which a user interacts with a software application.
Keyword	Non-controlled terminology; language extracted from the content literally.
Keyword search	Query request for literal text as crawled and indexed by a search engine.
Lemma	A term in linguistic data processing that refers to the base or canonical form of a word in a running set of forms words (e.g., ride is the lemma for rode, riding, and ridden).
Lexicon	The vocabulary, including words and expressions of a language. A language's inventory of lexemes.
Linguistics	The scientific study of a language including the nature, structures, and variant meanings constituting the language.
Loaders	Software applications designed to transfer data from one database to another, often coupled with extractors and transformers.
Metadata	Explicitly defined labels for structuring content that describes any document or file regardless of the native format. In a library system: Bibliographic elements or fields. In a file format: Properties.
Morphology	The study of the structure and form of words in language including inflection, derivation and the formation of compounds. A component of linguistics.
Natural language processing	Use of computers to interpret and manipulate words imparted in the form of human (natural) languages.
Natural language query	Search expression posed as a question by a native speaker who asks for information through a software interface.
Navigation	Method of searching by traversing content with a device (e.g., mouse), or accelerator keys through a structured layer of content to reach other content (e.g., drilling down through a taxonomic structure). The navigation layer is a controlled vocabulary list, organized by categories or classes, often hierarchical in their conceptual arrangement.
NLP	See: Natural language processing
Normalization	Standardizing or making consistent through a process or processes to create uniform format, language, and structure for data that needs to be consistently and meaningfully stored in a database and/or aggregated and federated upon retrieval. See also: Federation
Ontology	An assembly of terms in which all possible relationships that might exist between and among terms to express all concepts are explicitly mapped.
Open source software	Software available without licensing costs and customizable by the acquiring organization or by a third-party. (e.g., Lucene)
Parse	The analyzed text, made of a sequence of tokens (words and phrases), to determine its grammatical structure with respect to a given (more or less) formal grammar. To process language in preparation for semantic linguistic analysis. See also: Syntax

Term	Description
Portal	Web-based page of links serving as points of entry to specific content, other websites, and applications.
Prompt	Interface symbol or text indicating that a user response is required to proceed with the transaction.
RDF	Acronym for Resource Description Framework. A W3C standard framework based on the XML standard for describing and interchanging electronic metadata. It is used for integrating various web-centric activities including: sitemaps, content ratings, stream channel definitions, search engine data collection (web crawling), digital library collections, and distributed authoring. A language for describing relationships among electronic content application resources using specific vocabularies (ontologies) to leverage existing knowledge models for information re-use.
Relationships	Connections or associations between terms expressing hierarchies, components, membership, operations, and other forms of connectedness.
Repository	A database or file structure for electronic content; entity within a searchable domain.
Results	The output content of data retrieved in a search.
Retrieval	Process of accessing content through the act of searching.
Search	Process classification for all software designed to retrieve content whether embedded in a larger application or a standalone package. The act of retrieving.
Search engine	Software with algorithms specifying how data is to be retrieved from one or more indices.
Search platform	Suite of software products that together enhance simple index searching with additional functions related to content (e.g., transformation, analysis, and reporting)
Security	In a search environment, the search engine functions that support access controls to content applying authorization and validation rules.
Semantic net	Abbreviation for semantic network. A network of vocabulary mapping all the conceptual relationships among terms.
Semantic search	Use of natural language or meaningful queries to find content through retrieval software designed to understand complex questions and the linguistic concepts in the target content.
Sentiment analysis	Tonal or judgmental evaluation of content based on linguistic processing of the text, typically to discover positive or negative expression.
Site search	Option using navigation or a search box to retrieve content only from a specific website (URL) domain.
Stemming	A form of fuzzy search linguistic processing that reduces a word to its fundamental root and looks for any word with that root. (e.g., a search for *stemming* would also retrieve *stem*, *stems*, and *stemmed*)
Structured content	Data stored in a database or explicit metadata associated with a piece of content stored in a software application.
Structured search	Use of pre-defined forms or explicit commands to give bounds to query criteria and parameters. (e.g., restricting the search for a word to the title field)
Synonyms	Words or phrases with a meaning that is the same as, or very similar to, another word or phrase. Equivalent terms in vocabulary lists or ontologies.
Syntax	A set of rules for combining words and phrases to construct sentences in natural languages.
Tag and tagging	Used for semantic labels or functional labels indicating the purpose of a topic or conceptual string. Different than cataloging, in which metadata values are being assembled congruently to the content. Tags usually reside embedded in the content for index processing.
Taxonomy	Hierarchically ordered list of terminology approved for tagging or categorizing a corpus of content. Also, often exposed in the search interface to form the framework for navigated search.

Term	Description
Text analytics	A set of linguistic, statistical, and machine learning processes that model and structure textual sources for business intelligence, exploratory data analysis, research, or investigation. Post processing of mined text to derive additional information from the values extracted during text mining.
Text mining	Extracting interesting and non-trivial information and knowledge from unstructured text. Interdisciplinary field that draws upon: • Information retrieval • Data mining • Machine learning • Statistics • Fact extraction • Computational linguistics
Thesaurus	A list of terms that are assigned simple relationships, cross references, scope notes, usage notes, and other directives. A thesaurus is often more comprehensive than a taxonomy but less complex than an ontology.
Transformers	In data and content management, tools to normalize or otherwise systematically change data.
Triples	RDF statements that consist of a subject, predicate, and object, which correspond to a resource (subject), a property (predicate), and a property value (object). A fundamental construct in natural language processing.
Unstructured content	Content not organized in a formal structure; files not in a database (e.g., a Word document)
Visualization	Graphical or image representation of data to reflect some understood relationships that reflect information or reveal knowledge about the data.
Web search	Retrieval from a domain of content exposed to a single or multiple websites.
XML	Acronym for eXtensible Markup Language. An infinitely customizable markup language for defining metadata tags and descriptions of kinds of content within or applied to a domain of content.

Vendor Directory for Semantic Software Technologies

We have researched and reviewed scores of lists of exhibitors, industry publications, press releases, and websites of companies and organizations that have appeared over the past five years in some context that is related to semantic technologies, as described in our report. Knowing that our study and this directory will be the basis for making decisions about where to look for tools, services, and products the directory includes the following:

- Companies licensing semantic software technology products for the enterprise;
- Semantic platform or component developers and system integrators;
- Web semantic search engines;
- Standards, governmental consortia, or industry organizations.

The companies listed are those that, at this writing, we believe offer out-of-the-box products already established in enterprises, components, services, natural language querying on the web, or opportunities for understanding more about semantic software through communities of practice. In some cases the placement of a company in one list or another was subjective; therefore the rationale and other qualifying criteria precede each table.

The directory is alphabetic by company name; when a product name has a registered trademark and appears routinely in lists or industry news, those are included with a cross-reference to the company name. URLs are usually for the home page of a company but occasionally the URL will lead directly to the aspect of their business that is directly classified as semantic software or service.

The category tags map to this list and are the principal semantic technology "applications" stressed in the marketplace by vendors themselves. As products are always evolving, morphing, and being subsumed by other products, we recommend a visit to vendor websites for a better understanding of current options. These category tags are just a start. The irony of finding the categorization process for semantic products so challenging is not lost on us, but marketing and packaging do not always help.

- Text mining or text analytics (TM or TA)
- Concept and entity extraction (C&E Extract)
- Concept analysis (ConceptA)
- Natural language processing (NLP) (Also used for semantic search)
- Federation (Federate)
- Auto-categorization (Auto-cat)
- Data normalization (Normalize)
- Sentiment analysis (Sentiment)
- Building and maintaining vocabularies (taxonomy/thesaurus/ontology) (Vocab)

Note: Some of the companies have more extensive offering, not explicitly in the semantic software space. Categories for those other applications are not listed.

Companies Offering Semantic Software Technology Products for the Enterprise

Differentiating companies with a product or products based on semantic processing from those who offer tools for developing semantic products is somewhat subjective. Those in Table 4 have established themselves as having a commercial viable, packaged product and a number of customers under support contracts for a year or more. Undoubtedly, some will be acquired or will close their operations in the next year or two. Others from Table 5 will move into this category, or may already have done so. The distinctions are minor in many cases. We suggest that there are many circumstances for buyers to consider products in both Tables 4 and 5 when they are trying to solve a particular business problem. Opportunities for successful solutions exist in both lists.

Table 4. Companies Offering Semantic Software Technology Products for the Enterprise

Company/Product	URL	Principal Categories
ai-one	http://www.ai-one.com	C&E Extract | ConceptA
Ariadne	http://www.ariadnegenomics.com	TM or TA | NLP
Attensity	http://www.attensity.com	TM or TA | Sentiment | Auto-cat
Attivio	http://www.attivio.com	C&E Extract | Sentiment | Auto-cat
Autonomy	http://www.interwoven.com/components/pagenext.jsp?topic=PRODUCT::METATAGGER	TM or TA | Auto-cat
Basis Technology	http://www.basistech.com	TM or TA | NLP
Bitext	http://www.bitext.com	NLP
Brainware	http://www.brainware.com	Auto-cat
Cambridge Semantics	http://www.cambridgesemantics.com	C&E Extract | Federate | Normalize
Cerebra Inc.	http://www.cerebra.com	TM or TA | NLP
ChartSearch	http://www.chartsearch.net/devel/index.php	TM or TA | ConceptA | Auto-cat | NLP
Clarabridge	http://www.clarabridge.com	TM or TA | NLP |Sentiment | Auto-cat
ClearForest (Reuters)	http://www.clearforest.com/solutions.html	TM or TA | NLP | Auto-cat
Clearwell Systems	http://www.clearwellsystems.com/electronic-discovery-products/index.php	Federate | Normalize
COGITO	See: Expert System	
Cognition	http://www.cognition.com	ConceptA | NLP | Auto-cat
Collexis (Elsevier)	http://www.collexis.com	C&E Extract
Collibra	http://www.collibra.com	Normalize
Concept Searching	http://www.conceptsearching.com/web	ConceptA | NLP | Auto-cat
Connotate	http://www.connotate.com	TM or TA | Auto-cat
Endeca	http://endeca.com	C&E Extract | NLP | Auto-cat
EntropySoft	http://www.entropysoft.net/cms/home	Federate | C&E Extract
Exalead (Dassault Systems)	http://corporate.exalead.com/enterprise/l=en	ConceptA | NLP | Auto-cat
Expert System	http://www.expertsystem.net/?lang=1	ConceptA | NLP | Auto-cat | Sentiment

Company/Product	URL	Principal Categories
I2E	See: Linguamatics	
Inbenta	http://www.inbenta.com/index.php/en	NLP
ISYS	http://www.isys-search.com	C&E Extract | Auto-cat
Lexalytics	http://www.lexalytics.com/index.php	TM or TA | NLP | Sentiment
Linguamatics	http://www.linguamatics.com	TM or TA | ConceptA | NLP
Luxid	See: Temis	
Metatomix	http://www.metatomix.com	C&E Extract | ConceptA | Auto-cat
Microsoft	http://www.powerset.com	NLP
MindServer	See: Recommind	
Mondeca	http://www.mondeca.com	Vocab
MuseGlobal	http://www.museglobal.com	Federate | Normalize
Netbreeze	http://www.netbreeze.ch	NLP
NetWeaver	See: SAP	
Nstein (OpenText)	http://www.nstein.com/en	TM or TA | C&E Extract | Auto-cat
OneCalais	See: ClearForest	
Ontoprise	http://www.ontoprise.de/en/home	ConceptA | Vocab
Ontos	http://www.ontos.com/o_eng/index.php#hframe2.33063220147531	TM or TA | C&E Extract
Recommind	http://www.recommind.com	C&E Extract | Auto-cat
RiverGlass	http://www.riverglassinc.com/index.php	NLP
Rosette Linguistics Platform	See: Basis Technology	
Sandpiper Software	http://www.sandsoft.com	Vocab
SAS (Teragram)	http://www.sas.com/text-analytics/index.html	TM or TA | Sentiment | Auto-cat
Semantra	http://www.semantra.com	NLP
Semaphore	See: Smartlogic	
Sinequa	http://www.sinequa.com/index.html	ConceptA | NLP | Auto-cat
Smartlogic	http://www.smartlogic.com	ConceptA | Vocab | Auto-cat
Temis	http://www.temis.com	TM or TA | ConceptA | Auto-cat
Teragram	See: SAS	
XBS	http://www.xsb.com	C&E Extract | Vocab
ZyLAB	http://www.zylab.com	TM or TA | C&E Extract | NLP

Source: Outsell, Inc.

©2010 Outsell, Inc. Reproduction strictly prohibited.

Semantic Platform or Component Developers and System Integrators

As noted in the introduction to Table 4, it is likely that readers seeking solutions for a particular business challenge may find that a packaged product, already deployed in the marketplace, may not be ideal. This list contains many products that address specific semantic issues and, with services from the developer, their partners, or a third-party semantic technology integrator, will be a better choice. Many on this list may already be close to qualifying for Table 4.

This group also contains several companies whose focus is building semantic middleware, for content enhancement, and those servicing semantic website searching. Only by doing a thorough analysis of the problem to be solved and then having honest discussions with a vendor about how their tools and services can address that need will the reader be able to decide whether further evaluation is appropriate.

Table 5. Semantic Platform or Component Developers and System Integrators

Company/Product	URL	Notes
Adaptive Semantics	http://adaptivesemantics.com	Sentiment analysis software developers
Aduna	http://www.aduna-software.com	Semantic tool integrators
Amtera Semantic Systems	http://www.amtera.com.br/index.html	Semantic search platform
ATG	http://www.atg.com/en/products/commerce_search.jhtml	Web search engine platform
CheckMi	http://www.checkmi.com/index.html	Ontology management and normalization services
Clarkparsia	http://clarkparsia.com	Semantic software tool developers
Ctrl	See: Pragmatech	
Cycorp	http://www.cyc.com	Ontology development
dbMotion	http://www.dbmotion.com	Platform support (healthcare)
Digital Harbor	http://www.dharbor.com/indexChange.html	Semantic system integrators
Documill, Inc.	http://www.documill.com/en	Visual search
Expressor	http://www.expressor-software.com	Semantic software integrators
Franz Inc.	http://www.franz.com	Semantic software tool developers
Health Language	http://www.healthlanguage.com	Vocabulary management
Infolution	http://www.infolution.com	Semantic search platform developers
Information Extraction Systems	http://www.infoextract.com	Entity extraction and NLP
Intellidimension	http://www.intellidimension.com	Semantic development tools
Intelligenx	http://www.intelligenx.com	Entity extraction and metadata management software
IQser	http://www.iqser.ch/home1	Semantic search platform developers
Jarg Corporation	http://www.jarg.com	Semantic indexing
Knowledge Based Systems Inc.	http://www.kbsi.com/Capabilities/Semantic.htm	Vocabulary management
Lingway	http://www.lingway.com/content/view/27/249/lang,en	Semantic platform developers

Company/Product	URL	Notes
MetaCarta	http://www.metacarta.com	Semantic search engine (geographic focus) with NLP
nexTier Networks	http://www.nextiernetworks.com	Security solutions
Ontotext	http://www.ontotext.com	Semantic technology development
OpenCalais	See: Thomson Reuters	
Oracle	http://www.oracle.com/technology/tech/semantic_technologies/index.html	Database support for semantic integration
Orbis Technologies	http://www.orbistechnologies.com/index.html	Semantic software services
Orcatec	http://www.orcatec.com	Semantic software developers
Patterns & Predictions	http://www.patternsandpredictions.com/poulin/product/centiment.shtml	Data mining and Sentiment analysis
Pertimm	http://www.pertimm.com/en	Semantic search platform developers
Pragmatech	http://www.pragma-tech.com	Semantic analysis support
Progress Software	http://web.progress.com/en/Product-Capabilities/semantic-integration.shtml.en	Database support for semantic integration
Project10X	http://www.project10x.com	Semantic technology consulting
punkt.netServices	http://en.punkt.at	Semantic tools integrators
PureDiscovery	http://www.purediscovery.com	Semantic search engine platform
Raytheon BBN Technologies	http://www.bbn.com/technology/knowledge/semantic_web_applications	Semantic software developers
Rebholz- Schuhmann Group	http://www.ebi.ac.uk/Rebholz	Concept and entity extraction
Revelytix	http://www.revelytix.com	Semantic platform developers
Saltlux	http://saltlux.com/en	Semantic software integrators and Platform Services
Schemalogic	http://www.schemalogic.com	Metadata management | Vocabulary management
Semantic Arts	http://semanticarts.com/Default.aspx?tabid=2158	Semantic software services
Semantic Designs	http://www.semdesigns.com/Company	Semantic system integrators
Talis	http://www.talis.com	Semantic web platform
TextWise	http://www.textwise.com	Concept and entity extraction
Thomson Reuters	http://www.opencalais.com	Semantic tools developers
TopQuadrant	http://topquadrant.com/index.html	Ontology development | Semantic platform services
WAND	http://www.wandinc.com	Vocabulary management
Zepheira	http://zepheira.com	Semantic software integrators

Source: Outsell, Inc.

©2010 Outsell, Inc. Reproduction strictly prohibited.

Web Semantic Search Engines

While the body of our study has focused on enterprise semantic software technologies, this market is clearly driven by expectations for a truly semantic web. Hundreds of sites on the internet already make use of semantic software technologies to provide semantically more relevant search results, including products from companies in Tables 4 and 5. Among them are scores of publishers, life sciences and healthcare sites, and e-commerce operations. Users encounter the benefits every time they encounter a site that is complex and yet the results are surprisingly accurate in response to their queries. The reason is probably some linguistic processing and superior metadata vocabulary management.

The scope of this study does not permit inclusion of all the cases that might illustrate semantic processing. Furthermore, the traditional web search engines (e.g., Google, Yahoo!, and AOL) have been adding semantic processing layers to their engines for years. Readers can detect many new functions and features that have improved relevance and these are probably driven by application of semantic background processing. Table 6 is a list of web search engines that were launched to answer questions using either natural language processing or auto-categorization as underlying technologies. Because they are all public, readers can test-drive and experience for themselves what happens when they ask a question versus typing in keywords. There is a learning process to posing queries effectively, and these are good test-beds.

Table 6. Web Semantic Search Engines

Company/Product	URL	Categories
ASK	http://www.ask.com	Semantic web search engine
Bizo Inc.	http://www.bizo.com	Web marketing platform
Clusty	http://www.clusty.com	Meta-search with auto-categorization
EasyAsk	http://www.easyask.com	Semantic web search engine
Evri	http://corporate.evri.com/solutions	Semantic search platform
Hakia	http://www.Hakia.com	Semantic web search engine
Kosmix	http://www.kosmix.com	Semantic web search engine | Federation
Orcatec	http://www.truevert.com	Semantic web search engine
Microsoft	http://www.bing.com	Semantic web search engine
Semantifi	http://www.semantifi.com	Semantic web search engine
SenseBot	http://www.sensebot.net	Semantic web search engine
Thomson Reuters	http://www.opencalais.com	Semantic web search engine
TrueKnowledge	http://www.trueknowledge.com	Semantic web search engine
Truevert	See: Oractec	
UKPMC	http://ukpmc.ac.uk/classic	UK PubMed Central

Source: Outsell, Inc.

©2010 Outsell, Inc. Reproduction strictly prohibited.

Standards, Governmental Consortia, and Industry Organizations

ANSI and ISO (national and international) standards for semantic technologies are sparse and selective, peripherally related to vocabulary management (thesaurus), indexing guidelines, and data interchanges. Table 7 provides a link to the ANSI/ISO store; it is best searched by keywords. The remainder of the sites are representative of the communities of practice and organizations formed to help place guidelines and industry standards around software with semantic roots or relationships. With so much diversity and innovation, the industry is many years from establishing uniform standards.

Table 7. Standards, Governmental Consortia, and Industry Organizations

Company/Product	URL	Categories
ANSI/ISO	http://webstore.ansi.org	US National Standards and International Standards Organizations
Apache UIMA	http://uima.apache.org	Unstructured Information Management applications
CALBC	http://www.calbc.eu	Collaborative Annotation of a Large-Scale Biomedical Corpus
DERI International	http://www.deri.org	Semantic Web Research Institute
DITA/OASIS	http://dita.xml.org/standard	Darwin Information Typing Architecture (Metadata)
GRO	http://www.ebi.ac.uk/Rebholz-srv/GRO/GRO.html	Gene Regulation Ontology
Health Care and Life Sciences Interest Group	http://esw.w3.org/HCLSIG	Subset of W3C for healthcare interest
IAOA	http://www.iaoa.org	International Association for Ontology and its Applications
lotico: The New York Semantic Web Meetup	http://semweb.meetup.com/25	Subset of W3C regional community
NIEM	http://www.niem.gov	National Information Exchange Model (Departments of Justice and Homeland Security)
OMG	http://www.omg.org	Object Management Group
OWL	http://www.w3.org/TR/owl2-overview	Web ontology language
RDF	http://www.w3.org/TR/rdf-primer	Resource Description Framework
Semantic Universe	http://www.semanticuniverse.com	Independent conferences and publishing group on semantic technologies
SPARQL	http://www.w3.org/TR/rdf-sparql-query	Syntax and semantics of the SPARQL query language for RDF
W3C	http://www.w3.org	International community developing and promoting standards for semantic web

Source: Outsell, Inc.

©2010 Outsell, Inc. Reproduction strictly prohibited.

Bibliography for Semantic Software Technologies – 2010

The following citations are organized by the broad topics for which the reader may want to gain more understanding. Links to articles, blogs, and websites were valid as of 06/07/2010. Of course there is the possibility that by the time you read this they may have changed. Try copying the title (in quotation marks) into your favorite search engine interface to see if the content is located elsewhere. Books link to a description and source for purchasing.

Linked authors have a significant role in the field or have written other works that you might find interesting.

Articles, Papers, and Web Postings on Semantic Technologies

Anicic, Nenad. ***An Architecture for Semantic Enterprise Application Integration Standards***, by Nenad Anicic, Nenad Ivezic, and Albert Jones. (Faculty of Organization Sciences, Belgrade and NIST). 2006 10p. [IN: Interoperability of Enterprise Software and Applications, Springer, London, 2006, pp. 25-34].

Arnold, Stephen E. ***Exclusive Interview with David Milward, CTO, Linguamatics***, 1p. Beyond Search, 02/16/2009.

Brindley, Lynne. ***Pioneering research shows 'Google Generation' is a myth***, research by a panel consisting of, Lynne Brindley DBE, (Chief Executive) British Library, Lord Triesman, Dr Malcolm Read, Ian Rowlands, … 35p. CIBER at UCL, 01/11/2008.

Cane, Alan. ***New techniques find meaning in words***, 2p. FT.com/Financial Times, 10/08/2008.

Chakravarthy, Anil S. ***Sense Disambiguation Using Semantic Relations and Adjacency Information***, from the Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Cambridge, Massachusetts, 1995, pp. 293-295.

Chickowski, Erica. ***Understanding Semantic Web Technologies***, 5p. Baseline, 08/05/2008.

Daume III, Hal. ***NAACL-HLT 2009 Retrospective,*** 3p. Natural language processing blog, 06/12/2009. Read more at: http://clear.colorado.edu/NAACLHLT2009/.

Erickson, Jonathan. ***Semantic Integration: Meeting The Challenge***, [Interview with Richard Keller, senior research computer scientist and group lead for the information sharing and integration group at NASA]. 1p. InformationWeek: Dr. Dobb's, 10/24/2009.

Gibbon, Dafydd. ***How to Make a Dictionary – Class Notes*** (word forms, morphology, lexicography, etc.), 2005-2006. University of Bielefeld.

Grimes, Seth. ***Breakthrough Analysis: Two + Nine Types of Semantic Search*, **Intelligent Enterprise, 01/21/2010.

Heires, Katherine. ***For Wall Street, a Matter of Semantics***, 3p. SecuritiesIndustry.com, 02/25/2008.

Kho, Nancy Davis. ***Customer experience and sentiment analysis***, 3p. KMWorld, 02/01/2010.

Marks, Oliver. ***The Semantic Enterprise***, 1p. ZDNet Collaboration 2.0, 01/04/2009.

McComb, Dave. ***CIO's Guide to Semantics***, v3. 16p. 05/27/2010.

McCreary, Dan. ***Entity Extraction and the Semantic Web*,** SemanticUniverse, 01/01/2009.

Miller, Paul. ***Semantic Technology's place in the enterprise; key, but low-key?***, 1p. The Semantic Web, 09/29/2008.

Miller, Paul. ***New Report Places Semantic Web 'On the Cusp' of Something Big***, 1p. The Semantic Web, 09/30/2008.

Miller, Ron. ***Semantic Search takes root in the enterprise***, 4p. InfoToday: Enterprise Search Sourcebook, 09/01/2008.

Monash, Curt. ***Where "semantic" technology is or isn't important***, Text Technologies, 12/29/2008.

Painter, Joshua. ***Semantic normalization: making sense out of health data***, Intel Software Network, 09/23/2008.

Radhakrishnan, Arun. ***9 Semantic Search Engines That Will Change the World of Search***, Search Engine Journal. 04/19/2009.

Rebholz-Schuhmann, Dietrich. ***Semantic Standardisation of the Scientific Literature***, The Rebholz-Schuhmann Group, 2009.

Shaw, Tony. ***SemTech 2009 Post-Event Media, Blogs, and Trip Report Items SemTech Universe***, 07/09/2009. Links to the following: Semantic Universe announces that videos and audio recordings from the Semantic Technology Conference are now available in its "SemTech 2009 Highlights" editorial issue.

Sheth, Amit. ***Semantic Meta Data for Enterprise Information Integration***, Information Management Magazine, July 2003 2p.

Uszkoreit, Hans. ***What is Computational Linguistics?***, 2p. University of Saarland, Germany, 2003.

Weinberger, David. ***The dream of the Semantic Web***, 2p. KMWorld, 03/01/2009.

Slide Presentations

Boeri, Bob. ***Improving Findability Behind the Firewall***, 28 slides. Enterprise Search Summit 2010, NY, 05/2010.

Dahlgren PhD, Kathleen. ***The Puzzle of Semantic Technologies***, 24 slides. Infonortics/Search Engine Meeting, 04/01/2009.

Doszkocs, Tamas. ***Semantic Search Engines for Consumer Health***, 29 slides. Search Engine Meeting (Information Today), 04/26/2010.

Rappaport, Avi. ***Federated vs. Aggregated Search Architectures***, 19 slides. Enterprise Search Summit 2010, NY, 05/2010.

Soubbotin, Dmitri. ***The Variety of Goals and Applications of Semantic Approach to Search***, 20 slides. Semantic Engines, (sic)New York (Boston), 2009.

Reference Resources

Association for Computational Linguistics (association)

Computational Linguistics (journal)

Encyclopedia of Linguistics

Guarino, N., ed. ***Applied Ontology, An Interdisciplinary Journal of Ontological Analysis and Conceptual Modeling***, edited by N. Guarino and M.A. Musen. IOS Press, Amsterdam.

Hedden, Heather. ***Accidental Taxonomist***, May, 2010. Information Today. ISBN 978-1-57387-397-0. 472p.

IAOA. ***The International Association for Ontology and its Applications, Ontolog virtual panel,*** June 18th, 2009, [chaired by] Nicola. Guarino, ISTC-CNR, Laboratory for Applied Ontology, Trento, Italy, 41 slides, IAOA, 06/18/2009.

King, Brandy. ***Finding the Concept, Not Just the Word: A librarian's guide to ontologies and semantics***, by Brandy King and Kathy Reinold. Chandos Publishing, 09/01/2008, ISBN: 1843343193. 202p.

Lebeth, Kai. ***Semantic Networks in a Knowledge Management Portal***, Springer Berlin / Heidelberg. 2001. ISBN: 978-3-540-42612-7.

Semantic Exchange [Website dedicated to the Semantic Web]

Semantic Universe [Website dedicated to Semantic Applications and Technologies]

Vendor Deep Dives

COGNITION GIVING TECHNOLOGIES NEW MEANING™

Cognition

Representative Customer Insights

We use Cognition to gather information in the biomedical domain and it is indispensible for doing my research. Specificity is its strength – the synonymy of common verbs and objects is excellent.

Cognition has top quality people with great attention to detail. There is a legacy of building the Lexicon by following a clear decision-making process that accounts for the quality of the dictionary and concept relationships.

History

Cognition Technologies has been in business since February of 2003, a successor to Intelligent Text Processing (ITP). Kathleen Dahlgren, Ph.D., Cognition's founder and CTO is the connection between these companies. She brought to both companies six years of experience at IBM where she led a research team to create the first prototype of a Natural Language Understanding System.

Cognition's Semantic Natural Language Processing (NLP) technology was developed by a research group of over 30 professionals, many of them lexicographers, and computational linguists led by Dr. Kathleen Wallace, Dr. Daniel Albro, and Dr. Brian Potter. Development was funded by IBM, the US Army, and investors including Southern California Tech Coast Angels. Cognition is currently self-sustaining through software licensing revenue.

Based in part on ITP work funded by an SBIR award that resulted in a patent in 1998, Cognition's Semantic NLP was launched in 2004 as a commercial prototype. From that core a suite of applications and demonstration sites were established.

Cognition is focused on providing its semantic natural language processing technology to software developers and companies building applications that need deep linguistic text processing capabilities. The company has business relationships with the Powerset division of Microsoft, Merrill Corporation, and University of Texas Medical School. A dozen development tools have been packaged to support semantic text mining, text analytics, categorization, and search applications.

The experience and expertise of the Cognition staff provides high-level professional services for any enterprise requiring linguistic processing support. Cognition works with companies to improve their e-commerce and business presence on the web by enhancing retrieval precision and recall in any application.

Cognition's toolset, in both scope and depth, contributes to its emerging leadership position in the semantic software market as a partner for building both enterprise solution applications and better semantic web searching. Cognition's semantic map alone, painstakingly crafted with advanced grammatical rules, contains over ½ million base word forms and provides a much needed jumping off point for developers. Using it to build value-added software applications, suppliers to the marketplace will deliver more precision in search results with better recall using natural language processing. Cognition NLP is a building block for Web 3.0.

There is also recognition in the technical press of Cognition's leadership in the semantic net development arena. Representative comments are found in these write-ups: Cognition Releases the Largest Semantic Map of English Language; Cognition Is Hard at Work Building the Semantic Web; Cognition.com - Semantic Searches. Cognition is based in Culver City, California.

Descriptions of the Offerings

Cognition's Semantic NLP combines algorithms and process from both statistical and symbolic NLP. The applications and software tools offered by Cognition to developers of web products and services and enterprises seeking to improve internal retrieval are designed to leverage language with its millions of nuanced relationships and rich (but not always common) expressions.

Cognition's Semantic NLP and associated products are packaged to support development of other products. As well, Cognition has the deep expertise to work with developers and implementers addressing semantic problems on a consulting basis. Having devoted a couple of decades to establishing the Cognition tool framework, these experts can make a substantive difference in defining a project or product for success.

Cognition semantic products fall into three major categories (search, categorization, and special applications), plus an application programming interface (API) and services.

Semantic Search

Searching in an NLP framework requires substantial content processing paired with Cognition's linguistic management architecture to match the meaning of the inquiry (search request) with accurate and complete content results. Cognition's Lexicon (semantic net) delivers full value with the tools in this architecture that are designed to fully leverage its complexities. This is the suite of search products:

- *The Cognition Parser* assigns grammatical structure to sentences to discover concepts defined in the governing Lexicon (the dictionary, ontology, meaning contexts, and meaning thesaurus). It operates on these document types: HTML, XML, plain text, Word, WordPerfect, RTF, PDF, Power Point, and other common document formats.

- *The Cognition Indexer* creates searchable "indices" of concepts extracted by Cognition Parser, reading each sentence, phrase, and word before assigning meaning to words based on context. Meaning attributes (up to 15 per word) are associated directly with each word in the indices. Examples of meaning attributes are "cats have tails", "cats are felines", a synonym of "cat" is "kitty", etc. Due to aggressive use of compression algorithms, indices are typically half the size of the data indexed, despite having many pieces of information for each token. Indexing speed is linear with the number of files indexed.
- *CognitionSearch* controls access to indexed data using a variety of search argument syntaxes and methods: complex natural language queries, standard Boolean, advanced "Linguistic Boolean," fuzzy, pattern, and Soundex name search queries. It is supplied with functions to support relevance ranking, search term highlighting and linking navigation. Retrievals are sub-second on a terabyte index.
- *The Cognition Broker* is a network performance optimization tool for the Semantic Search engine software in a large, multi-server environment.
- *The Cognition Spider* supports the integration of external data with internal repositories by crawling the web or remote sites.
- *The Cognition Interface* is applied to starting up CognitionSearch by providing a default interface for searchers. It delivers all the expected functions including query delivery to the search engine, and displaying results with highlighting of words and phrases in the retrieved documents. It aids term disambiguation by displaying the sense and definition of a term chosen by CognitionSearch beside the alternatives, supporting changes in word senses from the user, and displaying spell-checking information with a method for accepting spelling choices from the user.
- *The Cognition Ranker* improves the quality of machine translation from foreign languages to English. This software produces hypothesis translations in a ranked order. The Cognition Ranker parses the hypotheses, and ranks them for semantic and syntactic plausibility. Similarly, it can rank statistically-based speech recognition software output (for English) for semantic and syntactic plausibility to improve final choice of interpretation for speech.

Concept Driven Topic Discovery and Document Categorization

The success of search can be dramatically improved by providing enhanced content categorization, which in turn benefits from concept discovery. Cognition offers tools to automatically identify key document content concepts directly from concepts in Cognition's Semantic Map or mediated by a third party topic/vocabulary resource.

- *Cognition Topic Discovery* makes a document-by-document analysis of all concepts to formulate a concept map. It then calculates a salience score (how important is this concept to this document) for each concept in each document. Used for evaluating content across a project or case, Cognition Topic Discovery is delivered with reports that can be configured to show all concepts ranked by salience score, or just those scoring above some predetermined threshold. Leveraging Cognition's Semantic Map, the module uses word groupings: words related via synonymy (e.g., "deed" and "title") OR ontology (e.g., "heart" as a "circulatory organ") to determine a concept's saliency score.

- *Cognition Categorization* is similar to Cognition Topic Discovery, scoring each document with respect to a predefined set of topics using either vocabulary in Cognition's Semantic Map or a third party resource.

Search Results Applications

Delivering high precision, high recall in a form that is immediately understood by users is the ultimate determinant of search success. Cognition's product portfolio includes search output applications, each with specific market focus.

- Cognition Foldering applies concept analysis and categorization to a special domain (e.g., a document management system corpus specific to a particular legal case) in which the issues and documents are most likely to be relevant. For example, assume a review of all material documents that could reveal the relationships between the U.S. government agency for Minerals Management and BP. The application assumes a preliminary review of case issues that identifies the most relevant complex (multi-concept) topics that must be isolated for final review from millions of documents that have minimal relevance. Relevant documents are placed into multiple folders based on processing queries, each expressing a complex Boolean construct. One construct, or "foldering" query might be (*Minerals Management Service personnel* AND *oil industry* AND *permit approvals*) while a second could be (*BP* AND *Offshore Energy and Minerals Management* AND *regulatory oversight*) and so on.

  Each part of the Boolean is interpreted linguistically and automatically enhanced with alternative expressions contributed by the *Cognition Lexicon* plus unique case-specific language (e.g., MMS synonym for Minerals Management Service). Finally, with all documents irrelevant to this particular case removed, the remaining documents are placed into folders for each query construct, per the conceptual project framework.

Two products provide support for e-commerce web-based applications. They identify and rank documents containing text similar in conceptual content to a target query or contextual content.

- *The Cognition Ad Matcher* leverages semantic analysis to match concepts (phrases) found in advertising copy to relevant media content, optimizing the similarity between an ad and companion stories in a publication.
- *Cognition More Like This* analyzes and interprets internal or external documents within a particular domain for conceptual similarity.

Many business purposes drive the need for statistical information about a content domain. Whether determining the entity types and their relationships or evaluating the conceptual mapping of topics that are present, text analytics give support to leveraging content.

When high-level semantics are applied to text analysis, accuracy, and value of the output is increased. This is the purpose of the *Cognition Text Analytics* application. This application leverages Cognition's Semantic NLP to generate a collection of reports that present statistics from a content corpus. Data sets (words, stems, senses, phrases, concepts) are reported for their frequency, significance, and salience. The application is an embeddable function that can be used to enhance Cognition Interface for Cognition Searcher.

Cognition *Professional Tools* and *Consulting Services* are important for tailoring the Semantic Map to the linguistics of a specialized project.

- *Cognition Customization* includes text analysis tools, semi-automated lexical acquisition (a professional service), and client-controlled lexical customization. These tools and services apply to the *Cognition* Lexicon, comprised of a Lexical Dictionary, Taxonomy, Word Meaning Context database, and Word Meaning thesaurus. Collectively these include millions of word forms, proper nouns, concepts, concept groups, and relationships/linkages between them. Maximizing the linguistic power of Cognition's NLP is the intent of their customization services.
- *The Cognition API* is the application that supports advanced integration for Cognition's Semantic NLP software. This is also used by enterprise customers who have an internally developed data repository or content management system. The Application Programming Interfaces (API) for central components of the software includes Search, Index, and Review functionality. Sample scripts and API libraries are also available for C++, Python, Perl, Ruby, Java, VB/ASP/ActiveX, and PHP4.

Strengths

In 1998, Cognition Technologies (through its predecessor company) was awarded a patent (USP 5,794,050) entitled, *A Natural Language Understanding System*. This seminal patent includes 30 claims. The patent describes a linguistic method for controlling the explosion of potential parses (syntactic structures) to reduce parsing time exponentially. Most computational parsers are noted for their lack of scalability as the size of a corpus reaches millions of documents.

As explained by Cognition, "this patent deploys common sense reasoning and naive semantics to avoid fruitless paths and back-tracking in the computational parser. It would prevent, for example, a computational parser from trying to build up a parse of *The woman on the rock cried*, in which *the rock cried* is a sentence level part. It prevents this fruitless path by noticing that it is semantically implausible for a "rock to cry." It is unique in its linguistic algorithms that exhaustively apply the rules of grammar with rich semantic representations.

Cognition's Semantic NLP makes technologies and applications more human-like in their understanding of language, thereby resulting in more robust applications, greater user satisfaction, and new NLP capabilities available for exploitation. For any organization planning to deliver Web 3.0 applications, the suite of tools that Cognition has built to embed and enable NLP will be a good place to start.

High precision and high recall are the best performance indicators of Cognition Searcher. *Precision* is the proportion of all relevant documents in the retrieved set, while *recall* is the proportion of all the desired (or relevant) documents in the corpus that were actually retrieved. This means not missing relevant documents and, simultaneously, not retrieving irrelevant documents. Cognition Searcher overcomes the problems with other retrieval systems that are based on pattern-matching or statistical relevance or hybrids of the two.

Cognition Foldering exceeds conventional clustering methods by eliminating irrelevant documents from conceptual groupings. Culling out these documents reduces manual review time and cost, typically by 50%.

Cognition More Like This is based on *conceptual* similarities among documents rather than straight similarity of the strings of text in the documents.

Cognition's conceptual linguistic processing is performing at the time of indexing; the result is the speed of Cognition Searcher because the computationally intensive reasoning is already done.

Cognition products put developers in control over their own application product possibilities, allowing them significant flexibility, whether managing indexing performance using the Cognition Broker to direct the set-up of distribution configuration across servers or using Cognition Customization and Cognition API to design and integrate a new product in a new industry.

Problems Solved

The Powerset division of Microsoft has licensed Cognition's Semantic Map for its dictionary and grammar rules, comprehensive term *disambiguation* and correlation of *concept relationships*. The application is Bing and the licensing arrangement provides for updates to data in the Semantic Lexicon so that new terminology is regularly delivered for sustained high retrieval relevancy.

For Merrill Corporation Legal Solutions, Cognition's Semantic NLP provides *conceptually precise and comprehensive document retrieval, categorization, and organization* based on specific case requirements. It culls irrelevant documents from the organized results, the benefit being significant reduction in time to process case documents with improved accuracy in results. In one recent legal case Cognition Foldering reduced the e-discovery documents by over 50%. Merrill Legal solutions is available to clients as a hosted service.

Semantic MEDLINE, providing access to 18 million National Library of Medicine abstracts, was released in 2008. It is powered by Cognition's Semantic NLP and is free to the public.

Using Cognition Broker, overhead requirements can be improved through its automatic load balancing and automatic discovery of network layouts and addition of any new server. It also provides fault-tolerance by automatically compensating in case of server failure and allows you to start or stop servers at any time without modifying any configuration files.

Cognition's Semantic NLP dictionary is constantly updated with new language or neologisms. Support is provided for users to add categories of nouns in a special file, new nouns in classes, such as lists of kinds of video recorders, motorcycles, or roses. The addition of more complicated special vocabulary is introduced into Cognition's Semantic NLP dictionary as a fee-based service.

Cognition's NLP query interface provides a "View Concepts" window to give searchers an opportunity to select an alternative meaning from a list of drop-down choices if the initial Cognition meaning selection is not the one intended by the searcher. Additionally, advanced searchers can apply Boolean constructs using *AND, OR, WITH, NOT WITH, and AND NOT*. For example, the elements between the Boolean operators are treated as conceptual queries. If you ask: "(strike in baseball) AND NOT (walk in baseball)" – it will search on "strike", meaning a state of the game baseball, rather than meaning "hit" or "labor walkout" and will exclude the concept of "walk" meaning a state of the game of baseball rather than "walk" meaning "to put one foot in front of the other."

Strategic Advantages and Competitive Positioning

Cognition has entered the competitive semantic software technology marketplace, relatively recently, to capitalize on its original patented linguistic parsing technology (USP 5,794,050) and more recent provisional filing (US Application No. 20,070,106,499) for "handling of a linguistic dictionary." Cognition has taken serious steps by securing investments that enabled packaging its technologies in commercial applications for this complex market.

In just two years, Cognition has gained a toehold in three significant markets: legal, health sciences, and semantic web searching, all using Cognition's Semantic NLP. These are three areas where precision and completeness of results are required. The enormous quantities of information needing to be processed quickly and accurately have rendered inadequate and insufficient conventional string searches, standard Boolean and statistically-based relevancy ranking methods. The type of innovation established by Cognition in linguistic processing has the potential to supplant legacy methods of search.

Cognition has a reputation among experts in the semantic technology community who know and understand the substantive work that has gone into building their semantic net, the Cognition Lexicon. Through speaking engagements and consulting services to companies just beginning to leverage semantic techniques and tools, they are positioning themselves as thought-leaders on how to dramatically improve natural language processing. And they have the tools to embed their technology in numerous applications.

Cognition has arrived on the scene with the competency and professional expertise to educate, train, and work side-by-side with companies ready to embrace semantic search. Whether companies want to make existing products semantically rich or to create new products that are semantically state-of-the art, innovators who are knowledgeable in this field will select Cognition as a logical partner.

Futures

Cognition says, "We look at what we're doing as a significant component to the Semantic Web. The focus is on semantically enhancing other technologies. Cognition is building a solid leadership team with strong experience in building businesses coupled with deep technical expertise in a very challenging arena. Investors have taken notice and responded with funding."

Even in a period of business stresses, Cognition is gaining traction. Kathleen Dahlgren has a solid reputation in the field of linguistics as her publications and development work attest. She inspires a level of confidence not only for her insights into how to make linguistics work correctly, but other professionals find her practical and collegial approach to problem solving refreshing. Her leadership on the development of tools and on consulting engagements is noteworthy.

Customer Testimonials

A weakness of most search engines is that they are not very good at finding information that is unique and rarely looked for. Cognition solves that problem.

We tried working with some other statistically-based products for categorizing documents and they just did not give good results. We always have to deal with new topics and different queries. Cognition helps us with new terminology and makes it easy to work through the additions to the vocabulary.

Our clients work in a very fast-paced environment; everything has to be done quickly and efficiently. We are able to deliver excellent results without overwhelming them with an overly complex process to get there, using Cognition to mine for the concepts they need.

I have used other semantic nets and they are ragged and inconsistent. Cognition has superior disambiguation techniques and the rules of grammar are applied in a totally consistent manner.

The tools for doing language acquisition and building up the semantic net are very efficient. I have worked on it in collaboration with Cognition and found that they have a very straightforward approach to updating the Lexicon. The structure is great and fast, and the tool for updating is terrific.

People do not understand how much routine work is needed to solve hard science problems but it is essential to begin by finding out what is already "out there." Cognition supports that kind of exploration and research through tons of literature, and you can be confident about the results.

Corporate Facts

Corporate Headquarters: 6133 Bristol Parkway, Suite 350, Culver City, CA 90230

Sales and Support Contact Information: Stephen Lief 646-330-9918

Officers: Bill Collins, Chairman, Kathleen Dahlgren, PhD (CTO and Founder), Stephen Lief, Director of Sales, Daniel Albro, PhD (Chief Scientist).

Status: The Company is privately held. It has been funded to date by investments from individuals, angel investors, and venture capital firms.

EXPERT
SYSTEM
SEMANTIC INTELLIGENCE

Expert System

Representative Customer Insights

The Cogito platform has enabled us to speed up our competitive intelligence filtering processes to capture "weak signals" that indicate an impending change in our marketplace. These present opportunities for us to move quickly and adjust our business strategy before the competition.

We selected Expert System because we already knew that we wanted, true semantic technology and not just text analysis with statistical inferencing.

History

Expert System builds its semantic net product, *Cogito*, on 20+ years of experience with linguistic technology. In fact, their roots reflect a foundation built on providing "spell checkers" for Microsoft Office. The relationship continues; they are the only Italian software house that is a Microsoft partner. In 2007, when Expert System announced Cogito SIMS (semantic intelligent management system for partners and OEMs), their legacy of engineering and development in language technology was paying off for the privately held and consistently profitable firm. At the end of 2009 they were able to announce revenues of +10M Euro, 100+ employees, and 27% EBITDA, a 44% annual growth.

Cogito core technology is based on a *semantic net*, Expert System's *Sensigrafo*, a graphic knowledge representation of language. It is optimized for automated language processing. Currently, Expert System offers support for the Cogito semantic network of 350,000 words and 2.8 mil relationships. This core network includes all words in English and relationships among them. Similar semantic networks are available for Italian, German, French, and Arabic. Expert System estimates that about 75% of the knowledge concepts in its English semantic net have equal representations in other languages and the other 25% have been carefully considered and added for each language to capture the nuances. Several hundred people over the years have contributed and continue to contribute to Cogito's knowledge concepts and language relationships.

Expert System has expanded its reach into over a dozen vertical markets where its rapid (weeks) deployment model, adaptability for ingesting industry and enterprise specific vocabularies, and out-of-the-box linguistic analysis capabilities help speed application implementation for clients. In comparison, other applications take a year or more and a dozen people to implement. Expert System's approach is significantly more efficient than shallow text analytics tools that apply statistics, heuristic rules, and morphological recognition to solve similar semantic problems.

The Cogito product line is packaged for a diverse range of semantic challenges. This is noteworthy because of the scope and complexity of semantic problems to be tackled and the desire of most enterprises to take on each one in a logical business sequence. Expert System has demonstrated its business savvy in recognizing the way enterprises need to operate by leveraging its rich linguistic framework, and packaging Cogito uniquely for each semantic application. Applying semantic tools in small bites is smart and logical from a business perspective.

Expert System is making serious inroads in energy, homeland security, electronics, telecommunications, automotive, finance, media, and life sciences industries. Customers see significant improvement in the semantic relevance of retrieved content using Expert System's tools and professional linguistic services for enriching their out-of-the-box semantic net with domain specific terminology.

Expert System offers targeted solutions for: categorization and classification of documents, sentiment analysis, entity extraction and facet creation, disambiguation, and NLP intelligent self-help queries. These solutions are designed to contribute semantic intelligence throughout an enterprise depending on function and business case. Customer services, competitive intelligence and faster categorization of large quantities of unstructured content for research and professional internal use each benefit from Cogito point solutions.

Expert System has been achieving awards and market positioning since 2007. Here are a few on the list: Gartner - Information Access Magic Quadrant - 2009, 2007; Search Engine Watch Awards, Most Innovation New Search Engine - 2009; Codie Awards, Best Enterprise Search - 2009

Descriptions of the Offerings

Expert System engages the marketplace with a steady program of pre-packaged applications to meet individual enterprise semantic challenges. Some find widespread adoption in specific industries and others apply to business functions in all verticals. The following characterizes individual use cases for each product.

Cogito Family of Products

Cogito Semantic Search is for generalized enterprise search across a broad general spectrum of content repositories. This search engine goes beyond traditional keyword search engines by leveraging Cogito's semantic net to find *meaning-based concepts*. Where content repositories are large and lack adequate metadata, Cogito has both relevancy and performance advantages over traditional enterprise search engines.

Cogito Categorizer supports the organization and classification of corpuses of content that exceed what humans can realistically tag. The Categorizer can be customized to include concepts and entities unique to a particular domain. Enterprises can apply internal vocabularies or thesauri, particularly in heavy research environments such as life sciences.

Cogito Discover brings a package solution to enterprises with large quantities of un-governed structured or unstructured content. Using its advanced linguistic framework, Discover extracts, transforms irregular language across data stores, tags documents, de-duplicates, and normalizes content to place it into a semantically uniform framework or database. It is ideal for merging and integrating databases in similar domains that have been built and maintained separately. After being appropriately tagged, content can be fed to the Categorizer or other applications for further analysis and tagged for more accurate retrieval.

Cogito Monitor is ideal for tracking and quickly viewing, in clear displays, the sentiment or tone of discussion about a company, product, or an industry. By applying Monitor across public websites and feeds of social sites, enterprises will immediately detect opinions and weak signals of trends that affect them directly or suggest the need for strategic action.

Cogito Focus brings web content together with enterprise content to improve the reach of a query. The product enables retrieving external and internal information and bringing results into a common view to correlate and analyze the information in a unified context. It also supports the integration of web definitions to "train" your semantic net to refine search results based on the meaning you intend. The visualization features and ability to create new facets "on-the-fly" promote highly dynamic interactions with the content and facilitate exploration of data. Recently, Focus has adopted the IPTC (International Press Telecommunications Council) classification scheme to include its ontology of categories.

Cogito Answers is an application using natural language processing (NLP) to deliver customer service activities via the web, mobile devices, or through e-mail. Answers is designed to integrate Cogito's linguistics technology and semantic net with a knowledge base of unstructured content that contains the answers. Then it can correctly interpret questions asked in plain English.

Cogito Intelligence Platform combines text mining for extraction into data bases, semantic analysis for multilingual content, detection of weak signals, automatic classification and correlation across disparate repositories, with advanced visualization tools. The target market is government agencies and competitive industries needing intelligence processing from incoming feeds and external repositories.

Cogito SIMS Semantic Intelligence Management System is the first advanced linguistic Software Development Kit (SDK) specifically designed to develop applications to find, organize, select, and correlate with accuracy a high level of quality unstructured data coming from different sources. SIMS' ability to effectively understand the meaning of terms using Cogito's semantic net streamlines the development of semantic applications that need to increase relevancy in retrieval of unstructured text. It is also a stepping stone to building applications to normalize and automatically enhance metadata – named entities, relations, and events related data trapped in a document.

SIMS SDK is directed to OEM customers and system integrators. It is simple to use and it includes a declarative language and intuitive user interfaces to enable programmers who do not possess extensive computational or linguistic experience to develop advanced search, categorization and text analysis applications

Strengths

Expert System is an Italian company and that presented challenges for gaining product traction in the English-speaking world. As already noted, the company is approaching two decades of experience in this language domain through their linguistic support for Microsoft products. But embedded software tools, even those with millions of installations have a stretch to build commercial recognition among direct buyers.

Next, Expert System launched a systematic program of product releases, beginning in 2007, coupled with new management presence in the UK and US and began their steady growth in markets outside Italy. Readers can see for themselves the record of activity by viewing their Press Release archives, not so much for the marketing messages, but as evidence of strategic awareness and action.

There are seven specific areas in which Expert System clearly demonstrates market understanding and responsiveness:

- *Packaging* – Semantic technologies are complex to explain and sell to business buyers. Taking individual business cases that they know their Cogito platform and semantic net can solve, Expert System has wrapped solutions into individual applications that speak to business challenges. It already has public stories to tell about every application.

- *Business Use Cases* – Cogito products address business cases that are on everyone's target for leveraging semantic technology. Over the past decade or two those challenges have been tackled by the largest enterprises with tools that require purchasing high priced software licenses, plus services to implement over months or years. This approach doubles or triples the original license cost. Taking a direct approach to each application and establishing reasonable paths to evaluation, testing, procurement, and quick implementation, Expert System places its products in a highly competitive position for even SMBs.
- *Customizable and Compact Semantic Net* – Cogito's core semantic net, Sensigrafo, provides a comprehensive vocabulary with a small footprint (~50 MB), which can be built up using any industry or domain specific terminology.
- *User Interfaces* – Semantic applications require tuning and administration. Cogito products are delivered to be used in real world environments in which subject, text analytics, and linguistic experts can apply their knowledge to curation and customizing the semantic net.
- *Quick Installation and Superior Performance* – Four to eight weeks is typical for implementing and testing a new corpus of content. Typical semantic speeds are in the order of 120KB/sec with a standard quadcore CPU.
- *Efficiency Gained for Teams* – Deployment and tuning is usually performed by professional linguists, language engineers and programmers. An implementation is easily supported by a team of two to four. Expect that internal terminology experts groups of up to 15 people before installation can usually be reduced to one or two.
- *Hosted Services for Applications* – Monitor, Focus, and Answers can also be offered as hosted solutions to address the growing needs for some functions in the enterprise (e.g., Marketing or Customer Care) to minimize the hassle to develop and support the architecture required to run the applications.

Problems Solved

Expert System's products offer tools for all fundamental semantic problems:

- Natural language search;
- Text mining and analysis;
- Sentiment analysis;
- Disambiguation of terminology in context;
- Auto-categorization.

An implementation of the Cogito Semantic Search Engine after installation might begin with this training process for a new corpus. This is an iterative process, making adjustments to improve semantic understanding in a specialized domain. Expert System offers services to perform the operations, or the enterprise experts can do it themselves.

Here are a couple of examples of Cogito solutions applied:

- *RCS*, the leading publishing company in Italy with businesses in press, book, news, and multi-media, uses the Categorizer to manage feeds of hundreds of articles each day. Two principal benefits are higher accuracy classifying articles and the ability to perpetually process content, 24X7X365. The Categorizer worked so reliably that RCS merged categorization output (subjects, places, people, and company entity names) with its editing system to provide a normalizing and checking function across their content. RCS has reduced operating costs, while improving quality and consistency across content.
- Cogito Monitor, which was deployed for the auto industry to monitor sentiment by scanning blog feeds, is used by *Honda* and *Pirelli.* It looks for information about customers' attitudes and product problems to aid product managers, dealers, manufacturers, and consumers.

Visit the product market and customer pages of the Expert System website for more application case studies.

Strategic Advantages and Competitive Positioning

Expert System has the linguistic depth with its semantic nets to compete in multiple languages, which gives them coverage in most of Europe and the Middle East. As well, they have demonstrated competency in competitive vertical markets.

They have lengthy and deep experience with semantic analysis. This means understanding the four cornerstones of semantics: morphology, grammar, logic, and meaning.

Sensigrafo, Expert System's semantic net, embodies some unique features that allow superior semantic analysis and disambiguation of text supporting deep text analysis and high performance (accuracy and speed). Unique features include:

- Expanded definition of lemmas (base form of a words) including all synonymy, and semantic relations;
- Categories of attributes;
- Differentiated domains of usage;
- Differentiated sets of attributes;
- Small footprint, 50MB of memory;

Each node in the semantic network has a unique ID and a proprietary way to access the data to speed performance and support in memory processing.

The company offers its own ETL (extraction, transforming, loading), converters, crawlers, and integration points with popular programs such as Microsoft FAST, SharePoint etc. to ease the use of semantic tools with existing IT investments. Expert System also integrates with many third-party ETL tools.

Expert System Semantic retrieval will return all documents that explicitly match an NLP request, and leverages the entire semantic net framework to include content containing concepts that are semantically synonymous.

Expert System native database retrieval includes MSQL and integrates with other DB technologies (e.g., Oracle 11g and Franz AllegroGraph).

Futures

Having established a presence in many of the largest multi-national corporations in information intensive industries (e.g., ENI – energy, Honda – automotive, and Homeland Security – government), Expert System has opened opportunities for expansion in each organization and vertical market. Once an enterprise has committed to semantic software for solving difficult search and discovery problems, it has begun the process of understanding:

- What types of questions can be answered;
- How to formulate queries;
- Practices for improving and maintaining a domain enhanced semantic net;
- The expertise resources that are most suited to implementation and ongoing usage.

When companies begin to build up expertise with the Cogito suite of products, experienced and expert users make inside selling a natural progression.

Recent investment in product marketing and presentations across industries, including major information technology conferences, have brought the Expert System story to every region of North America. They are consistently present and aggressively gaining a toe-hold outside of Europe.

Partnerships with Expert System are an opening for others in the semantics industry; their long experience with Microsoft shows that they can bring significant linguistic technology value to even the largest software companies. They have strengths that are "embeddable" and will be sought after on that front.

Consultancies and system integrators seeking search solutions at a reasonable price point will be an area that Expert System can exploit. Capgemini and Avande are two such partners. Others that make a serious commitment to supporting the products will have a springboard for multiple specialized opportunities that would not otherwise be possible.

Customer Testimonials

For us, using Cogito Answers, the NLP query option for our mobile devices is huge boost to our support efficiency. Average relevance to our customers for a question they pose is 93% and we save double-digit dollars for every query processed successfully that does not go to a service person.

We are trying to learn everything we can about our marketplace, competitors and customers. The (Cogito) knowledge extraction tools are excellent for finding information that is important to us, and that goes into a database for further analysis.

Two of the questions Cogito helps us answer are "Who are we doing business with?" and "What do we know about them?" This helps us in sourcing, contracting and re-selling within an organization.

With Cogito, we are able to add our own unique concepts to the semantic net, as needed, and then use the linguistic processing to help us normalize a lot of very unstructured content.

We have very aggressive schedules for proving the precision of the retrieval; with several testers trying to find problem areas, we really appreciate Expert System's flexibility and accommodation to help us with the process.

Corporate Facts

Corporate Headquarters:
Via Virgilio, 56/Q - Staircase 5, 41123 Modena - Italy
Phone: +39 059 894011
Fax: +39 059 894099
info@expertsystem.net

Sales and Support Contact Information:
United Kingdom
Office: +44 (0) 1590 612855

USA
Expert System US
J. Brooke Aker
Office: +1 860 728-8000
Mobile: +1 860 614-2411
baker@expertsystem.nettwitter.com/brookeaker

Officers: Marco Varone (Founder, CTO and President), Stefano Spaggiari (Founder and CEO), Andrea Melegari (Partner and Vice President Intelligence Division), Marcello Pellacani (Partner and Vice President Corporate Division), Luca Scagliarini (Partner and Vice President Strategy and Business Development), J. Brooke Aker (CEO of the US subsidiary)

Employees: 100+

Pricing: Pricing is done on a traditional software licensing model with recurring maintenance and support in the following years. Prices are calculated by server capacity and customer requirements (e.g., simple categorization is a lighter load on the server than a full semantic markup of every document). Cogito can handle 60-100 thousand documents per day for an average server. Starting points for the license model are less than $100k. A SaaS model is also available using hosted and cloud computing infrastructures. The SaaS model is less expensive up front but will be roughly equal to the licensing model within 5 years.

Status: Privately Held

Linguamatics

Linguamatics

Representative Customer Insights

> *We were looking for a product to extract information from millions of unstructured documents in published literature to answer scientific research questions. I had used the product at a previous company and made a strong push for it at my current employer where we continue to expand and improve its application.*

The first thing that captured our attention was Linguamatics' flexibility and support for analyzing and correctly interpreting completely new research questions and producing accurate results ... excellent linguistic processing and terminology disambiguation.

History

Linguamatics' name reflects its roots in linguistic and natural language processing. Founded in 2001 with headquarters in Cambridge, UK and US operations in Massachusetts, the company has 30+ employees focused on I2E software development and services.

In an interview with *Business Weekly UK*, Executive Chairman John Brimacombe describes its history, beginning with the development of the I2E platform by four colleagues who worked together as computational linguists and computer scientists at SRI International. Two of the founding team continue in leadership positions with the company, Dr. David Milward, CTO, and Dr. Roger Hale, COO.

The company has been profitable since its founding, with principal funding for product development and expansion coming from license and professional services revenues. The only other funding has come from UK and EU governmental research grants.

Beginning in 2003 Linguamatics has gained a solid position in the pharmaceutical industry with the deployment of its technology in many of the top-20 companies. It has well-established customer relationships with Pfizer, GSK, AstraZeneca, Roche, Bayer, Amgen, Biogen Idec, and *Boehringer Ingelheim.* In the industry it leads in text mining biomedical literature, applying agile natural language processing (NLP) to extract facts, weak signals, sentiment, and novel information relationships in real time. I2E concept mining and explicit relationship extraction across extremely large unstructured corpuses of high-value content is highly flexible and adaptable to rapidly answer unique questions.

Linguamatics participates in numerous collaborative relationships across its customer base, through business partnerships such as Accelrys, ChemAxon, and research collaboratives such as the *Collaborative Annotation of a Large-scale Biomedical Corpus* (CALBC). The latter provides a Silver Standard Corpus that can be used for evaluation of different systems and terminologies. The range of applications in which Linguamatics holds an NLP leadership position in biotechnology and pharmaceuticals underscores its broad reach and influence in the industry. A portfolio of case studies, and publications in peer reviewed and industry publications by Linguamatics founders illustrate their high credibility in the field.

Recognition of its business and technical success continues to accrue. Linguamatics was named *Company of the Year 2009* by the University of Cambridge Computer Lab Ring Hall of Fame Awards at Queens' College, Cambridge, UK. At BIO-IT World 2008 I2E was named *Best of Show* award winner *for life science information applications*, and at the 2010 Search Engine Conference in Boston David Milward's paper, *From Document Search to Knowledge Discovery: Changing the Paradigm*, was selected for the Evvie Award, the best Search Engine Study Paper at the conference.

Linguamatics has exceeded 50% growth for the past five years with sales revenue in 2009 over $5 million.

Descriptions of the Offerings

Linguamatics I2E, version 3.1

Linguamatics I2E is a software tool to be directly applied to answering specific complex questions from a very large, domain specific corpus of unstructured and/or semi-structured content. Its most unique strength is the application of natural language processing (NLP) to find intelligent information to answer questions that have never been asked before. As one Linguamatics veteran customer expressed it, unlike other semantic software that depends on pre-structured linguistic constructs and previously defined concepts, the flexibility and agility of the software supports totally new ways of framing queries and the disambiguation required to answer them. This is described as *meaning-based queries* that are correctly interpreted by I2E, returning facts that are highly relevant, and structured with all discovered relationships in tact with source content.

Using NLP-based text mining against any corpus of content, structured databases or unstructured files, together with domain specific ontologies, I2E returns the documents with structured relationships that may never have been defined before. These results present opportunities for further text mining and analysis. The synergistic terminology lists or ontologies may be pre-existing industry-based or proprietary or a mixture. I2E supports and works best when tuned and updated with customer language that will improve both query interpretation and retrieved results.

I2E has been steadily enhanced since breaking into the major pharmaceutical companies in 2003. Strong customer relationships feed a steady flow of product improvement ideas to meet needs as users push the technology, gain experience with the product and expand their vision of how it can be used. The latest major release I2E 3.1, October, 2009, features the following improvements:

- State-of-the-art NLP-based querying of a *greater choice of document types* including XML, HTML, plain text, Microsoft Word (doc and docx), PowerPoint (pptx), and PDF. I2E 3.1 also provides automatic updating and indexing of *MEDLINE* data.
- *Enterprise deployment* to a wide range of users with high performance querying, query publication via web services and portals, provision of query libraries, and *integration with enterprise collaboration tools.*
- The new *I2E Chemistry* option with *chemical substructure and structure similarity search* powered by *ChemAxon* for enhanced chemical text mining.
- *Enhanced results reporting*, with results previews, improved document rendering with hits in context, and the new *Results Curator* for review, analysis, and annotation of results. *Flexible hyperlinking* from extracted entities to web resources such as gene identifiers, glossaries of biomedical terms, and chemical structure visualization provide further scope for exploring results.

System Deployment

I2E features a client-server architecture built using industry-standard Java and C for use by single users, project teams, or in enterprise-wide deployments.

Organizations have full control over source content to be mined, definition of queries, and results output.

Web services provide access for example to I2E smart queries for use in portals and web parts.

System Requirements

Client-server configuration:

- Java version, client and server: Sun Java 1.5 (Java SE 5.0) or later
- Supported server OS: Windows 2003/8 server, Linux (Red Hat or Suse), Solaris (SPARC)

I2E Standalone:

- Java version: Sun Java 1.5 (Java SE 5.0) or later
- Supported OS: Windows XP/Vista

Strengths

Linguamatics offers solutions to mining text for answers to queries that are not easily discovered with other tools, and certainly not from conventional search engines. Its principal strength is ability for indexing huge quantities of unstructured content in a form that enables rapid query resolution with exceptional results relevance. Because it works with any terminology specific to a particular discipline and supports routine subject expert curation for enhancing that terminology, its applications are unlimited.

High-value content in a very knowledge intense domain, pharmaceuticals, was a smart place to first put the product to test. This is an industry where complex information problems secure technology funding. Informatics departments with subject matter experts and computational linguists are also prominently employed in this industry. They are the professionals who understand the power of a product like I2E and how to leverage it.

As a technology that works in highly heterogeneous platform environments with complementary software tools, these features of I2E stand out:

- *I2E provides for real-time, interactive querying* providing a mechanism that fits well in a fast-paced, research-intensive R&D team. As it has matured, the usage interface has improved and users are becoming more sophisticated about potential applications. Not unlike hyperlinking as a mechanism for allowing the curious to follow unstructured paths through a maze of information on the web, the agility of I2E presents unlimited possibilities for research empowerment. Via exploration of information resources coupled with bench science, scientists will be able to seek and test possible theories more rapidly.

- *I2E supports users at levels beyond the NLP queries posed by experts*; power users with advanced business analytical skills will export results to Excel for further analytic evaluation and occasional users will benefit from simpler but equally powerful access via smart queries.
- *I2E supports and collaborates efficiently with numerous other technologies*, a reflection of its practical foundation that acknowledges a multi-vendor, multi-platform environment for solving high value, complex information problems.
- *I2E is rapidly deployable recovering ROI quickly;* it delivers at least a 10X improvement in the processing of queries and delivering relevant information over conventional automated search tools. This contributes to financial gains from increased productivity, and better and timelier decision support (e.g., one customer recently stated that with a couple of hours work with I2E they made a project saving, which repaid their one-year investment in the software).
- *I2E discovers novel relationships among entities* and also makes it possible to have tailored, faceted search for each customer. Discovering new relationships across content types, among previously unrelated entities, has the benefit of suggesting new constructs for queries that can lead to opportunities for discoveries.
- *I2E's deep mining provides new ways of getting at weak signals;* weak signals are another way to gain insight that might reveal research and discovery opportunities or a competitive threat.

Problems Solved

Life science problems solved: relationship and fact discovery are coming from these critical areas of biomedical research via I2E processing:

- Earlier discovery (pathways, target ID, validation)
- Gene-disease interactions
- Biomarkers
- Systems biology
- Pharmacogenomics
- Toxicity, safety, pharmacovigilance
- Patent analysis

Competitive intelligence leading to licensing opportunities, key opinion leadership, trend, and sentiment analysis can be pursued by answering specific questions like:

- Which companies are working on technology C?
- What compounds are available for in-licensing in a disease area?
- Which research groups are my competitors collaborating with?
- What are key opinion leaders saying about our product?

- In the marketing intelligence domain, I2E has been applied for its ability to do sentiment analysis, providing partner 81qd with technology to discover knowledge critical to optimal life cycle management planning.

Decision Support for R&D can find answers to define "what target research should we invest in?" by asking things like:

- Which proteins interact with protein X?
- What genes are linked to disease Y?
- What are potential biomarkers for condition A?
- What dosages of compound B cause adverse reactions?
- What alternative indications exist for my compound?

Strategic Advantages and Competitive Positioning

Linguamatics has a strong position as it aggressively pursues its technology roots. Roger Hale, Linguamatics COO, in this 2005 article, *Text Mining: Getting more value from literature resources*, lays out the concepts behind *Interactive Information Extraction (I2E)*. Certainly the message has resonated within the informatics community of major adopters, top ten pharmaceutical companies. Professional interchanges within the industry give Linguamatics a very prominent position. Their customers have written and spoken in numerous professional forums about the company and I2E.

In conversation with David Milward (CTO) and Phil Hastings (Director, Business Development), we are impressed with the focus on professionals in the field who are at the frontlines of molecular modeling and other biochemical analytical challenges. Their respect for the scientific process and attention to making I2E work well for these scientists, speaks to a high level of business integrity. It is surely a reason for the positive commentary received from customers who know that they are listened to by Linguamatics.

Biotechnology, semantic technology, and search engine meetings provide exposure to new market opportunities and give Linguamatics platforms for educating potential users about their unique semantic software technology. While natural language processing has been evolving for decades, Linguamatics' value proposition has been front and center for several years. Having a practical product application that is relatively turnkey for solving known problems has given them a competitive edge.

Press and meeting exposure among professionals in the fields of search, text mining, and text analytics is opening many opportunities for partnering with complementary technologies and finding vertical markets ready to embrace a truly new software design for exploiting massive content stores.

As noted by Dr. Milward and Dr. Hastings, they are usually approached first by informatics or library science professionals who have identified I2E through their professional relationships and meetings. Defining and explaining the business impact of semantic technologies and linguistic processing requires clear differentiation against conventional search technologies. Linguamatics is committed to helping professionals in the process of making the case to their management. They understand how these complex technologies need to be positioned inside enterprises.

Futures

With I2E's initial success well-established in the life sciences R&D arena, Linguamatics is building relationships and beginning engagements in other linguistically expansive domains.

Other chemical, scientific, and technical industries are beginning to understand the possibilities of deep text mining, and how it can improve their scientific research. Linguamatics has plenty of growth opportunity. Phil Hastings emphasizes that Linguamatics is working to maintain focus on the pharmaceutical industry and not trying to do everything at once. That said, he also noted plenty of related opportunity in healthcare and homeland security domains.

The partners cited and press mentions on their website will give readers a good idea of where Linguamatics has been and where it is evolving. They are clearly seeking growth through new business channels and earlier in 2010 added a management position for this function.

Co-founder David Milward sees gains for Linguamatics as companies recognize the importance and need to extract value from what they and the industry already know. Management is requiring their IT groups to find technologies to solve large-scale data and text mining problems. Once they begin the deployment and adoption, users' desire to find answers to more complex and difficult questions quickly scales. According to Dr. Milward, "We know Linguamatics can add value on both the business side and on the science side."

One more news item illustrates how Linguamatics is capturing the imagination of potential adopters of sentiment analysis applications based on text mining where "tweet" analysis revealed opinions of "Twitterati" relating to the British elections.

Customer Testimonials

Linguamatics has been extremely responsive and demonstrated that (even at its early stage) it is competent to work with large and influential customers

We were looking for an NLP product but wanted it to be lightweight and give us the flexibility we needed. Linguamatics specializes in agility and I knew it would address multiple research initiative problems.

No other products had the same set of capabilities, especially processing over 17 million abstracts and documents, some a 1,000 pages in length.

Using a small team of business and informatics people, an IT project manager, and a text analytics expert, the payback has been excellent. Our researchers are experiencing a 10X to 1000X time savings to answer very complex questions over conventional literature searching, and with more relevant results.

The ability for the researcher to explore questions that no one else has asked and get back relevant results is impressive.

Text and data mining is a learning process, and as we become more skilled at applying the technology it will render more and more (positive) business impact. We need to emphasize the business benefits, not the technology benefits, and Linguamatics recognizes that is their mission, too.

We have limited support staff for developing smart queries and analyzing results but want to continue to enrich the terminology with new concepts. I2E minimizes the human curatorial effort in building up the ontology to improve the NLP. An added benefit was automatic entity tagging of the publications database against our proprietary thesaurus (proteins, diseases, adverse events), thus enhancing its value.

The Linguamatics technology has not been fully exploited; my observation is that most companies deploy it for a single project and could dramatically expand their vision of all the queries that would be satisfied. Finding more informatics specialists with multi-disciplinary backgrounds will help push use into new areas.

People who have a systems view of how things work tend to be better at semantic linking possibilities and the need for these people is growing.

Once Linguamatics came out with version 3 of I2E, our earlier concerns with its usability and interface tools were eliminated, and we could realize a big advantage over other products that we had tried and were just not flexible enough.

The product flexibility allows us rapid deployment to build small corpuses of content, experiment and take them down after we have captured and stored results in a database.

The company has been very open to new development ideas and projects. Also, when we share with them usability issues or flaws in the product, they receive the feedback and are quick to fix issues. Their deliveries of new releases have been very comprehensive.

Corporate Facts

Corporate Headquarters:

Linguamatics, St John's Innovation Centre, Cowley Rd, Cambridge CB4 0WS UK
Tel: +44 1223 421360 (main)

North American Regional Headquarters:

Riverside Center, 275 Grove Street, Suite 2-400, Newton, MA 02466, USA.

Sales and Support Contact Information:

North America
Susan LeBeau
E-mail: susan.lebeau@linguamatics.com
Tel: +1 774 571 1117

Europe and ROW
Phil Hastings
E-mail: phil.hastings@linguamatics.com
Tel: +44 1223 421360

Officers: Mr. John M. Brimacombe (Executive Chairman), Dr David Milward (Chief Technology Officer), Dr Roger Hale (Chief Operating Officer), Dr Phil Hastings (Director, Business Development)

Employees: 30+

Pricing: I2E pricing varies depending on a number of factors such as number of users and product options licensed. Professional services may also be considered as an option. Please contact Linguamatics for a requirements assessment and pricing options.

Status: Privately Held

Lynda Moulton
Senior Analyst and Consultant
gilbane@outsellinc.com

At Outsell, independence and objectivity are at the core of our business and our values. Read our Ethics & Integrity Policy

Outsell is committed to minimizing our impact on the environment and acting in a sustainable way. Read our Green Policy

Outsell is the only research and advisory firm focused on the publishing and information industries. Our international team provides independent, fact-based analysis and actionable advice about competitors, markets, operational benchmarks, and best practices, so our clients thrive and grow in today's fast-changing digital and global environment.

Call +1 617.497.9443
Fax +1 617.497.5256
763 Massachusetts Avenue
Cambridge, Massachusetts 02139

Call +1 650.342.6060
Fax +1 650.342.7135
330 Primrose Road, Suite 510
Burlingame, California 94010
info@outsellinc.com
www.outsellinc.com

Call +44 (0)20 8090 6590
Fax +44 (0)20 7031 8101
25 Floral Street
London WC2E 9DS

Copyright 2010. All rights reserved.

OUTSELL™
Advancing the business of information