SANDIA REPORT
SAND2003-2916
Unlimited Release
Printed September 2003

Natural Language Processing-Based COTS Software and Related Technologies Survey

Michael G. Stickland, Shelley M. Eaton and Gregory N. Conrad

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2003-2916
Unlimited Release
Printed September 2003

Natural Language Processing-Based COTS Software and Related Technologies Survey

Michael G. Stickland
Advanced Decision Support Applications

Shelley M. Eaton
Software and Information Engineering Department

Gregory N. Conrad
Advanced Decision Support Applications

Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1138

Abstract

Natural language processing-based knowledge management software, traditionally developed for security organizations, is now becoming commercially available. An informal survey was conducted to discover and examine current NLP and related technologies and potential applications for information retrieval, information extraction, summarization, categorization, terminology management, link analysis, and visualization for possible implementation at Sandia National Laboratories. This report documents our current understanding of the technologies, lists software vendors and their products, and identifies potential applications of these technologies.

Table of Contents

Acronyms

API	Application Programming Interface
IDE	Integrated Development Environment
COTS	Commercial Off-The-Shelf (Software)
CPR	Corporate Policy Requirement
KPP	Knowledge Preservation Project
NLP	Natural Language Processing
NTK	Need To Know
PLSI	Probabilistic Latent Semantic Indexing
SDK	Software Development Kit
WFS	Web Fileshare
XML	Extensible Markup Language

Natural Language Processing-Based COTS Software and Related Technologies Survey

Introduction

Problem Statement

The current standard technologies used for indexing, searching for, and retrieving information from unstructured text in electronic documents are ones that index keywords that were either manually associated with the documents—the document's metadata—or that were automatically extracted from them using fairly simplistic automated methods. These search engines then retrieve documents by matching user-provided search words to the keyword index. This type of technology is failing to keep up with the current knowledge management needs of organizations for several reasons. A primary reason for this failure is that keyword-based search engines—such as those provided by Verity, Stellent, and Google—do not adequately take into account the context(s) in which these keywords are used within each individual document. Subsequently, searches performed with them can produce search results containing hundreds or thousands of documents, many of which are probably irrelevant to the actual information being sought, that the knowledge seeker must then manually sift through to find what they are looking for. Another factor in this failure is that a vast quantity of time and resources can be needed when manually creating document metadata, with or without producing the desired benefit to the organization.

In managed document collections, knowledge workers may spend great amounts of time attempting to determine an appropriate set of keywords to use for a particular collection of documents, and to developing the processes necessary to sufficiently assign them. These efforts are quite costly since they depend on the tedious manual efforts of trained knowledge workers who must attempt to consistently and accurately assign keywords to documents. Even with this level of effort, the full range of topics contained in comprehensive documents may still not be adequately covered by the assigned words; and since knowledge workers are usually not subject matter experts, they often are unable to accurately and completely assign the appropriate terms.

In unmanaged collections, the knowledge provider must spend time trying to express the contents of a document using only a limited number of words. In the best case, due to the unverified nature of this method, the keywords could cover the full range of topics, or else they may cover only the main topic, or they may not be entered at all. In a worst-case scenario, documents may never leave the desktops of knowledge providers simply because they don't want to take the time to deal with entering metadata, thus leaving the organization without any access to their information. To further exacerbate the issue, since the unmanaged entry of keywords allows for the use of non-standardized keywords, it is virtually impossible for a knowledge seeker to know which words will be effective to use in their searches.

Goals

The goals of this survey project were to discover, catalog, and examine natural language processing (NLP) based software and related technologies that are currently commercially available to the business community. We set out to determine their capabilities, the kinds of problems they can address, examples of how they can be applied, and whether or not they may potentially provide more advanced knowledge management capabilities to Sandia National Laboratories.

Scope

The software we surveyed is not intended for desktop installation. The survey focused on infra-structure-level solutions that would integrate into a corporate computing environment, or into custom software applications, providing enhanced capabilities for more robust performance than are currently employed. We did not include speech recognition technologies, Internet search tools, email crawlers, document or email routing, or low-level linguistic development tools in the products we evaluated. However, you may find some of these capabilities in the products we identified. We did include non-lexical analysis based technologies, such as those that use statistical or pattern-matching approaches, as long as they were being used to process unstructured text sources. Additionally, we included tools that don't perform any NLP functions at all, but that provide enhanced methods for visualizing, navigating, or machine learning, reasoning, and analysis of textual knowledge.

Survey Approach

We started by conducting an extensive web search in an effort to identify as many NLP based software products and related technologies as we could find. Once a candidate technology was located, we acquired the available product literature and/or technology white paper (via downloads from web sites and/or direct requests to vendors), and reviewed them for the sole purpose of determining which technology category or categories each fit into. We then documented which technology categories each product belonged to, and which products were produced by each vendor. We also collected a list of vendor contact information.

The vendors were then solicited for additional information about their NLP based products and technologies using an email questionnaire we developed for that purpose. Next, a more in-depth review of the literature was conducted —including vendors' responses to the questionnaire—in order to determine which products the team would attempt to examine further. Since time and resources were limited, we would only be able to conduct cursory hands-on evaluations or view vendor-provided web demos of a limited number of products. Therefore, the perceived usefulness, technological advancement, and maturity of each product, as well as the availability of evaluation software or web demos, were taken into account when deciding which to pursue further. At least two evaluation software products from each technology category were requested from vendors. When on-site evaluations were not available or practical, web demos were sought.

We did not attempt to perform qualitative evaluations and comparisons of the products we chose to look more closely at. Instead, we sought to gain a much better understanding of each technology category, the diversity and viability of the products within them, and the potential benefits to Sandia National Laboratories of adopting them. These were simply the first steps taken towards a true understanding of the products and technologies that will emerge to be the most effective at providing the greatest benefit to users of specific future applications.

NLP-Based Technologies

Natural language processing technologies differ from traditional information retrieval technologies in that they attempt to "understand," at some level, the semantic context of the text being processed, as opposed to search engines that only search on keywords without taking context into account. Understanding the semantic context that words and phrases are being used in is very important for enhancing a software system's ability to effectively process unstructured textual information, just as it is for our capacity to do the same. Many of these technologies utilize linguistic-based approaches to gain an understanding of the text, a number of them employ statistical or pattern-matching based methods, others rely on rule based logic processing, and some exploit multiple styles. Yet, as if indifferent to this distinction, many in each class are capable of semantically processing unstructured text, and are being commercialized to perform a wide number of different functions. Thus, it seems that only time will tell us which approach(es) will prove to be the best. In the meantime, we'll be keeping an open mind to all of them.

Linguistic-based techniques involve low-level activities, such as word segmentation (tokenization), that identifies individual units of text (words, word particles, abbreviations, punctuations, etc.) and stemming that identifies the true stem (base form) of individual text tokens (i.e. swim, swimming, swims, and swam are all forms of swim). In addition, part-of-speech tagging identifies the part of speech (noun, verb, adjective, etc.) of each word within the context it's being used and tags them accordingly. This all leads to the identification of proper nouns and noun and verb phrases that represent the entities, concepts, events, and their relationships that are contained within the text.

Statistical and pattern-matching techniques use mathematical models to identify entities and concepts. They can also be used to group like documents into clusters, map individual documents to pre-defined topic categories, or map user queries to relevant documents or portions there of. Algorithm types used in these methodologies include Bayesian Probability, Neural Networks, Support Vector Machines, and K-Nearest Neighbors, explanations of which are beyond the scope of this document. Finally, rule-based systems can spell out very specific, fine-grained definitions of the criteria used to process a document. They rely heavily on subject matter experts to define them and usually only work with keywords, but can incorporate decision trees and Boolean Logic for greater accuracy.

There are a wide variety of products that implement these different techniques, and most of them fall primarily into one or more of five specific NLP technology categories. Additionally, as was previously stated, we also included non-NLP based tools in our survey for a total of seven technology categories. The first five of these—Information Extraction, Terminology Management, Categorization, Summarization, and Question Answering—represent NLP based technologies, while the other two—Visualization and Link Analysis—represent technologies that can be used to provide enhanced user interfaces for accessing

and understanding the information resulting from the use of NLP technologies, and/or to perform advanced forms of analysis on it. The table below offers brief descriptions of each of these seven categories.

Technology	Description
Information Extraction	• Identifies nouns, noun phrases, and other lexical units from texts • May identify entity types, i.e. names, places, organization, etc. Can identify relationships between concepts • May extract relationships between entities, or facts about events • Can be foundational for categorization, summarization, and question/answering technologies
Terminology Management	• Allows for the development and maintenance of knowledge organization structures (i.e. taxonomy, ontology, dictionary, thesaurus) which can enable or enhance many other NLP technologies
Categorization	• Identifies the main topic, or all significant topics, in a document • Maps documents to predefined topic categories based on content
Summarization	• Identifies the main, or most significant, topics in a document • Produces a summary or abstract of a document by excerpting and displaying the sentences or sentence fragments that best portray its content
Question/Answering	• Understands the meaning of the natural language questions posed by users, and of the concepts and sentences in source documents • Maps semantic understanding of user questions to previously understood sentences contained within source documents, which (potentially) contain answers to them
Visualization	• Visually represents knowledge organization structures and/or relationships between entities, concepts, or data • Can provide easier navigation, clearer understanding, and/or quicker discovery of information
Link Analysis	• Provides features for analyzing links between, and learning and reasoning about entities found in texts • Can be used to facilitate the discovery of new information, or to build advanced applications for problem solving, event probability prediction, etc.

Table 1: NLP-Based Technology Categories.

As well as falling into the seven categories mentioned above, these technologies may also have enabling or dependency relationships with one another. As shown in Figure 1, Information Extraction, Categorization, Summarization, and Question Answering all depend on one or more of a few general classes of core technologies. In addition, though the diagram doesn't depict this, the latter three may also depend on Information Extraction. Some Terminology Management tools, such as those that perform automatic taxonomy generation, also rely on these core technologies. The core technologies may, in return, depend on the knowledge organization structures (i.e. taxonomy, ontology, etc.) and rule bases created and maintained using the Terminology Management tools.

The Visualization of NLP-derived information relies on the output from using Information Extraction, and/or Categorization (i.e. possibly hierarchical, topically related clusters of documents), which depends on knowledge organization structures, again produced using Terminology Management software. Likewise, in order for Link Analysis technologies to do their job, they must also be provided with the output produced by Information Extraction techniques, including the relationships which link entities, concepts, events, and facts to one another. As can be seen, there can be quite a few technologies needed to produce a single NLP or related software solution.

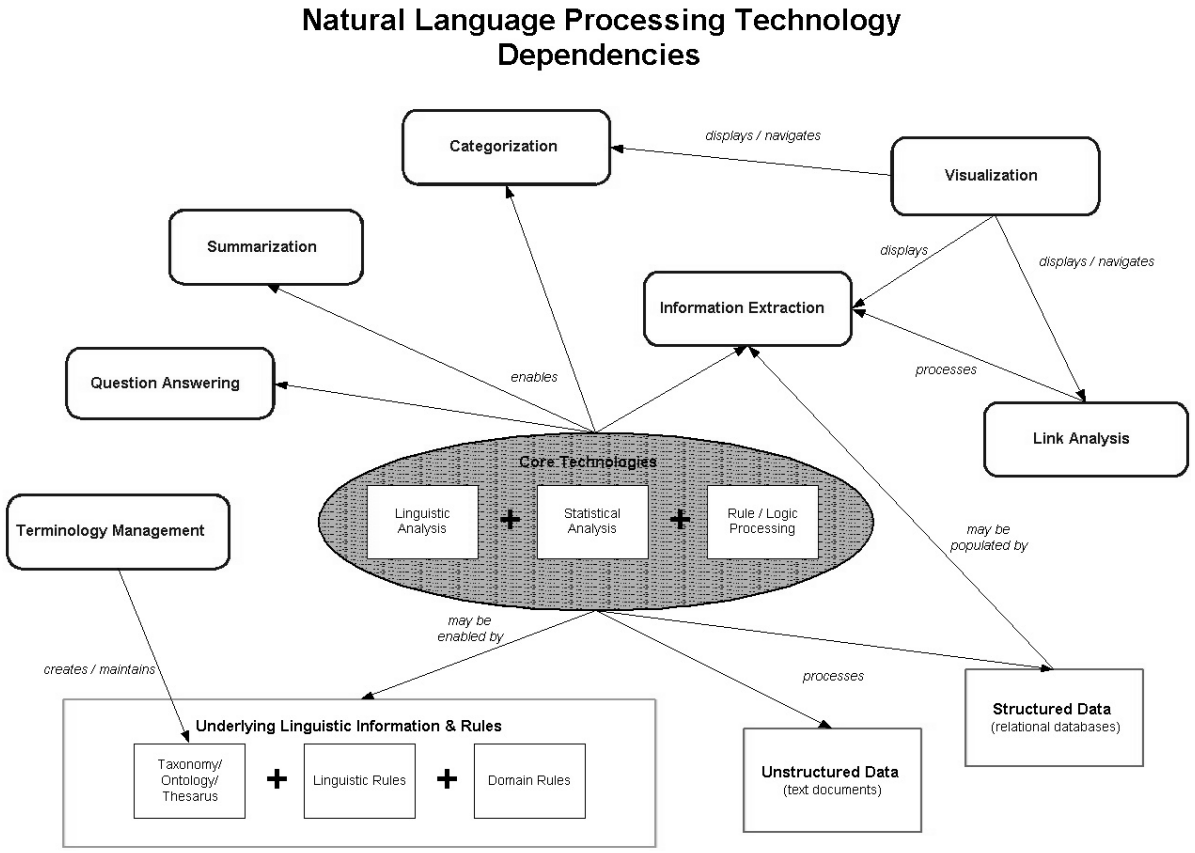

Figure 1: NLP Technologies and their interdependencies.

Information Extraction

Several varieties of semantic element types can be extracted from unstructured text sources. These include entities, concepts, events, facts, and relationships. One type of element commonly extracted is the names of entities like people, places, companies, products, components, materials, and chemicals. In addition to being extracted, they can also be identified by their entity category type (i.e. person, location, organization, etc.). Another semantic element type that is frequently extracted is concepts. Like named entities, concepts are often represented by phrases containing more than one word that when taken together provide greater meaning than when taken separately. Examples of these include “national forest,” “high school reunion,” and “green chili chicken enchilada.” Entity and concept extraction is a valuable capability because it provides a foundation for superior semantic machine understanding of text documents, thus allowing for more accurate search results and contributing to the success of other NLP technologies.

Identifying which entities are extracted, and how they are semantically tagged, is something that can be tailored for each individual domain and for specific applications.

There are extraction technologies that perform tasks that are more complex as well. These involve a level of understanding beyond that of recognizing entities and concepts: they require the discovery of relationships between them or of the facts concerning specific events. Many types of relationships can be extracted, with one common type being affiliations between people and organizations, as well as their positions within them (e.g. ORG-NAME: Sandia National Laboratories, OFFICER-TITLE: President, OFFICER-NAME: C. Paul Robinson). It is also possible to identify and associate the facts surrounding an event with the event itself. One example of this would be to extract information about product failures, which could include the name of the failed component, the system it's used in, the reason for the failure, and the date it occurred (e.g. EVENT: component failure, COMP-NAME: sprocket axle, SYS-NAME: chain drive unit, FAILURE-REASON: corrosion at base, FAILURE-DATE: July 22, 2002). This ability to extract relationships and facts can be used effectively to convert unstructured information into structured information; and this can be used to develop advanced applications for such areas as competitive intelligence, national security, decision support, and defect tracking, to name a few.

Terminology Management

Terminology Management tools can either be standalone applications or embedded in a larger NLP solution. They are used for the automatic (unsupervised), assisted (supervised), or manual construction and population of a knowledge organization structure (i.e. taxonomy, ontology, thesauri, etc.). These knowledge organization structures can be used to enable or enhance the capabilities of other NLP based technologies, by providing them with a known collection of terms and their semantic context(s). Products in virtually all of the other NLP technology categories utilize assets maintained by these tools. Note that regardless of whether they provide automatic, assisted, or manual terminology management, all of these tools should provide user interfaces to allow knowledge workers to update the knowledge organization structure(s) being managed.

Taxonomy Creation.

There are several approaches that can be used to define the nodes in taxonomies: Rules of one form or another, groups of related noun phrases, and sample document sets are all ways of manually defining them. Automatic and assisted taxonomy creation tools will determine or suggest the nodes for you based on the topics they find in target document collections. To accomplish this, they may first use linguistic methods to identify the entities and concepts in the documents, statistical or pattern matching processes to do this, or may rely solely on extracting keywords. In any case, statistical or pattern matching techniques are then used to determine or suggest the main themes and sub-themes of information contained within these complete, or training set, collections of documents. The identified nodes, or topic categories, in the generated taxonomy are then populated with the extracted entities and concepts, or keywords that best represent their themes. These topic profiles are what Categorization software tools map individual documents to when determining to what topic category(s) they belong.

Reusability

While conducting this survey, we realized that an important feature of any terminology management tool, whether standalone or one embedded, is the ability to import and export knowledge organization structures along with the information that populates and describes them. The reason we feel this is important is one of reusability. Since it can take significant resource expenditures involving experienced

knowledge workers and subject matter experts to develop and refine the knowledge organization structure for a single domain, the ability to reuse this information is highly desirable.

Categorization

Like automatic and assisted taxonomy generation software, Categorization software products identify the entities and concepts in documents and use statistical or pattern matching approaches to comprehend the main themes included in them. However, instead of determining the main themes in a collection of documents, they map the contents of individual documents to one or more categories contained in a previously defined taxonomy or topic list. Accordingly, Categorization technologies depend on the knowledge organization structure definitions created using Terminology Management tools, since this is what documents are mapped to.

The results produced using Categorization tools can be used in a variety of ways. User interfaces that exploit categorization technologies provide topical browser views of document collections, or add semantic context to search results. Categorizing the same documents in multiple ways can provide different views into them for different groups of knowledge workers, or for different tasks. Other uses for categorization include document and email routing applications, profiling the interests or areas of expertise of people by analyzing the documents they write together with the ones they read, and analyzing existing or suggesting appropriate need-to-know controls for documents.

Summarization

The underlying techniques used by Summarization tools are very similar to the technologies that empower Categorization and automatic and assisted taxonomy generation. The main difference is found in what document Summarizers do with their understanding of the content of a document. Summarization algorithms identify the sentence fragments or sentences that best represent the main theme of a document and produce a concise summary or abstract describing it. These document summaries allow a knowledge seeker to quickly ascertain the main theme(s) of a document, within the context of some of its own text. One example of this is when the summary is presented as the content associated with a document in a search results set. Another is when it's used to populate one of the metadata fields associated with a document in a repository. However used, an intelligent, concise summary can vastly improve a knowledge seeker or worker's ability to determine the relevance of the information in a document for their current needs.

Question Answering

Question Answering solutions attempt to provide the knowledge seeker with precise answers to their questions. What this class of tools does is analyze the sentences within a collection of unstructured text sources and the natural language questions posed by knowledge seekers and then attempt to map the questions to one or more sentences in source texts that contain answers to them. To accomplish this, they too can rely on one or more of the linguistic, statistical, pattern matching, or manually authored rule-based methods for NLP. However, Question Answering software can produce much more precise retrieval of very specific information than traditional search engines can and represents a big step forward for information retrieval technology.

This leap in information retrieval performance is not without cost. While most Question Answering systems can perform better than traditional search engines can right out of the box, they perform best when

augmented with domain-specific terminology and rules. Question Answering technologies require the resources of skilled knowledge workers to set up and maintain the underlying knowledge organization structures that enable their ability to intelligently understand text at the sentence level within a domain.

Question Answering technologies can be used for any applications where users need to query repositories for specific information, including customer self service, call centers, corporate internal and external websites, and reference materials.

Visualization

Visualization technologies provide alternate ways of viewing and navigating information that is beyond traditional browser-based access to content taxonomies or document directories. They present the user with graphical representations of related information that can provide a level of understanding that is not otherwise easily achieved.

Visualization software graphically represents information in knowledge organization structures, such as those created using taxonomy creation and categorization tools, or the results of information extraction activities. Visualization technologies often provide the user interface for information discovery and retrieval, and link analysis applications. Examples include star trees, time lines, and hierarchical structures.

Link Analysis

Link Analysis tools allow discovery of relationships amongst entities. The visualization can represent an aggregate of all of the relationships found for entities, not only within an entire document, but also across a collection of documents. If one part of a document talks about Person A belonging to Organization B and another document discusses Person C's relationship to Organization B, Link Analysis technology will show that there is a relationship between Person A and Person C that goes through Organization B.

Some technologies included in this category can also perform machine learning and reasoning functions. This type of capability can be used in a variety of advanced artificial intelligence applications like logistics planning, event probability prediction, and intelligence gathering.

Software Product Reviews

In order to gain a broad understanding of NLP based software capabilities and the general maturity level of the industry, we reviewed at least one software package from each of the technology categories. These reviews included web demonstrations, conference calls, training classes, and evaluation software installations. While none of these reviews were in depth, we acquired insights into how the technologies could fundamentally change how Sandians search and retrieve knowledge. In addition, we found that it was possible to integrate existing Sandia knowledge, such as the library's thesaurus data, with some of the software products.

With the help of the Library, we compiled a test set of 100 unlimited distribution, unclassified SAND reports all dealing with the topic of solar energy, in either Adobe Acrobat or Microsoft Word formats. The goal was to test evaluation software using our own data, rather than depend on the vendor's prepared set of data to demonstrate software functionality. One company, InQuira, was able to use this test set of data and process the information with their question answering software. A product that we were able to process

this information with was LexiMine from SPSS, which provides text mining and link analysis capabilities. Results of these tests were not available at the time of this publication. Attempts to evaluate other software packages using our data failed because the software products proved to be too complex to configure without the vendors' support, which was not available for this project, or time was too short for us to do so.

The following are brief descriptions of the software vendors' products that we reviewed, along with a short listing of the products' capabilities and benefits. A complete listing of software products that we found in our survey is shown in Appendix A: Software Technologies by Category. Appendix B is a catalog of the software products by company or organization. Company/organization contact information is listed in Appendix C.

Attensity Corporation

By extracting event information from unstructured text, Attensity software is able to support advanced knowledge management activities, including link analysis, machine learning, and trending, for business and intelligence communities. Attensity uses a linguistic approach to locate specific, known, event information and transform that information to structured text or relational tables. MOAB is the core natural language processing engine. VisualMOAB is a visualization tool which generates a graphic parse tree of sentences and shows what's going on within the MOAB engine's "black box." The Alta Server is available in the commercial market and incorporates the MOAB engine, while the Malta server is designed for intelligence and government groups. The Unter tool is for domain rule creation and dictionary management.

Capabilities and Benefits

- Provides document collection, conversion, filtering, meta-tag handling, event extraction, XML output, dictionary creation, and event definition creation.
- Can cross sentence and paragraph boundaries and can create document-level event extraction.
- Handles anaphora (person name = he, she, etc.).
- Processes unstructured text at a rate of up to five megabytes per minute.
- Deals with poor grammar, misspelled words, shorthand, unknown terms, etc.
- Works on Windows, Linux, and Unix platforms and ODBC/JDBC compliant databases.
- Extracts data from multiple data sources, including databases, email, and the web.
- Provides for broad or narrow (less detailed vs. more detailed) event definitions.
- Depends on context and clarity of text for higher accuracy results, along with well-defined domains, to avoid problems with disambiguation.

Discern Communications, Inc.

Discern is an automated Question Answering application that answers users' natural language text queries. The software's natural language processing functions identify key topics and entities in the text, such as people, companies, products, problem conditions, times, locations, and other attributes. These words and concepts are indexed and stored, so they can later be retrieved by the users. This software grew from over ten years of SRI International (formerly Stanford Research Institute) natural language research for the Department of Defense.

Capabilities and Benefits

- Questions can be spoken or typed, and can either use full sentences or just keywords.

- Provides answers at the sentence level instead of providing just a list of related documents for the user to read through.
- Multiple instances of domain-specific lexicons can be deployed in a distributed architecture to handle multiple domains.
- Discern is now developing crawlers for web sites and directory structures.
- Integration with repositories is available through customization.
- The NLP engine is rule and statistics based.
- Multiple interfaces, including mobile phones and hand-held devices, can be employed.
- Provides access to multiple data sources, such as databases or unstructured text.
- Terminology is managed in the lexicon/dictionary as synonym sets.
- Taxonomy relationships (e.g. is-a) are managed separately in a tree structure that maps to the atomic terms in the lexicon.
- Dialog context is not currently kept by the system, but Discern is working towards that feature in the future.

InQuira, Inc.

InQuira™ 6 is a Question Answering software product that enables users to ask a question in natural language. The software then interprets the intent of the query and automatically provides the answer and may point to other relevant documents, web links, or information. In addition, it provides sentence-level answers in context, if the context has been defined. Documents developed with some sort of rigor (e.g. Sandia's CPRs) are usually easier to retrieve answers from than non-structured texts (e.g. emails).

Capabilities and Benefits

- Access controls are implemented based on content and may be customized.
- Collects information about users and their searching experiences in order to point out areas of improvement.
- Uses an ontology along with language and business rules to create packaged dictionaries that can be extended with industry and company layers.
- Industry extensions can be combined and company extensions can be segmented. This is the language model and is available for programmatic access.
- Offers several ways to search to provide flexibility to the user, from natural language to key-word searches.
- Works with unstructured context, including text in databases, as well as using parametrized or fielded data.
- Offers dictionary management, accuracy reporting and management, and systems administration functionality.
- Scales to very large organizations because InQuira's question-processor function can be replicated over multiple servers and its architecture is concept-based.

Invention Machine Corporation

Invention Machine offers a suite of applications geared toward helping engineers solve problems and develop creative designs in R&D environments. The software employs a large commercially available

library of scientific effects to assist engineers during the conceptual design stages of product development. It employs semantic indexing technology, which is based on natural language processing.

Capabilities and Benefits

- TechOptimizer™ employs analysis algorithms to diagnose the weak elements in a system or a manufacturing process, and comes up with problem-solving goals.
- The TechOptimizer™ Function Analysis and Trimming applications help engineers to identify the core problems in systems and processes.
- The Effects Knowledge Base contains over 8,000 animated effects and examples to aid in problem solving.
- TechOptimizer™ Prediction provides a process to engineers to design new products, proposes modes, and resolves contradictions in engineering systems.
- Goldfire™ platform is a scalable enterprise solution that employs natural language processing technology to enable retrieval of engineering knowledge, both from internal and external sources.
- CoBrain™ allows centralized access to the Internet, Deep Web, Patent Office databases, and corporate resources. It indexes unstructured information for later retrieval.
- Goldfire Intelligence™ captures information that can later be analyzed and shared.
- Knowledgist™ is a tool designed for knowledge workers to manage search and extraction of useful data from electronic information, utilizing NLP-based technologies.

Inxight Software, Inc.

Inxight Software offers a suite of integrated tools to manage, extract, categorize, summarize, and analyze information from various places, including document repositories and the web. These tools include Categorizer™, Categorizer Executive™, LinguistX Platform™ software development kit (text analysis for information retrieval applications), and Smart Discovery™ (metatext extraction which includes Summarizer™, ThingFinder™, Concept Linker™, and Similarity Finder™).

Inxight Categorizer™ automates classifying, analyzing, and administering electronic text. The tool automatically updates information when documents are added, changed, or retrieved from repositories. In addition, the tool updates the categorization scheme as new topics are identified or changes are made.

Inxight Categorizer Executive™ is an administrative tool for knowledge workers that includes a graphical user interface for creating and maintaining training sets.

Summarizer™ extracts a summarized version of the text. Knowledge workers can determine the relevancy of a retrieved document using a summary instead of downloading and reading the entire document. The user specifies length and topics of interest.

Thing Finder™ identifies a document's proper nouns such as companies, products, places, and people and then places them into appropriate categories. The user then searches based on these categories. The user chooses from 29 standard categories or creates custom categories.

Concept Linker™ aids in searching a collection of documents for a keyword or phrase. The software generates a list of other words and phrases that are related to the keywords. The user then conducts searches using these related words. The software can create indexes of related concepts to narrow searches.

Similarity Finder™ is a service that finds subject and content similarities in the multiple documents.

Star Tree Studio is a visualization product that allows navigation through related information.

Capabilities and Benefits

- Allows the text index data to be used for other purposes.
- Machine-learning algorithms, morphology, syntax, semantic, and contextual analysis are employed in 12 languages.
- Provides self-learning to update categorization as more documents and categories are added.
- Provides optimal high- and low-confidence thresholds for each category, using the training set of data.
- Supports over 70 common word processing file formats including HTML, PDF, MS Office Suite, email, Lotus Notes, and plain text.
- Connects to any database that supports XML-based queries, e.g., Oracle, Sybase, and Microsoft SQL Server database management systems.
- Provides API coded in Java and C with XML output and enables complete integration into portal environments, internal applications, or third party software.
- Integrates with systems and applications, by using XML, Java, JSP tag library.
- Support for four languages: English, French, German, and Spanish, with other language support planned for the near future.

Megaputer Intelligence, Inc.

TextAnalyst™ uses linguistic and neural network technologies to extract information from unstructured text to increase the speed and accuracy of search and retrieval activities. This allows users to quickly understand the information in a document repository.

Capabilities and Benefits

- Distills the meaning of texts by creating an accurate semantic network of the information, which can then be used for further analysis.
- Summarization of texts is achieved by using both linguistic and neural network text processing methods.
- The user can exclude or include specific words or phrases to guide the search for information.
- Hyperlinks are used to relate the semantic network concepts to the source documents and individual sentences.
- Provides automatic creation of a hierarchical topic structure that represents the semantics of the analyzed text, with the more important subjects placed closer to the root of the tree structure.
- Weak relationships in the semantic network are identified, which allows users to group documents and topics.

Primus Knowledge Solutions, Inc.

Primus® Answer Engine indexes information in documents and databases to provide Question Answering search retrieval, employing natural language processing technologies. The software also provides the capability to identify and analyze patterns of use by knowledge seekers for the purpose of continuous improvement. Answer Engine is one of a suite of products, including eServer (knowledgebase technology), Interchange (communications management), and eSupport (web-based front end for customer self-service solutions). This software appears to be well suited for call centers, help desks, and other support organizations. This software was developed over approximately 20 years at Princeton University.

Capabilities and Benefits

- Answer Engine reports on queries that were successfully answered as well as those that were not.
- Primus® Answer Engine allows searching of information in any of 225 document formats.
- Answers to questions are highlighted within the content.
- Integrates with an organization's existing search engines.
- Collects user feedback to rate answers, based on user initiation and also uses results relevancy thresholds to determine if the question was answered correctly.
- Answer Engine runs as a standalone technology or in conjunction with the Primus® eServer knowledgebase and allows searches from multiple documents and databases on a variety of platforms.
- Compatible with J2EE and UNIX.
- Extracts and indexes unstructured data and then transforms the text to HTML.
- Reports on queries successfully answered as well as those that were not.
- At this time, does not maintain previous query topic context for subsequent searches.
- Supports text boxes in image files, but not OCR-based searches.

Recommind, Inc.

Recommind's MindServer™ Categorization software can automatically create taxonomies and thereby provide structure to information and identify relationships amongst that information. The software does not depend on a training set of data to learn how to categorize the information. Concepts can be placed into multiple categories (or aspects) that aid in disambiguation. MindServer™ Categorization can handle structured, unstructured, and expert (taxonomies, categories, metadata) data sources. MindServer™ Categorization integrates with Recommind™ Information Retrieval Solution so that knowledge seekers can browse and retrieve related information.

Capabilities and Benefits

- Extraction, categorization, and retrieval capabilities use Probabilistic Latent Semantic Indexing (PLSI) and statistical algorithms.
- Handles multiple taxonomies on the same data.
- Imports existing or custom-built taxonomies or automatically generates a taxonomy from the text without the need for a data training set.

- Allows manual intervention to modify the taxonomies and make recommendations for new categories.
- Manages multiple file repositories in multiple locations.
- Handles over 300 file formats.
- Offers an API.
- The software is compatible with Java, and C++ oriented software applications, and runs on both Unix and NT operating environments.
- MindServer™ Retrieval does not require taxonomies or structures for retrieval of text and structured information.
- User can determine how the search results are displayed.

Saffron Technology, Inc.

SaffronNet is a knowledge discovery application that identifies relationships between entities (people, places, things, and behaviors). Discovery of these relationships and why they are significant is enhanced by the software's ability to score documents based on their relevance to a specific task. In addition, the software is able to learn the context and conditions that form these very detailed relationships. SaffronOne enables the application to process large sets of complex data, automatically updating entities and relationships in the network. SaffronNet also keeps track of the locations of the source documents that were used to identify these relationships so that users can access these source documents.

Capabilities and Benefits

- "Saffron associative memories" are created and managed in the software. These associative memories act like storage mechanisms that allows applications to read and write information. An associative memory is not similar to a highly structured relational database that uses indexed-based searching. Instead, Saffron stores associations representing the co-occurrence of items in a particular context.
- Provides a visualization tool where users can control how the information is displayed and what relationships have more significance.
- Users do not have to create complex Boolean searches, but instead can use information contained in documents to create a new search to find other meaningful relationships in the network.
- Information can be parsed using Saffron technology or the company's existing extraction tools.

SPSS LexiQuest Ltd.

LexiQuest Categorize extracts textual information from target documents and uses this information to categorize them. To do this, the tool matches the entities extracted from the target documents against descriptions of categories organized according to a taxonomy. When matching categories are found for the documents the application allows the company to either apply meta-tags or simply forward the documents to application repository. Essentially LexiQuest Categorize allows a user to visualize content relationships of terms found in a database of documents. Once interesting relationships have been discovered, the tool allows the user to view the specific document content that describes the relationships. The taxonomy does not maintain ontological data other than synonyms.

Capabilities and Benefits

- The initial category descriptions are obtained using sets of learning documents, which serve to "teach" SPSS LexiQuest Categorize what you would like to see in each category. From these, a series of terms are found and weights assigned based on the uniqueness of the term. For example, a common word like "view" would receive a low weight as it will be found in documents assigned to many categories whereas a phrase like "fuselage" would receive higher weighting due to its more exclusive use in the aerospace category.
- Currently available dictionaries are English, French, and German.
- LexiQuest Guide is software that accepts queries in everyday language and returns results that are more accurate than keyword-based search engines by transforming ordinary questions into complex Boolean search strings.
- LexiQuest Categorize is accompanied with LexiQuest Taxonomy Manager that allows for the manual maintenance of the taxonomy used by the Categorize tool. Adding or removing terms and adjusting their weights modify the taxonomy can. Term can also have synonyms specified.
- The Categorize tool comes with an extensive SDK allowing applications to take advantage of the taxonomy.

LexiQuest Mine provides an approach for isolating, discovering and processing information by using linguistic and statistical information processing techniques. Processed information can be rendered graphically, allowing easy identification of specific events in a flow of information.

- SPSS LexiQuest Mine's core dictionary and grammars enhance the linguistic functions of the tool, where dynamic clustering algorithms allow users to interactively adapt the tool's graphics and output to their specific needs.
- Analyzes data from both internal sources and across the Internet, including HTML, XML, MS Office, PDF, plain text, and RTF.

Stratify, Inc.

The Stratify Discovery System™ automatically adds structure to unstructured data by organizing and classifying that data. Then users can retrieve information by querying the categorized data. The Discovery System™ allows manual intervention to create meaningful taxonomies of topics, analyzes documents in order to appropriately classify into the taxonomies, and uses the results for numerous business applications.

Capabilities and Benefits

- Can use an existing taxonomy, either home grown or an industry standard, and automatically extend or expand it.
- Can generate metadata for each document and map that metadata to canonical (approved) metadata.
- Allows integration of the software's classification technology into an organization's applications.
- Documents may be classified into multiple categories as long as they meet the threshold for each.
- Classifies documents in over 200 formats, including HTML, plain text, Microsoft Office, and Adobe PDF.

- Supports classification in French, German, Italian, and Spanish as well as English.
- Collects documents from file servers, the Internet, intranet sites, Lotus Notes, and Microsoft Exchange through its optional crawlers.
- Provides an optional subscription-based hierarchy of almost 15,000 business, technology, and news topics.
- Refines and optimizes taxonomies and training sets automatically.
- Imports industry-standard or customized taxonomies in XML, or directly from an organized file system or web site.
- Allows as much or as little human oversight of taxonomy creation and document classification as desired through the integrated Taxonomy Manager interface.
- Provides Boolean classifier in addition to statistical, keyword, and source classifiers, enabling the importing of taxonomies with predefined rules.
- Incorporates personalization features that learn users' interests automatically and provide them with matching documents.
- Enables collaborative workflow with topic-level security.

Text Analysis International, Inc.

VisualText is a general development environment for natural language processing, pattern matching, and more. It is an IDE for developing deep text analyzers, and features NLP++, a new programming language for quickly elaborating grammars, patterns, heuristics, and knowledge. Sample application types that VisualText can support are described below.

Sample Applications:

- Information Extraction - Systems that accurately extract, correlate, and standardize the important content of text.
- Shallow Extraction - Systems that accurately identify names, locations, dates, and other atomic features of text.
- Indexing - Indexing text for quality search capabilities on the World Wide Web and other electronic text sources.
- Filtering - Systems that are both accurate and fast, to determine if a document is relevant.
- Categorization - Systems to determine the topic of documents.
- Data Mining - Finding important nuggets of information in voluminous texts.
- Test Grading - Reading and matching prose tests with idealized answers.
- Summarization - Building a brief, accurate description of the contents of a text.
- Automated Coding - Resume processing, medical reports, and police reports, for example.
- Natural Language Query- The ability to ask a computer questions using plain text.
- Dissemination - Routing documents to people or locations that require them.

Capabilities and Benefits

- Provides a rich set of tools to facilitate the knowledge engineering process.
- Implements the learning phase as a knowledge engineering effort conducted by humans and facilitated by the software.

- Provides a simple process to gather texts for a training corpus to use during the learning phase.
- It is possible to tailor the behavior of the system precisely to the desired preferences for interpreting the training corpus.
- Includes a tool to assist with the formulation of production rules. By providing a list of examples, or by highlighting examples in text, you can request that VisualText generate a rule that recognizes the examples.
- Navigation is simple between views of text, views of a parse tree, and the rule that produced a parse tree vertex.

Applications of NLP-Based Technologies

With the diversity of work at Sandia National Laboratories and the reliance on accurate and complete information to accomplish that work, more innovative, automated methods to search and retrieve knowledge are needed. By exploiting NLP-based technologies, Sandia can fundamentally change how it generates, stores, uses, and manages information. These technologies can make the automation of manual processes achieveable and create previously unattainable applications for using knowledge in our everyday tasks. Listed below are just a few possibilities for applying NLP based technologies across Sandia.

Area	Specifics
Administration	• Track contract compliance • Determine records management disposition categories • Answer questions from Corporate Business Rules (CPRs), Capital Equipment User's Guide, etc. • Visually represent funding allocation • Summarize reports for Review & Approval process • Create abstracts for published reports
Analysis	• Allow knowledge workers to more efficiently find solutions to problems • Rapidly formulate more informed make/buy decisions • Identify inconsistencies or discrepancies in data
Engineering	• Allow engineers to identify root cause of stockpile surveillance problems • Provide traceability to show pedigree for weapon certification • Identify inconsistencies and weaknesses in engineering processes • Federated access to information contained in multiple Knowledge Preservation repositories • Answering questions for weapons intern training participants

Information Management	• Provide consistent Sandia/domain-specific linguistic information • Allow federated access to Sandia's information repositories • Determine metadata for describing Web FileShare documents • Create and populate taxonomies for SAND reports, CPRs, manuals and guides, etc.
Security/Counter Intelligence	• Identify potential aggregated classified information • Enhance analyst's ability to identify security vulnerabilities under certain conditions • Display structure of access control groups to aid counter intelligence activities • Assist in determining NTK based on document content • Define data extraction parameters and extract data from large volumes of unstructured text, such as newspapers, journals, and websites, to populate a database for further analysis

Table 2: Examples of how NLP-based Technologies could be employed at Sandia

Conclusions

NLP-based software is an emerging commercial industry that has viable technologies in each of the identified categories. These technologies have wide applicability to Sandia National Laboratories and to the Nuclear Weapons Complex. Based on our survey, we believe that Sandia would greatly benefit from these technologies to search for information buried in textual documents, correlate this information with other structured or non-structured data, and populate databases for analysis.

In fact, these technologies are so fundamental to enabling a leap forward in how we search and view enterprise knowledge, that we can only begin to envision the potential applications. In addition to document and information management, other applications may include decision support, situational awareness, and vulnerability analysis. As these applications are explored, we are confident that even more uses will emerge.

This survey allowed us a peek into the possibilities of employing these technologies. Future steps would include software prototypes in each technology category for specific scenarios, such as topically categorizing Web Fileshare contents, semantically indexing and summarizing the transcripts of videos managed by the Knowledge Preservation Project (KPP), integrating KPP information with data in corporate databases, and collaborating with Technical Library staff members to develop a central ontology of Sandia's terminology and technologies.

Even though full exploitation of our information repositories is a corporation- or complex-wide issue, implementation of these technologies could be started in smaller segments, such as with the NWSBU repositories of information. Finally, as with other high-value propositions, deployment of these technologies will require a multi-year commitment to invest in the resources necessary to procure and support these capabilities, so that they can be used to fundamentally improve how Sandians interact with and gain knowledge from the information contained within our corporate repositories.

Glossary

Anaphoric reference. Referring back to or substituting for a previous word or group of words, such as substituting “he” for “John Smith.”

Categorization. Associating a document with one or more categories within one or more predefined or automatically generated taxonomies or hierarchical structures, based on the main topics in the document.

Clustering. 1) Unsupervised discovery or grouping of related information or facts within documents, to reveal knowledge or previously unknown relationships. 2) The use of repelling and attracting forces to identify documents that are more or less alike. Clusters of related information may be visually documented.

Data Mining. Advanced analytic techniques and technologies, including association, sequence or path analysis, classification, clustering, and forecasting, that find hidden and potentially useful information, patterns, and relationships amongst structured databases and data files.

Disambiguation. Removing the possibility that a word or phrase could have more than one meaning or interpretation.

Information Extraction. Identifying and selecting nouns and noun phrases, which may be keywords, relationships, concept, and facts, from unstructured text. Examples include people, places, events, and organizations. The extracted text can then be subsequently used for activities that require structured input, such as categorization, link analysis, visualization, and answering questions.

Knowledge Discovery. Revealing of previously unknown insights and understanding from apparent disjointed or incohesive data.

Knowledge Management. Processes through which organizations create, disseminate, and utilize intellectual assets, often realized through the use of computerized tools.

Knowledge Mining. Extends data mining concepts to include identifying and analyzing relationships amongst data and unstructured text in order to reveal knowledge.

Knowledge Construction/Population. Manual, semi-automatic, or automatic construction and/or population of an ontology, thesaurus, dictionary, syntactic structure, or other meta representation.

Content Tracking. Notifying interested parties if certain information changes. Tracking systems may include user profile capabilities.

Lexicon. A repository of terms, expressions, concepts, and relationships within a specialized field or subject area.

Morphology. The patterns of word formation, including inflection, derivations, and composition.

Natural Language Processing (NLP). Software development technologies that allow common forms of human communication to interface with electronic information, such as voice or everyday language instead of computer languages. Examples include spoken-language interfaces to databases and plain English question-answer technologies.

NLP Development Tools. Natural language processing software that includes APIs or development kits for developing new solutions by using or integrating them into other applications.

Non-Structured Data. Text that lacks a defined arrangement, organization, or content, such as email documents and reports.

Ontology. An organized set of concepts, rules, and relationships that specifies and describes a domain.

Precision. Degree of accuracy for search results.

Query Analysis. Part of question answering technology that interprets the user's natural language question and performs the query based on that interpretation.

Question Answering. A category of natural language processing-based software technologies that retrieves information at the sentence level in response to a natural language question, sometimes logically inferring that which is beyond what is explicitly stated in the text.

Routing. Categorizing the content of a text, mapping that content to the appropriate recipient, and directing the text to the recipient usually through some kind of predefined workflow. An example is routing email inquiries to the appropriate person or department in a software company.

Structured Data. Text that is organized into a defined arrangement or structure and has constraints on the content, such as databases and flat files.

Semantics. The meaning, or an interpretation of the meaning, of words and phrases.

Semantic Network. A set of interconnected concepts extracted from text, where the concepts and the relationships between the concepts are weighted by their relative importance.

Summarization - Document. Natural language processing based software technology that produces a condensed version of a text by excerpting key points from that text in the form of phrases, sentences, partial paragraphs, or full paragraphs, enabling readers to browse quickly though volumes of information.

Summarization - Knowledge. Capsulates extracted text that is most significant to a person's query.

Syntax. Principles or rules guiding the formation of grammatical sentences in a language.

Taxonomy. A hierarchal framework or structure used to classify and arrange objects, such as terms or keywords.

Text Mining. A broad term used for the application of statistical and/or natural language processing based software technologies to analyze and organize large sets of text. Examples include information extraction, categorization, visualization, and machine learning.

Translation. Rendering text from one language to another so that both versions have the same meaning.

Visualization. Graphical representations of related information to provide viewers understanding that is not as easily achieved with textual representations. Examples include star trees, time lines, and hierarchical structures.

Appendix A: Software Technologies by Category

1. Categorization (Includes Clustering & Routing)

a. Analyst Workbench	(SRA International, Inc.)
b. DataSet	(Intercon Systems Ltd.)
c. Insight Discover	(TEMIS-GROUP)
d. Insight Discoverer™ Clusterer	(TEMIS-GROUP)
e. Inxight Categorizer	(Inxight Software, Inc.)
f. KM suite	(SRA International, Inc.)
g. LexiQuest Categorize	(SPSS, Inc.)
h. Lextek Profiling Engine SDK	(Lextek International)
i. MatchPoint	(TripleHop Technologies)
j. MasterText	(InsightSoft-M)
k. MindServer Categorization	(Recommind Inc.)
l. MindServer Software Developer's Kit (SDK)	(Recommind Inc.)
m. Qclassifier	(Quinary)
n. RouteX Document Classifier and Router	(Lextek International)
o. SKILL CARTRIDGES™	(TEMIS-GROUP) [Categorization Dictionary and Rule Templates]
p. Stratify Classification Server	(Stratify, Inc.)
q. Stratify Discovery System	(Stratify, Inc.)
r. TextAnalyst	(Megaputer Intelligence Inc.)
s. TextAnalyst COM	(Megaputer Intelligence Inc.)
t. TopicalNet's Classifier	(TopicalNet, Inc.)
u. TopicalNet Network Gatherer	(TopicalNet, Inc.)
v. UltraSeek Classification Solutions	(Verity, formerly Inktomi)
w. VisualText	(Text Analysis International, Inc.)

2. Content Tracking

a. askOnce	(Xerox Corporation - MKMS)
b. Copernic Enterprise Search	(Copernic Technologies, Inc.)

3. Information Extraction

a. Aerotext	(Lockheed Martin Management & Data Systems)
b. Alembic	(The MITRE Corporation)
c. Alembic Workbench	(The MITRE Corporation)
d. Analyst Workbench	(SRA International, Inc.)
e. Attensity	(Attensity Corporation)
f. BBN IdentiFinder	(BBN Technologies)
g. FASTUS	(SRI International – AIC)
h. InfoXtract	(Cymfony, Inc.)
i. Insight Discoverer™ Extractor	(TEMIS-GROUP)
j. Inxight LinguistX Platform	(Inxight Software, Inc.)
k. Inxight SmartDiscovery	(Inxight Software, Inc.)
l. Inxight Thing Finder	(Inxight Software, Inc.)
a. Knowledgist	(Invention Machine Corp.)
b. MasterText	(InsightSoft-M)
c. NetOwl Extractor	(SRA International, Inc.)
d. PowerIndexer	(Xanalys Incorporated)
e. Quenza	(Xanalys Incorporated)
f. Syntalex Engine API	(Context Ltd)
g. TextAnalyst	(Megaputer Intelligence Inc.)
h. TextAnalyst COM	(Megaputer Intelligence Inc.)
i. TextPro	(SRI International – AIC)

j. VantagePoint	(Search Technology, Inc.) *Works with text data from bibliographic databases*
k. VisualText	(Text Analysis International, Inc.)
l. Xerox Terminology Suite (XTS)	(Xerox Corporation - MKMS) [Multilingual]

4. Information Search / Retrieval

a. askOnce	(Xerox Corporation - MKMS) [Federated] [Multilingual]
b. ConSearch	(Management Information Technologies, Inc.)
c. Copernic Enterprise Search	(Copernic Technologies, Inc.) [Federated]
d. DataSet	(Intercon Systems Ltd.)
e. Desktop Knowledge	(Knowledge Management Software, Inc.)
f. Domain Knowledge	(Knowledge Management Software, Inc.)
g. EasyAsk Precision Search	(EasyAsk Inc.)
h. EasyAsk Search Advisor	(EasyAsk Inc.) [Guided]
i. Enterprise Search Server	(Intelliseek, Inc.) [Federated]
j. Information Retrieval	(The MITRE Corporation)
k. InQuira 6	(InQuira, Inc.)
l. MasterText	(InsightSoft-M) [Federated]
m. MatchPoint	(TripleHop Technologies) [Federated]
n. MindServer Retrieval	(Recommind Inc.) [Federated]
o. Online Miner	(TEMIS-GROUP) [Federated] [Multilingual]
p. Portals Division	(Knowledge Management Software, Inc.)
q. Readware Information Processor	(Management Information Technologies, Inc.)
r. Readware SDK	(Management Information Technologies, Inc.)
s. TextAnalyst	(Megaputer Intelligence Inc.)
t. TextAnalyst COM	(Megaputer Intelligence Inc.)
u. TopicalNet Network Gatherer	(TopicalNet, Inc.) [Federated]

v. Universal Knowledge Suite	(Knowledge Management Software, Inc.)
w. Verity UltaSeek Products	(Verity, formerly Inktomi)

5. Knowledge Generation (Reports)

a. Knowledgist	(Invention Machine Corp.)
b. TextRoller	(InsightSoft-M)

6. Knowledge Organization Structure Construction & Management

(i.e. Ontology, Taxonomy, Thesaurus, Rule Base, etc.)

a. UltraSeek Classification Solutions	(Verity, formerly Inktomi)
b. KM suite	(SRA International, Inc.)
c. MindServer Categorization	(Recommind Inc.) [Automatic]
d. Readware Knowledge Workshop	(Management Information Technologies, Inc.)
e. Stratify Discovery System	(Stratify, Inc.) [Automatic]
f. Stratify Taxonomy Manager	(Stratify, Inc.) [Automatic]
g. Xerox Terminology Suite (XTS)	(Xerox Corporation - MKMS) [Multilingual]

7. Link Analysis

a. SaffronNet	(Saffron Technology, Inc.)
b. SaffronOne	(Saffron Technology, Inc.)

8. NLP Development Tools

a. Aerotext	(Lockheed Martin Management & Data Systems)
b. Brevity Document Summarizer	(Lextek International)
c. Inxight LinguistX Platform	(Inxight Software, Inc.)
d. Inxight Star Tree	(Inxight Software, Inc.)
e. Inxight Summarizer	(Inxight Software, Inc.)
f. Inxight Thing Finder	(Inxight Software, Inc.)

g. Lextek Profiling Engine SDK (Lextek International)

h. MindServer Software Developer's Kit (SDK)(Recommind Inc.)

i. Miscellaneous tools (New York University, CSc Dept.)

j. Readware SDK (Management Information Technologies, Inc.)

k. RouteX Document Classifier and Router (Lextek International)

l. SABLE (New York University, CSc Dept.)

m. Stratify Classification Server (Stratify, Inc.)

n. Syntalex Engine API (Context Ltd)

o. TAIParse (Text Analysis International, Inc.)

p. TextAnalyst COM (Megaputer Intelligence Inc.)

q. Tipster Architecture (New York University, CSc Dept.)

r. VisualText (Text Analysis International, Inc.)

s. XeLDA (Xerox Corporation - MKMS)

9. Question Answering

a. CoBrain (Invention Machine Corp.)

b. Discern's Dynamic Context (Discern Communications, Inc.)

c. GoldFire Intelligence (Invention Machine Corp.)

d. InFact (Insightful Corp.)

e. InQuira 6 (InQuira, Inc.)

f. Primus Answer Engine (Primus Knowledge Solutions, Inc.)

g. Question Answering System (Cymfony, Inc.)

10. Summarization

a. Brevity Document Summarizer (Lextek International)

b. Copernic Enterprise Search (Copernic Technologies, Inc.)

c. Copernic Summarizer (Copernic Technologies, Inc.) (Uses Extractor)

d. Extractor	(Institute for Information Technology)
e. Inxight Summarizer	(Inxight Software, Inc.)
a. Knowledgist	(Invention Machine Corp.)
b. KM suite	(SRA International, Inc.)
c. MasterText	(InsightSoft-M)
d. Analyst Workbench	(SRA International, Inc.)
e. TextAnalyst	(Megaputer Intelligence Inc.)
f. TextAnalyst COM	(Megaputer Intelligence Inc.)
g. VisualText	(Text Analysis International, Inc.)

11. Text Mining

a. Knowledgist	(Invention Machine Corp.)
b. LexiQuest Mine	(SPSS, Inc.)
c. PolyAnalyst	(Megaputer Intelligence Inc.)

12. Translation

a. Xerox Terminology Suite (XTS)	(Xerox Corporation - MKMS)

13. Visualization

a. Analyst Workbench	(SRA International, Inc.)
b. Inxight Star Tree	(Inxight Software, Inc.)
c. Inxight VizServer	(Inxight Software, Inc.)
d. LexiQuest Mine	(SPSS, Inc.)
e. Online Miner	(TEMIS-GROUP)
f. PolyAnalyst	(Megaputer Intelligence Inc.)
g. Quenza	(Xanalys Incorporated)
h. Semio (now Stratify)	(Stratify, Inc., formerly Semio Corporation)

i. Stratify Discovery System (Stratify, Inc.)

j. TopicalNet Network Gatherer (TopicalNet, Inc.)

k. Watson (Xanalys Incorporated)

14. Web Information Retrieval

a. Copernic Agent (Copernic Technologies, Inc.)

b. Fetch Agent Platform (Fetch Technologies, Inc.)

c. TextAnalyst for MSIE (Megaputer Intelligence Inc.)

d. WebAnalyst (Megaputer Intelligence Inc.)

Appendix B: Software Technologies by Company / Organization

1. Applied Semantics, Inc.

2. Attensity

a. Attensity	(Information Extraction)

3. BBN Technologies

A Verizon Communications company

a. BBN IdentiFinder	(Information Extraction)

4. BTexact Technologies

5. Center for Natural Language Processing – Syracuse University

6. Clairvoyance Corporation

7. ClearForest Corp

8. The Condilla Group - University of Savoie

9. Context Ltd (Syntalex)

a. Syntalex Engine API	(NLP Development Tools, Information Extraction)

10. Copernic Technologies, Inc.

a. Copernic Agent	(Distributed Web Information Retrieval)
b. Copernic Enterprise Search	(Document Summarization, Federated Information Search / Retrieval, Content Tracking)

c. Copernic Summarizer	(Document Summarization) (Uses Extractor)

11. Cymfony, Inc.

a. InfoXtract	(Information Extraction)
b. Question Answering System	(Question Answering)

12. Discern Communications, Inc.

An SRI AIC spinoff

a. Discern's Dynamic Context	(Question Answering)

13. EasyAsk Inc.

a. EasyAsk Precision Search	(Information Search / Retrieval)
b. EasyAsk Search Advisor	(Guided Information Search / Retrieval)

14. Fetch Technologies, Inc.

a. Fetch Agent Platform	(Web Information Retrieval Agents)

15. H5 Technologies Incorporated

16. Institute for Information Technology, National Research Council of Canada

a. Extractor	(Document Summarization)

17. Invention Machine Corporation

a. CoBrain	(Question Answering)
b. Scientific Effects Knowledge Base	[Over 8,000 animated effects and examples to aid in problem solving.]
c. GoldFire Intelligence	(Question Answering)
d. Knowledgist	(Information Extraction, Summarization, Text Mining)

e. TechOptimizer	["Weak Link" Analysis of systems or processes for Engineering.]
f. TechOptimizer Function Analysis & Trimming	[Core Problem Analysis of systems or processes for Engineering.]

18. InQuira Inc.

a. InQuira 6	(Information Search / Retrieval, Question Answering)

19. Insightful Corp.

a. InFact	(Question Answering)

20. InsightSoft-M (A Russian company.)

a. iExactAnswer	(Information Extraction, Federated Information Search / Retrieval)
b. MasterText	(Document Summarization, Knowledge Clustering, Information Extraction, Federated Information Search / Retrieval)
c. TextRoller	(Knowledge / Report Generation)

21. Intelliseek, Inc.

a. Enterprise Search Server	(Federated Search, etc…)

22. Intercon Systems Ltd.

a. DataSet	(Document Categorization, Information Search / Retrieval)

23. Invention Machine Corporation

a. CoBrain	(Question Answering)
b. Scientific Effects Knowledge Base	[Over 8,000 animated effects and examples to aid in problem solving.]
c. GoldFire Intelligence	(Question Answering)

d. Knowledgist	(Information Extraction, Summarization, Text Mining)
e. TechOptimizer	["Weak Link" Analysis of systems or processes for Engineering.]
f. TechOptimizer Function Analysis & Trimming	[Core Problem Analysis of systems or processes for Engineering.]

24. Inxight Software, Inc. (A Xerox PARC spinoff company.)

a. Inxight Categorizer	(Document Categorization)
b. Inxight LinguistX Platform	(NLP Development Tools, Information Extraction)
c. Inxight SmartDiscovery	(Information Extraction)
d. Inxight Star Tree	(NLP Development Tools, Visualization / Visual Navigation)
e. Inxight Summarizer	(NLP Development Tools, Document Summarization)
f. Inxight Thing Finder	(NLP Development Tools, Information Extraction)
g. Inxight VizServer	(Visualization / Visual Navigation)

25. Knowledge Management Software, Inc.

a. Desktop Knowledge	(Information Search / Retrieval)
b. Domain Knowledge	(Information Search / Retrieval)
c. Portals Division	(Information Search / Retrieval)
d. Universal Knowledge Suite	(Information Search / Retrieval)

26. Lextek International

a. Brevity Document Summarizer	(NLP Development Tools, Document Summarization)
b. Lextek Profiling Engine SDK	(NLP Development Tools, Document Categorization / Routing)
c. RouteX Document Classifier and Router	(NLP Development Tools, Document Categorization / Routing)

27. Lockheed Martin Management & Data Systems

a. Aerotext	(Information Extraction)

28. Management Information Technologies, Inc. (MITi) (a.k.a. Readware)

a. ConSearch	(Information Search / Retrieval)
b. Readware Information Processor	(Information Search / Retrieval)
c. Readware SDK	(NLP Development Tools, Information Search / Retrieval)
d. Readware Knowledge Workshop	(Knowledge Organization Structure Construction/Population)

29. Megaputer Intelligence Inc.

a. PolyAnalyst	(Text Mining, Visualization)
b. TextAnalyst	(Document Categorization / Clustering, Information Extraction, Information Search / Retrieval, Document Summarization)
c. TextAnalyst COM	(NLP Development Tools, Document Categorization / Clustering, Information Extraction, Information Search / Retrieval, Document Summarization)
d. TextAnalyst for MSIE	(Web Information Retrieval)
e. WebAnalyst	(Web Information Retrieval)

30. The MITRE Corporation (Natural Language Processing Group)

a. Alembic	(Information Extraction)
b. Alembic Workbench	(NLP Development Tools, Information Extraction)
c. Information Retrieval	(Information Search / Retrieval)

31. New York University, Computer Science Department (The Proteus Project)

a. SABLE	(NLP Development Tools) (SABLE: Scalable Architecture for Bilingual Lexicography) (Only referenced from this site, not developed there)
b. Tipster Architecture	(NLP Development Tools) (Only referenced from this site, not developed there)
c. Miscellaneous tools	(NLP Development Tools)

32. Nuance Communications

Develops Speech Processing Solutions

33. Open Text Corporation

34. Plugged In Software

35. Primus Knowledge Solutions, Inc.

a. Primus Answer Engine	(Question Answering) (Uses Inxight's LinguistX Platform.)
b. Quinary	
c. Qclassifier	(Document Categorization) (Uses Inxight and Open Text Corp. technologies)

36. Quinary

37. Recommind Inc.

a. MindServer Categorization	(Automatic Knowledge Organization Structure Construction/Population, Categorization)
b. MindServer Retrieval	(Information Search / Retrieval, Categorization of search results)

c. MindServer Software Developer's Kit (SDK)(NLP Development Tools, Categorization)

38. Saffron Technology, Inc.

a. SaffronNet	(Knowledge discovery application that accesses existing systems, internal and external data sources, and applications)
b. SaffronOne	(Real-time learning engine that analyzes massive amounts of data and performs analytics to expose the key patterns in data)

39. Search Technology, Inc.

a. VantagePoint	(Information Extraction) (Works with text data from bibliographic databases)

40. SPSS, Inc.

a. LexiQuest Categorize	(Document Categorization)
b. LexiQuest Mine	(Text Mining, Visualization)

41. SRA International, Inc.

a. Analyst Workbench	(Information Extraction, Summarization, Categorization, Visualization)
b. KM suite	(Automatic Knowledge Organization Structure Construction/Population, Summarization, Categorization) (Doesn't appear to be a single product.)
c. NetOwl Extractor	(Information Extraction)
d. NetOwl Summarizer	(Document Summarization)

42. SRI International - Artificial Intelligence Center (AIC)

a. FASTUS	(Information Extraction)
b. TextPro	(Information Extraction) (For Macintosh computers)

43. Stratify, Inc.

a. Stratify Classification Server	(Categorization, NLP Development Tools)

b. Stratify Discovery System	(Automatic Knowledge Organization Structure Construction/Population, Categorization, Visualization/Category Browsing)
c. Stratify Taxonomy Manager	(Knowledge Organization Structure Construction/Management)

44. TEMIS SA (TEMIS-GROUP)

a. Insight Discoverer™ Categorizer	(Document Categorization) (Uses XeLDA, the Xerox Linguistic Development Architecture)
b. Insight Discoverer™ Extractor	(Information Extraction) (Uses XeLDA, the Xerox Linguistic Development Architecture.)
c. Insight Discoverer™ Clusterer	(Document Categorization) (Uses XeLDA, the Xerox Linguistic Development Architecture.)
d. Online Miner	(Federated Multilingual Information Search / Retrieval, Visualization) (Uses XeLDA, the Xerox Linguistic Development Architecture)
e. SKILL CARTRIDGES™	(Knowledge Organization, Categorization) (Categorization Dictionary and Rule Templates)

45. Text Analysis International, Inc. (TextAI)

a. VisualText	(NLP Development Tools, Information Extraction, Categorization, Summarization)
b. TAIParse	(NLP Development Tools)

46. Textology

a. Textology Categorizer

b. Textology Extractor

c. Textology Summarizer

d. Textology Concept Linker

e. Textology Meta Tagger

47. TopicalNet, Inc.

a. TopicalNet's Classifier	(Document Categorization)

b. TopicalNet Network Gatherer	(Federated Information Search / Retrieval, Document Categorization, Visualization)

48. TripleHop Technologies

a. MatchPoint	(Federated Information Search / Retrieval, Document Categorization)

49. Verity

a. UltraSeek	(Search / Retrieval, Document Categorization)

50. Xanalys Incorporated

a. ChartViewer	(Visualization) (Distributable viewer for Watson link-charts)
b. PowerCase	(Case Management for Investigators)
c. PowerIndexer	(Information Extraction)
d. Quenza	(Information Extraction, Visualization)
e. Watson	(Visualization)

51. Xerox Corporation - Multilingual Knowledge Management Solutions (MKMS)

a. askOnce	(Federated Multilingual Information Search / Retrieval, Content Tracking)
b. XeLDA: Xerox Linguistic Development Architecture	(NLP Development Tools)
c. Xerox Terminology Suite (XTS)	(Multilingual Information Extraction, Automatic Multilingual Knowledge Organization Structure Construction/Population, Translation)

Appendix C: Technology Company / Organization Contact Information

Technology Company / Organization	Contact information
Applied Semantics, Inc.	Applied Semantics, Inc. 2644 30th Street, 2nd Floor Santa Monica, CA 90405-3009 Phone: (310) 460-4000 Fax: (310) 450-6686 Email: sales@appliedsemantics.com Website: www.appliedsemantics.com
Attensity	Attensity 90 South 400 West, Suite 600 Salt Lake City, UT 84101 Phone: (801) 532-1125 Fax: (801) 532-1164 Website: www.attensity.com
BBN Technologies *A Verizon Communications company*	Speech & Language Processing Department BBN Technologies 70 Fawcett Street Cambridge, MA 02138 800-295-7897 Email: bbnt-marketing@bbn.com Website:www.bbn.com/speech
BTexact Technologies	B Texact Technologies Adastral Park Martlesham Ipswich IP5 3RE UK Phone: +44(0)1473 607080 Fax: +44(0)1473 607700 Website: www.btexact.com
Center for Natural Language Processing – Syracuse University	Center for Natural Language Processing School of Information Studies Syracuse University 4-206 Center for Science & Technology Syracuse, New York 13244-4100 Phone: 315-443-5484 Fax: 315-443-5806 Liz Liddy, Director, liddy@syr.edu Lois Elmore, Administrative Assistant, laelmore@syr.edu Website: www.cnlp.org/index.asp

Clairvoyance Corporation	Clairvoyance Corporation 5001 Baum Boulevard, Suite 700 Pittsburgh, PA 15213-1854 Phone 412-621-0570 Website: www.clairvoyance.com
ClearForest Corp.	ClearForest Corp. 15 E. 26th Street, Suite 1711 New York, NY 10010 Tel: 212.432.1515 Fax: 212.432.1929 Website: www.clearforest.com
The Condillac Group - University of Savoie	The Condillac Group University of Savoie - Campus Scientifique 73 376 Le Bourget du Lac cedex - France Phone: +33 (0) 4 79 75 87 79 Email: roche@univ-savoie.fr Website: ontology.univ-savoie.fr
Context Ltd (Syntalex)	Context Ltd Suite 3, Grand Union House 20 Kentish Town Road London NW1 9NR UK Tel +44 (0)20 7267 8989 Fax +44 (0)20 7267 1133 Website: www.syntalex.co.uk
Copernic Technologies, Inc.	Corporate Offices Copernic 360 Franquet #60 Sainte-Foy QC G1P 4N3 Canada Phone: 418-527-0528 Fax: 418-527-1751 Website: www.copernic.com
Cymfony, Inc.	Cymfony, Inc. 600 Essjay Road Williamsville, NY 14221 Ph: 716-565-9114 Fax: 716-565-0308 Website: www.cymfony.com
Discern Communications, Inc. *A SRI AIC spinoff*	Discern Communications, Inc. 333 Ravenswood Avenue Menlo Park, CA 94025 Phone: 650.859.5700 Fax: 650.859.3027 Email: general@discern.com Website: www.discern.com

EasyAsk Inc.	EasyAsk 119 Russell Street Littleton, MA 01460 Phone: 978-486-8860 Fax: 978-486-0868 Sales: 800-425-8200 Technical Support: 978-486-8386 Email: info@EasyAsk.com Website: www.englishwizard.com
Fetch Technologies, Inc.	Fetch Technologies, Inc. 4676 Admiralty Way, 10th Floor Marina del Rey, CA 90292 USA +1 (310) 448-9148 Sales: Sales@fetch.com Website: www.fetch.com
H5 Technologies Incorporated	H5 Technologies Incorporated 520 Third Street, Third Floor San Francisco, CA 94107 Tel: 415.625.6700 Fax: 415.625.6799 Email: info@H5Technologies.com Website: www.h5technologies.com
Institute for Information Technology (IIT), National Research Council of Canada	Institute for Information Technology National Research Council of Canada Bldg. M-50 1200 Montreal Road Ottawa, ON K1A 0R6 Canada Telephone: (613) 993-3320 Fax: (613) 952-0074 Email: ilo@iit.nrc.ca Website: iit-iti.nrc-cnrc.gc.ca
Invention Machine	Invention Machine Corporation 133 Portland Street Boston, MA 02114 Tel: 800-595-5500 (USA/Canada only) Tel: (617) 305-9250 Fax: (617) 305-9255 Email: info@invention-machine.com Website: www.h5technologies.com
InQuira Inc.	InQuira Inc. 851 Traeger Avenue, Suite 125 San Bruno, CA 94066 T: 650-246-5000 F: 650-246-5036 Email: sales@inquira.com Website: www.inquira.com

Insightful Corp.	Insightful Corp. 1700 Westlake Ave. N., Suite 500 Seattle, WA, 98109 1-800-569-0123 (206) 283-8802 Website: www.insightful.com
InsightSoft-M *A Russian Company*	Email: development@insight.com.ru Website: www.insight.com.ru
Intelliseek, Inc.	Intelliseek, Inc. 1128 Main Street, 4th Floor Cincinnati, OH 45202-7236 Phone: (513) 618-6700 Fax: (513) 618-6702 Email: infoemailintelliseek.com Website: www.intelliseek.com
Intercon Systems Ltd.	Intercon Systems Inc. 790 Penllyn Pike, Suite 302 Blue Bell, PA 19422 Phone: (888) 202-9801 Email: intercon@ds-dataset.com Website: www.ds-dataset.com
Inxight Software, Inc. *A Xerox PARC spinoff*	Inxight Software Headquarters 500 Macara Avenue Sunnyvale, CA 94085 Tel. 408.738.6200 Tel. 888.414.4949 Fax 408.738.6203 Email: sales@inxight.com Website: www.inxight.com
Knowledge Management Software, Inc.	Knowledge Powered Solutions plc Rutherford House Pencroft Way Manchester Science Park Manchester M15 6SZ UK Phone: +44 (0) 161 227 1100 Freephone: 0800 028 0022 Fax: +44 (0) 161 227 1101 Email: info@kpsol.com Website: www.kmsoftware.com

Lextek International	Lextek International 1051 Fir Provo, UT 84606 Voice: 801.375.8332 Fax: 801.373.5342 Email: sales@lextek.com Website: www.lextek.com
Lockheed Martin Management & Data Systems	Lockheed Martin Management & Data Systems PO Box 8048 Philadelphia, PA 19101 Phone: 610-531-7400 Website: mds.external.lmco.com
Management Information Technologies, Inc. (MITi) *a.k.a. Readware*	Management Information Technologies, Inc. 6 NW 27th Terrace Gainesville, Florida 32607 Phone: 877-514-5092 Fax: 352-335-6385 Email: mitioke@readware.com Website: www.readware.com
Megaputer Intelligence Inc.	Megaputer Intelligence Inc. 120 West 7th Street, Suite 310 Bloomington, IN 47404 USA Tel: (812)-330-0110 Fax: (812)-330-0150 Website: www.megaputer.com
The MITRE Corporation *Natural Language Processing Group*	The MITRE Corporation 202 Burlington Road Bedford, MA 01730 Voice: 781-271-2000 Fax: 781-271-2352 Website: www.mitre.org
New York University Computer Science Department *The Proteus Project*	The Proteus Project Computer Science Department New York University 715 Broadway, 7th floor New York, NY 10003 Tel: 212-998-3497 or 212-998-3003 Fax: 212-995-4123 Website: www.cs.nyu.edu/csweb
Nuance Communications *Develops Speech Processing Solutions*	Nuance Communications 1005 Hamilton Court Menlo Park, CA 94025 Phone: 650-847-0000 or 888-NUANCE-8 Fax: 650-847-7979 Website: www.nuance.com

Open Text Corporation	Open Text Corporation 185 Columbia Street West Waterloo, Ontario Canada N2L 5Z5 U.S.A. Headquarters 2201 South Waukegan Road Bannockburn, IL 60015 North America Sales: 1-800-499-6544 E-mail: info@opentext.com Website: www.opentext.com
Plugged In Software	Plugged In Software 800 NE Tenney Rd, Ste 110-301 Vancouver, WA 98685 Phone: (360) 828-8280 Fax: (877) 290 6687 Email: info@pisoftware.com Website: www.pisoftware.com
Primus Knowledge Solutions, Inc.	Primus Knowledge Solutions, Inc. 1601 Fifth Avenue, Suite 1900 Seattle, Washington 98101 USA Telephone: 206.834.8100 Fax: 206.834.8125 Email: sales@primus.com Website: www.primus.com
Quinary	Milan Office via Pietrasanta 14 20141 Milan – Italy T +39 (02) 3090 1500 F +39 (02) 3090 1501 Email: info@quinary.com Website: www.quinary.it
Recommind	Recommind, Inc. 1001 Camelia Street Berkeley, CA 94710, USA Tel: (510) 558-7899 Fax: (510) 525-2351 Email: info@recommind.com Website: www.recommind.com
Saffron Technology, Inc.	Saffron Technology, Inc. 1600 Perimeter Park Dr., Suite 150 Morrisville, NC 27560 Phone (919) 468-8201 Fax (919) 468-8202 Email: info@saffrontech.com Website: www.saffrontech.com

Search Technology, Inc.	Search Technology 4960 Peachtree Industrial Blvd., Suite 230 Norcross, Georgia 30071-1580 Voice (770) 441-1457 Fax (770) 263-0802 Email: info@searchtech.com Website: www.searchtech.com
SPSS LexiQuest	SPSS Inc. 233 S. Wacker Drive, 11th Floor Chicago, IL 60606 Phone: (312) 651-3000 Email: sales@spss.com Website: www.spss.com/spssbi/lexiquest
SRA International, Inc.	SRA International, Inc. 4300 Fair Lakes Court Fairfax, VA 22033 Phone: (703) 803-1500 Fax: (703) 803-1509 Website: www.sra.com NetOwl Website: www.netowl.com NetOwl Sales Phone: (703) 803-1502 NetOwl Sales E-mail: sales@netowl.com
SRI International - Artificial Intelligence Center (AIC)	Artificial Intelligence Center SRI International 333 Ravenswood Avenue Menlo Park, CA 94025-3493 Email: aic@ai.sri.com Website: www.ai.sri.com
Stratify, Inc.	Stratify, Inc. 501 Ellis Street Mountain View, CA 94043 Phone: 650-988-2000 Fax: 650-988-2159 Email: sales@stratify.com Website: www.stratify.com
TEMIS SA *TEMIS-GROUP*	TEMIS France 59, rue de Ponthieu 75008 Paris France Tel. +33 - (0)1 58 56 48 00 Fax. +33 - (0)1 45 62 21 02 info@temis-group.com Website: www.temis-group.com

Text Analysis International, Inc. *TextAI*	Text Analysis International, Inc. 1669-2 Hollenbeck Ave. # 501 Sunnyvale, CA 94087 1-408-746-9932 1-877-235-6259 Email: info@textanalysis.com Website: www.textanalysis.com
Textology	Textology, Inc. 150 Post Street, Suite 400 San Francisco, CA 94108 Tel: 415-982-0555 x104 Fax: 253-369-0264 Email: info@textology.com Website: www.textology.com
TopicalNet, Inc.	TopicalNet Corporate Headquarters 800 West Cummings Park, Suite 2900 Woburn MA 01801 phone: 781-932-8400 fax: 781-932-2558 1-866-TOPICAL (867-4225) Email: info@topicalnet.com Website: www.topicalnet.com
TripleHop Technologies	TRIPLEHOP TECHNOLOGIES 45 W. 25th Street, 9th Floor New York, NY 10010 Telephone: (212) 243-4645 ext. 3055 Fax: (212) 243-4660 Email: sales@triplehop.com Website: www.triplehop.com
Verity, Inc.	Verity, Inc. 894 Ross Drive Sunnyvale, CA 94089 Telephone: (408) 541-1500 Fax: (408) 541-1600 Email:info@verity.com Website: www.verity.com
Wordmap Ltd.	Wordmap Ltd 26 Upper Borough Walls Bath, BA1 1RH United Kingdom Tel: +44 (0)1225 358 184 Fax: +44 (0)1225 358 183 Email: contact@wordmap.com Website: www.wordmap.com

Xanalys Incorporated	Xanalys Incorporated 95 Sawyer Road Three University Park Waltham, MA 02453 Tel: +1 877 XANALYS (1 877 926 2597) Fax: +1 781 736 1949 Email: us-sales@xanalys.com Website: www.xanalys.com
Xerox Corporation *Multilingual Knowledge Management Solutions (MKMS)*	Xerox Corporation 80 Linden Oaks Parkway Rochester, New York 14625 std: 716 264 5217 Email: sales_us@mkms.xerox.com Website: www.mkms.xerox.com

Distribution:

5	0139	P. J. Wilson, 9902
1	0451	J. L. Mitchiner, 6540
1	0451	S. M. Rinaldi, 6541
1	0631	H. M. Witek, 2910
1	0661	G. E. Rivord, 9510
1	0784	R. E. Trellue, 6500
1	0801	W. F. Mason, 9320
1	0899	E. C. Moser, 9616
5	1137	G. N. Conrad, 6544
1	1137	W. R. Cook, 6544
1	1137	R. L. Froehlich, 6545
1	1137	K. L. Hiebert-Dodd, 6545
1	1137	S. L. Humphreys, 6545
60	1137	M. G. Stickland, 6544
60	1138	S. M. Eaton, 6536
1	1138	D. P. Gallegos, 6533
1	1138	M. S. Tebo, 6536
1	1138	E. R. Young, 6532
1	1140	L. J. Ellis, 6502
1	1219	O. H. Bray, 5941
1	1219	I. Dubicka, 5941
1	9018	Central Technical Files, 8945-1
2	0899	Technical Library, 9616
1	0612	Review & Approval Desk, 9612
		For DOE/OSTI