Information Intelligence: Content Classification and the Enterprise Taxonomy Practice

June, 2004

DELPHI
GROUP

111 Huntington Avenue, Suite 2750,
Boston, MA 02199

Publication and distribution of this report is partially underwritten by the technology suppliers who have engaged with Delphi Group in the Spring, 2004 Taxonomy Research & Education Program. This program is part of an ongoing series of Delphi Group studies, publications, educational programs, and conferences which seek to establish thought leadership perspectives and best practices in the area of information intelligence. Profiles of these suppliers' technologies are included in this report.

Information Intelligence:

MARKET MILESTONE REPORT

Content Classification and the Enterprise Taxonomy Practice

June, 2004

Information Intelligence

Information intelligence refers to an emerging data and semantic infrastructure which will enable organizations to create a new generation of business applications. This new class of applications will build on the rich set of assets already available inside the organization.

In facing up to the challenge of managing the organization's increasingly merged data, content, and knowledge operations, executives in government and industry are beginning to re-engineer information-based business process. The key to their success lie s in learning the capabilities and practices of the information intelligence platform.

Taxonomy and classification provide a key new practice—an opportunity to change the investment return profiles for current IT assets as well as to deliver rapid returns on new application investment.

Report Scope

This Delphi Research report is one of an ongoing series examining the state of the technology and the level of business practice in the critical and related areas of information architecture, taxonomy/classification, and search and retrieval.

In particular, this report focuses on recent research in the areas of taxonomy and classification. The objectives of this report are to:

- Define taxonomy and classification and position their functionality within an integrated information architecture
- Analyze trends of taxonomy software technology development and implementation, based on a survey of over 300 organizations
- Provide examples of how these technologies are applied to tangible and pressing business problems
- Provide an education on examples of technology approaches to taxonomy generation and automated classification
- Describe a set of current example technology offerings to provide an overview of variety of approaches available in today's taxonomy and classification software market.

It is clear from the information presented in this report that taxonomy and classification provide a key new area of business practice—an opportunity to change the investment return profiles for current IT assets as well as to deliver rapid returns on new application investment.

DELPHI®
GROUP

111 Huntington Avenue, Suite 2750,
Boston, MA 02199

"Many companies have become locked into the view that IT can reduce transaction costs but then think of transaction costs as encompassing only the transfer of bits and data from one place to another. Viewed more broadly, transaction costs encompass such challenging business issues as the creation of meaning, the building of trust, and the development and dissemination of knowledge. These dimensions of transaction costs often represent significant bottlenecks to performance improvements and competitive advantage."

John Seely Brown, John Hagel III
Harvard Business Review, 2003

Contents

Table of Figures

Introduction

An Eternal Problem

"Everyday living is too fast, too busy, too complicated. More than at any time in history, it's important to have good information on just about every aspect of life. And, there is more information available than ever before. Too much in fact. There is simply no time for people to gather and absorb the information they need."

This quote speaks for a generation. We have experienced first hand the onslaught of infoglut, countless corporate intranets, and the vast resources of the World Wide Web. Virtually everyone reading this quote can empathize. Well, it may shock you to learn that this quote is not attributed to a contemporary, but written generations ago. The quote was first uttered by Britton Hadden, shortly before he founded *Time Magazine* with Henry Luce in 1929.

The experience of difficulties with cataloging and retrieving unstructured information effectively is an eternal problem. Not only has it been a pressing issue for generations, there is also evidence to believe that we will always seek better, more efficient approaches toward the ultimate goal of perfect recall and keen precision.

Over the last decade, the software industry that has set content creation

Defining Taxonomy and Classification

The creation of an information architecture, which may include a taxonomy, an ontology, search, and/or classification is often complicated by confusion over these elements themselves. It is therefore important to start with definitions of these functions.

Taxonomy – is a hierarchial or polyhierarchical listing of *topics* or subject categories. It may not include a definition of the topics, but only the hierarchical relationship of topics to one another. A taxonomy can incorporate content from both a thesaurus and an ontology. There are no standard file formats or approaches to taxonomy construction. A taxonomy is often used to provide a structured navigational path through a content collection.

Thesaurus – is a network of words and word meanings and relationships used to put conceptual definitions into context. It defines a lexicon and the relationships between words in that lexicon. A thesaurus may be a precursor to a taxonomy, in which the leading or preferred terms in a thesaurus are used to define the taxonomy structure. Thesaurus construction is defined by ANSI standard Z39.19. A thesaurus is often used to enhance the intelligence of a taxonomy and/or search tool by providing insight into word meanings and relationships.

Ontology – is a network of relationships that are self-describing and used to track how items or words relate to one another. For example, a "lives at" link or "works for" link in an ontology would be used to track these types of relationships and their corresponding values for listed individuals. Ontology is the framework of the semantic web, and permits intelligent navigation.

Classification – is the process of analyzing content and determining where in a taxonomy it belongs. The classification process typically results in the assignment of metadata tags. For example, the assigning of a Dewey decimal code to a book, based on the mapping of its content into the dewey decimal system is a form of classification.

and distribution on steroids has also attempted to keep pace with new automated approaches to categorization and retrieval, that leverage the speed and efficiency of the electronic information infrastructure.

In a recent survey conducted by Delphi Research with over 300 companies, 59% of respondents indicated that locating and accessing the information needed to do their job has gotten simpler and more effective over the last 2 years. Yet, 68% went on to state that retrieval is still difficult and time consuming, with 62% expressing dissatisfaction with overall information retrieval efficiency.

(The full findings of this Delphi Research are presented and analyzed in the section *Taxonomy & Classification Market Analysis*, beginning on p. 16 below.)

Toward an Information Architecture

Software that enhances the performance of search engines, text mining, ontology, summarization, content classification and taxonomy construction has had dramatic impact on lessening the burden of the knowledge worker, but the perfect solution is still not within our grasp. Part of the user dissatisfaction stems from information management solutions in organizations that have developed organically over a period of years. Most organizations do not have a centralized, strategically developed information architecture. Under such an architecture, multiple layers of information management functionality are orchestrated and coordinated as a service, capable of being deployed against any number of content sources and repositories. A well defined information architecture is the antithesis of approaches that link search and management to individual repositories, causing user frustration with jumping from one research environment to another. In the recent Delphi Research survey, 82% of respondents reported that they do not have access to a centralized single point of search and management across information sources.

A well thought out, orchestrated approach to content tracking, categorization, search, and retrieval can give the illusion of content integration. Though perhaps not the only or most efficient manner in every case, the provision of a singular navigational front end (e.g. taxonomy) and omnipresent search tool that collectively aggregate disparate content resources, can, from an end-user perspective, deliver the simple single point of access that many users strive for.

But therein lies the challenge to the business and IT communities. While the business side must determine the right levels of functionality needed, IT must develop approaches that simplify the delivery of such functionality and minimize the number of front-ends. Understanding myriad sources of unstructured content (e.g. web content, e-mail and on-line files), requires the orchestration and coordination of multiple disciplines and technologies working in concert. This is the result of a well planned information architecture.

Pivotal points in any such information architecture are the selection and implementation of taxonomy tools, categorization tools, and search tools, each ranked as critical to a information architecture by a plurality of research survey respondents.

Why Taxonomy

While we will revisit the definition of these functional components and their relationships, we have to first ask a more basic question: "Why the current interest in taxonomy in many business organizations?"

As previously stated, Delphi Group's research on user experiences with using content reveals that lack of organization of information is in fact the number one problem in information management and retrieval, in the opinion of business professionals. These professionals include customer service representatives, the sales team, the financial services professionals, the R&D engineers and senior executives. If these professionals are spending 20% of their time or more looking for information (as our survey results indicate), then this results in an opportunity cost and represents a runaway expense item in many organizations.

One of the defining challenges of this era of enterprise computing is just this: How do we find the relevant and pertinent information to do our jobs and make informed business decisions? The answer is at once obvious and elusive. We must harness the computer to help manage and retrieve content at the same rate at which it allows us to create and distribute that content.

Intelligent Information Processing

Why is it so difficult to effectively mine answers from unstructured content? To illustrate the answer to this question, lets use the example of searching for information about chips. A search on the Web for "chips" (using the Google search engine) returns 2,430,000 references. Even if only 1% of these documents were relevant, that is over 24,000 documents, which is far beyond the capability of most of us to wade through. Some of the documents contain information about – chocolate chips, potato chips, buffalo chips, wood chips, poker chips, an old TV series (ChiPS), or integrated circuits. "Chips" , like so many words can have multiple meanings, and become difficult to define when taken out of context. This example is easily applied across research environments. Most words and phrases can have multiple meanings - consider this short list: "java," "can," "branches," and "boot."

People can distinguish the specific meaning of a word or concepts based on the context the word is used in, but, computers cannot. Other methods must be used by computers to provide information intelligence (e.g. for precision in retrieval.)

Context

People explore concepts; computers (without the aid of sophisticated software) primarily search for key words. Relevancy is subjective to the individual who is performing the search. Only each individual can judge how relevant a particular bit of information is to what they are attempting to discover. The document may be too technical or out of date or too general for your needs. It could be only vicariously related to the subject of research, or completely misaligned. Context is the determining factor. Machines can't distinguish between "John Smith to marry Mary Jones" vs. "Reverend Billy Graham to marry Bruce Springsteen." Only individuals who can interpret text in the proper context can understand that Reverend Billy Graham will perform the ceremony and not be the recipient of Bruce's ardor.

Consider again the example of searching on "chips." The search may have been intended to focus on integrated circuit "chips" and not a recipe to make chocolate "chip" cookies. If you had only searched in a category such as computers or electronics, you would have found fewer documents, more precise and relevant information. If you have categories and hierarchical structures of information you will be able to narrow the search field and find relevant information faster.

Ambiguity

The beauty of language is that it has many words to describe the same thing. The corollary is that the same word may have different meanings. "Chips" is one example of an ambiguous word. "Java" is another. Java could be an island, a cup of coffee or a computer programming language. "Kick the bucket" is another phrase that could have multiple meanings depending on the context. Consider the sentence, "Joe kicked the bucket and the water spilled out." The question is: did Joe die? or did Joe violently place his foot in contact with a container of water? The context of the surrounding phrases in the document will clarify this. So the question becomes, how do you automate or simplify the categorization and classification of documents that are made up of richly ambiguous words and phrases?

Browsing vs. Searching

Effective search of content without benefit of a taxonomy requires that you know the words and phrases to use *before* you see what is in the collection of documents. Key word search assumes you know what you are looking for and that is often an erroneous assumption. Knowledge workers are not always exactly sure what they are looking for but, "when they see it, they know what it is." More than 20% of the day is involved in searching for information on the knowledge worker's computer system. About 70% of that time is spent browsing for information. 75% of the people surveyed during a Yahoo market research project preferred browsing to searching.

From our previous example about "chips," you may not have known there were many types of computer "chips," such as processor "chips," application specific (ASIC) "chips" or memory "chips." If there were categories of computer "chips" such as: processor, ASIC, and memory "chips"; you may find all the information you need is about flash memory "chips." In some instances, it is easier to discover information about a particular subject if you see it in the context of related information. Browsing encourages associative thought. Browsing in categories can guide you through the information discovery process.

Browsing is not superior to search, in reality they compliment each other, and should be used interchangably, based on the type of retrieval that is required in each situation.

People search for information in two basic and different ways - **searching/locating** and **browsing/discovery**. The first is the process used when you know what you are looking for. You know the answer, now you need to find more information about the subject. Keyword search with Boolean logic and traditional search engines are good for this type of approach. The application of thesauri and other concept-based tools and approaches to word search enhance this process. Using our previous example of "chips," the query could have been constructed on a full phrase such as "chocolate chips," or, quite differently, "computer chips." This approach to concept-based search would render more precise results. But ultimately, such approaches to retrieval of content do assume that the user has some knowledge of the subject and can formulate an effective query string. These scenarios are more about search and location and less about navigation and discovery.

Browsing, the second basic approach to finding content is more focused on discovery. In these instances, often the user does not know the answer they are looking for when they begin the search for information.

Let's use the example of doing research on categorization technologies. In the sidebar at left, we provide an example of a possible hierarchy for information about categorization technologies.

The availability of this taxonomy provides the researcher with great insight into the subject. The availability of this taxonomy eliminates the need for the researcher to completed understand the subject before issuing a query. It serves as a guide to the research process—even educates. Using this example, the user starting this search process can immediately realize that there are several alternative approaches to categorization. Further research could delve into the manual approaches, for example, or focus on the alternatives, an automated or hybrid approach. Further investigation of the taxonomy reveals - before a single question is posed, that there are both pros and cons to using a each of the approaches. The taxonomy allows browsing of these issues immediately, progressively revealing areas of interest to the researcher.

Methodologies used in the automatic process of categorization may not have been understood by the researcher initially. But the tracking of

A Sample Categorization Technology Hierarchy

Categorization Approaches

- Manual
  - Advantages
    - Human judgment
    - High accuracy
    - Disambiguation
    - ...
  - Disadvantages
    - Labor intensive
    - Inability to scale
    - Expensive resources
    - ...
- Automatic
  - Advantages
    - Handles huge volumes
    - Scales easily
    - Inexpensive resources
    - ...
  - Disadvantages
    - Rule/algorith fragility
    - Inaccuracies
    - Difficult to train
    - ...
- Hybrid
  - Advantages
    - High volume + accuracy
      - Human-guided rule sets
    - Incremental learning
    - ...
  - Disadvantages
    - Management challenge
    - Special skills needed
    - Maintenance effort required
    - ...

these by the taxonomy, makes them immediately discoverable by the researcher. By learning more about the various methods of automatic categorization (i.e., rules, example learning, semantic analysis and clustering analysis) you can better understand how each of the methodologies may be applied to your particular situation.

The point is that when this research was initiated, the researcher was not thinking about cost-justifying a software solution, nor were you aware of the various methodological approaches to the problem. This information was discovered without an time-consuming query/search/re-query process, but rather immediately discovered within the structure of the taxonomy. Browsing via a taxonomy in essence provides an education on the subject and lends insight into the issues or facets of the subject.

> Visual, hierarchical arrangements of subject categorization trigger associations and relationships that are not obvious when initially searching in most instances. This distinction is important and implies yet another reason why categorization can be critically important to the productivity of knowledge workers.

Visual, hierarchical arrangements of subject categorization trigger associations and relationships that are not obvious when initially searching in most instances. This distinction is important and implies yet another reason why categorization can be critically important to the productivity of knowledge workers. With the advent of web sites and portals, the usefulness of taxonomies has become even more powerful and popular. The taxonomy can expose the inner structure of a site or portal. It allows the user (e.g. a customer to a retail web site) to learn about or appreciate the content of the site. It provides an immediate education on the content and functions available on the site, and can guide the user through the site. This can be critical to the successful experience of a first-time visitor to the site.

An intrinsic benefit of the hierarchical structure of categorization is that links and summaries of information are rendered in the context of their unique “parent-child” relationships. Relevant information is more likely to be found when specific content foci are employed. Browsing the taxonomy provides this benefit in three distinct settings—dynamic, interactive/iterative and educational.

Browsing is dynamic. Information changes all the time. Virtually any search on a complex topic becomes a hunt for a moving target. In today’s business environments, information can change daily, weekly, monthly—or without warning. Changes to existing content are made, new content is added, dated content deleted. Indeed, respondents to the Delphi Research survey indicated that the number one source of frustration with search of on-line content is the fact that the content they search for is constantly changing, which both frustrates the user and reduces the effectiveness of simple search. Use of a taxonomy can provide a dynamic bookmark so to speak, a one-stop-shopping guide to all relevant content on a subject. Return to a subject node exposes the latest and complete collection of content about that subject area—each time the node is opened. This is a most powerful application of taxonomy, as it addresses what survey respondents cited as the number one cause of research frustration, the dynamic, volatile nature of information sources.

Browsing is also an interactive process. Navigation of a well-designed interface to information on a web site/portal automatically directs the researcher to other relevant topics. A search and browse through information about categorization software, for example, will uncover reviews, analysis, white papers and commentaries with information about other technologies, companies or related topics of information that may be worth investigating. In fact, if taxonomies were built on each facet of a site (e.g. one on subject matters the other on services and resources), the linking of the taxonomies in a matrix could provide cross selling or expanded education. For example, a user who discovers communities of practice regarding automated taxonomy tools via an initial investigation into these tools, could be exposed to and linked to other communities of practice available at this site on other related matters.

Navigation of a well-designed interface to information on a web site/portal, automatically directs the researcher to other relevant topics.

Information architects can indentify the major FACETS (subject matter areas) of a site and construct unique taxonomies or classifications to serve as navigation guides for each facet.

Browsing can also be an iterative and educational process. Repeating the process refines the focus while deepening the knowledge. Accessing relevant information and interrelated ideas and concepts supports a fundamental change in user focus and activity—from simply searching, to finding and discovering.

Towards an Integrated Information Architecture – Taxonomy and Search

Whether using navigation for discovery, research, education or using search for targeted retrieval, business users turn to search and taxonomy to find relevant information quickly and intuitively, to support better informed and hence more effective decisions and actions. Equipping enterprise knowledge workers with the tools to make smarter decisions is a strategic imperative in today's economy. Jakob Nielsen, the guru of usability, estimates that poor classification costs a 10,000 user organization $10M annually.

Jakob Nielsen, the guru of usability, estimates that poor classification costs a 10,000 user organization $10M annually.

To this end solution design needs to consider multiple ways of providing a taxonomy, and integrating it with search functionality. In so doing, multiple research environments can be supported.

Front-end, Back-end, To What End?

Maximum return on the investment of building a taxonomy does not come from viewing it as a stand-alone tool, but as a integrated strategically leveraged module of an integrated information architecture. Taxonomies can and should be integrated with other applications. New approaches to integrating taxonomies with search functions can provide high leverage and innovative features in these archtitecture projects.

In the most basic approach, a taxonomy can exist side-by-side, stand-alone to a search tool as a separate investigative alternative. The search tool provides content-based targeted search, the taxonomy a navigation path of discovery. Each function exists as an alternative to the user to support distinct approaches to research. It should be noted, however, that even in this "separate but equal" approach, taxonomy technology, and related thesaurus technology can be leveraged by the search engine to provide more accurate search results. The ability for search engines to search not just on keywords, but also on implied concepts and ideas can be implemented through the integration of lexical and linguistic techniques and knowhow captured in a thesaurus and or taxonomy.

The taxonomy can also be leveraged as a front-end to search. A user who is somewhat naive about an overall subject area and its many facets might begin the research process by navigating through a taxonomy. When a particular node of interest is discovered, a subsequent search could be executed—this time against only the content in this particular node of the taxonomy—to drill down and locate sources which contain occurrences of specific words, phrases and/or concepts. In this manner, a broad search is narrowed (precision increased) by discovering sub-topics of focus through the taxonomy. Once embedded in this area, search is used to further refine the investigation to a particular issue.

Conversely, a taxonomy could be integrated with a search tool as a back-end interface. Here the user might start with a broad-based word search. The results of the search would not be a uninformative listing of thousands of "relevant" sources. Through integration of a taxonomy, the results of the query could be displayed as a customized set of folders (derived from a taxonomy or a dynamic classification), which organize the content by related subtopics. This dynamically organized presentation interface—based on the integration of search and classification—provides the researcher with further insight regarding how the topic of focus is broken down into subtopics.

The Challenge of Taxonomy Design and Construction

The strategic deployment of a taxonomy in an organization's overall information architecture can provides many enhancements to information work. But at what cost? Taxonomy design and construction is not without a cost in technology resources, and, more important, in skilled human resources needed to develop the practice.

The concept of taxonomy evolved from the life sciences. In the scientific community, taxonomies were conceived as a way to organize and categorize life forms into a structured and controlled hierarchy. In this approach, a plant or animal is placed in a *single* spot describing its hierarchical relationship to other plants and animals.

The application of taxonomy as a means to organize business content is a much more complex issue. Since we are talking about semantics and language here, there is an inherent problem with using the word taxonomy to describe this type of technology. When applied to business content, as opposed to scientific classification, taxonomy is a conceptual organizational structure. Unlike the categorizing of life forms, categorizing business documents can and should be ambiguous. A document could and perhaps should be placed in multiple categories depending on the business context and the task environment or expertise of the user. This added level

The application of taxonomy as a means to organize business content is a complex issue.

Since we are talking about semantics and language here, there is an inherent problem with using the word taxonomy to describe this type of technology. When applied to business content, as opposed to scientific classification, taxonomy is a conceptual organizational structure.

Unlike the categorizing of life forms, categorizing of business documents of classification complexity in the business setting makes the design and construction of taxonomy structures all the more challenging.

The need to categorize or position content sources (e.g. documents, application data, rich media, etc.) into the resulting taxonomy requires careful, deliberate classification of each source. The assignment of metadata tags, for example, that enhance each content object and suggest "latching" the content to specific subject categories and taxonomy structures. The tagging effort represents another process that a business must undertake in order to obtain the benefits of a taxonomy. In some cases, this could be done manually. But this approach is not easily scalable. The speed of the classification is limited by the skills and number of individuals assigned to this classification task. Additionally, who within the organization will take on this role is an issue. Will authors (e.g. researchers, scientists, lawyers, doctors, senior management) be willing or available to perform this classification manually? Will a separate staff of content classification specialists need to be defined and trained?

Addressing these issues of taxonomy definition, construction and classification must have a dual focus:

- identify business requirements and trends
- understand and evaluate available technology alternatives.

To shed light on current business practice in the area of taxonomy, we turn now to the results of the Delphi Research 2004 survey on user experience with search and taxonomy technologies and practices.

Taxonomy and Classification Market Analysis

This section of the report provides an in-depth look at the reality of the taxonomy and content classification market. These findings represent a current real-world view into user expectations, requirements and experiences with these technologies and their role in enterprise information architectures.

Methodology/Survey Population

The charts in this section of the report present the findings from a survey vehicle initiated and distributed by Delphi Research in early 2004. The survey gathered responses from over 300 individual users and evaluators within firms worldwide. Approximately 60% of responses were from the U.S.

A full discussion and description of survey population characteristics in included in Appendix A at page 40.

(*A full discussion and description of survey population characteristics is included in Appendix A on page 40 of this report.*)

Respondents represent viewpoints from many vertical industries. The appeal of taxonomy and need for an information architecture appears to cut across horizontal industries. The same can also be said with regards to the size of the company in terms of both annual revenue and number of employees. Perhaps more importantly. the views expressed here span across user types. The roles represented in this survey population range from senior management in both IT and business functions to line of business managers, project managers and end users. Respondents also varied with regards to their current or anticipated role in the taxonomy project.

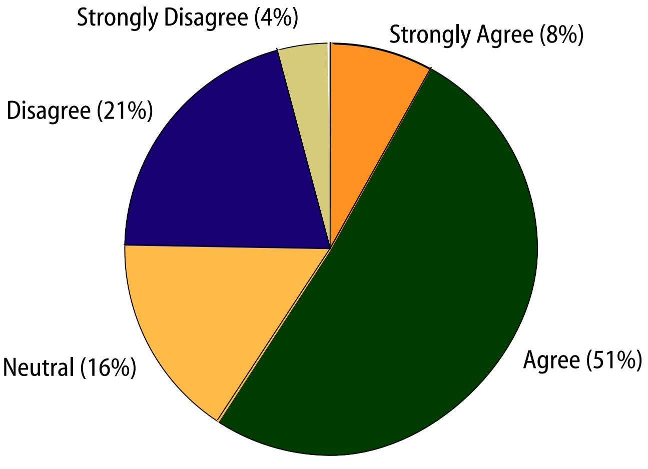
The 2004 survey follows a 2002 Delphi Research initiative on the same topic. In some instances, questions were intentionally repeated from the 2002 survey to demonstrate and validate evolving opinions, attitudes and behaviors within the taxonomy market. Where this was done, charts and data are presented with both years' findings juxtaposed to demonstrate contrast.

Although presented in this section of the report in their entirety, several specific survey findings and observations are presented elsewhere throughout the report to reinforce and illustrate concepts and trends, in addition to the quantified results represented within the charts.

Assessing the Current Situation

As noted in the introduction of this report, survey respondents reported that the ability to locate and retrieve information in order to perform their jobs more effectively has become simpler and more effective over the last two years. A full 59% of respondents agreed with this suggestion, with a significantly small minority of 25% disagreeing or strongly disagreeing.

"Finding the information I need to do my job has become simpler and more effective over the past 2 years."



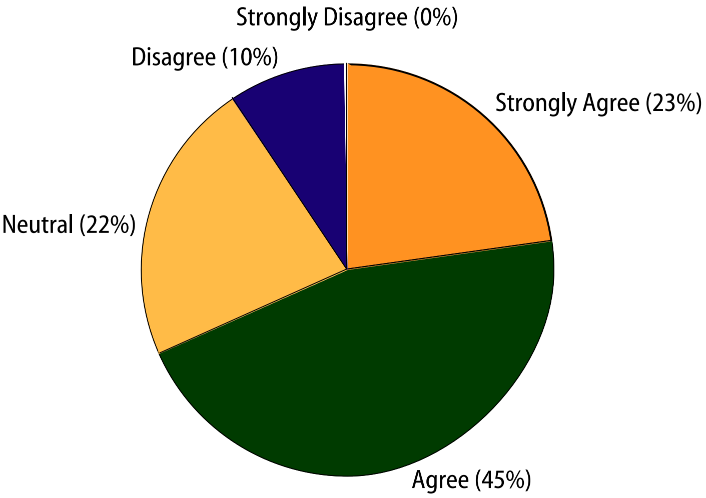
©2004 Delphi Group

Despite this strongly positive finding, a tempering note of reality is offered by an even stronger agreement amongst survey respondents that the process of locating and retrieving the information needed to effectively execute their jobs is still nonetheless difficult and time consuming. In this instance an even larger majority—68%—concur, and perhaps somewhat surprisingly not a single respondent strongly disagreed with this statement.

While users see some improvement in information retrieval over the last 2 years, their attitude towards its level of difficulty remained virtually the same. (Survey results on this question of the difficulties of search remained nearly the same in 2002 and 2004.)

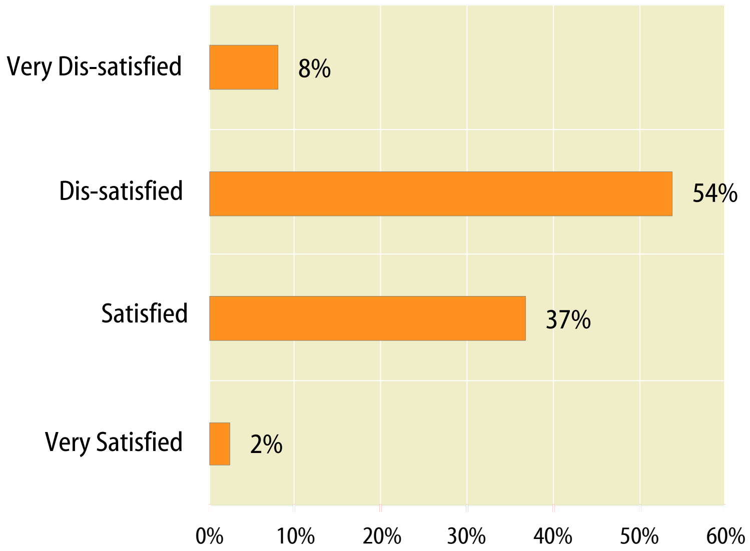
Further evidence of recent improvements being simply not enough to overcome user frustration with intelligent information organization and retrieval in their organizations is provided by the 62% of the respondents who reported that they

"Finding the information I need to do my job is difficult and time consuming."



©2004 Delphi Group

Satisfaction With Search Experience



©2004 Delphi Group

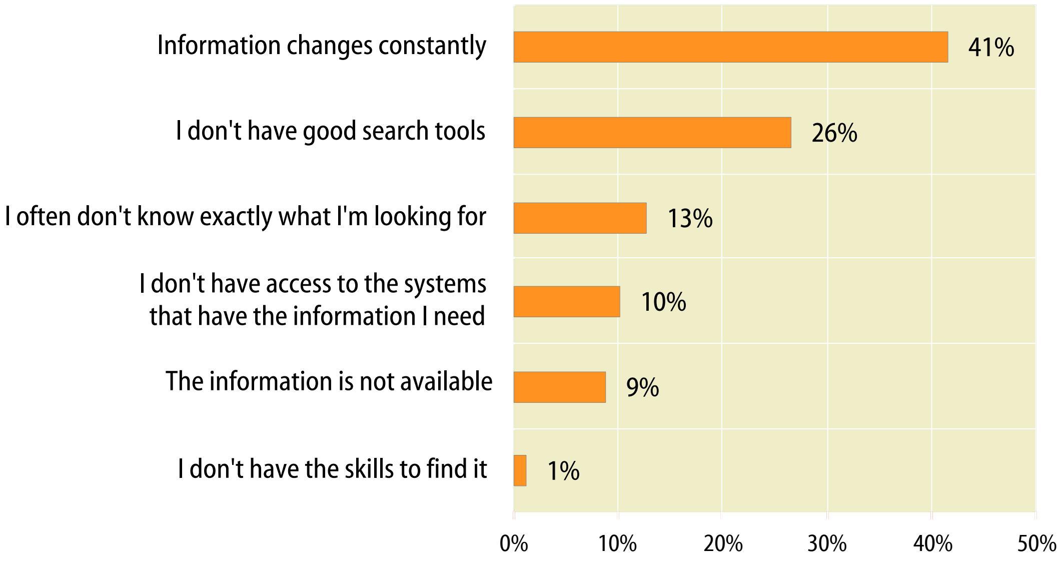
are dissatisfied or very dissatisfied with the overall search experience provided by their work environment.

When asked to identify the largest impediments to timely and effective retrieval of information, respondents overwhelmingly pointed to the fact that business content is constantly changing, and thus has to be continually relocated and retrieved, as the number one biggest impediment. Here is one area where the application of a centralized taxonomy or set of taxonomies within an information architecture would provide significant benefit. The consistent and constant classification of business content into a taxonomy structure would provide a persistent and reliable single point of access. Users could reliably return to known nodes of a taxonomy to pinpoint specific content and access its latest state.

The Quality of and the Reality of Search Habits

Imposing the element of time into the analysis sheds greater light on the magnitude of this issue. Business users spend a significant part of their work day pursu-

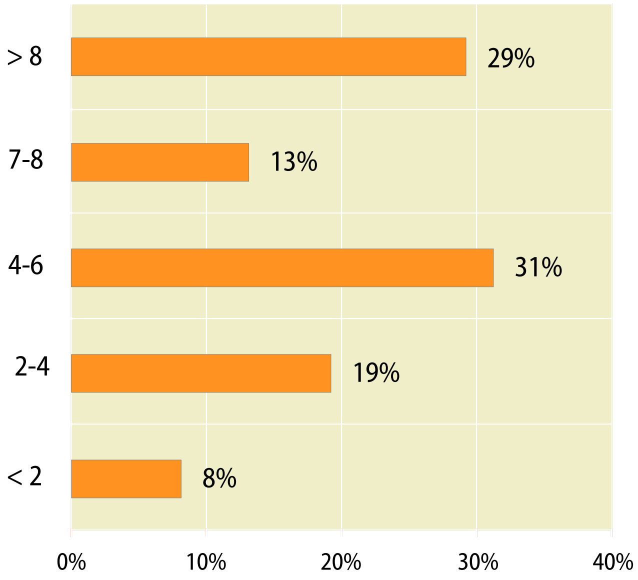
Impediments To Finding the Right Information



©2004 Delphi Group

ing relevant information. A full 42% of survey respondents report that 20% of their work week is spent seeking information. Another 31% percent spend between 10%–15% of their time each week looking for content. Clearly there is a business case to be made here regarding how knowledge worker productivity could be enhanced through faster/ more effective search and retrieval.

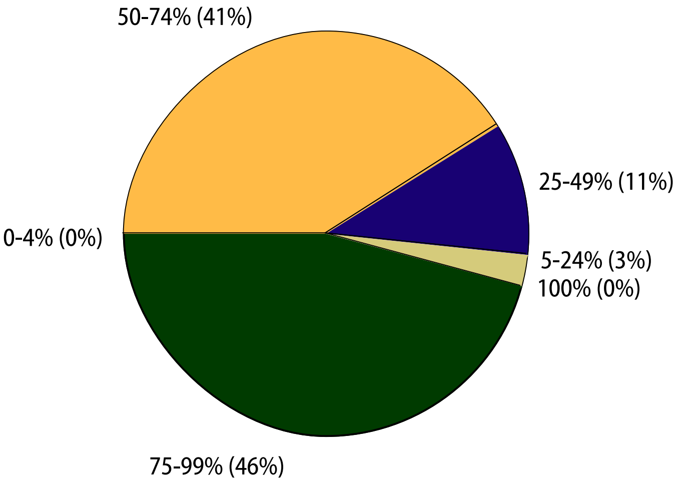
Hours per Week Spent Searching



©2004 Delphi Group

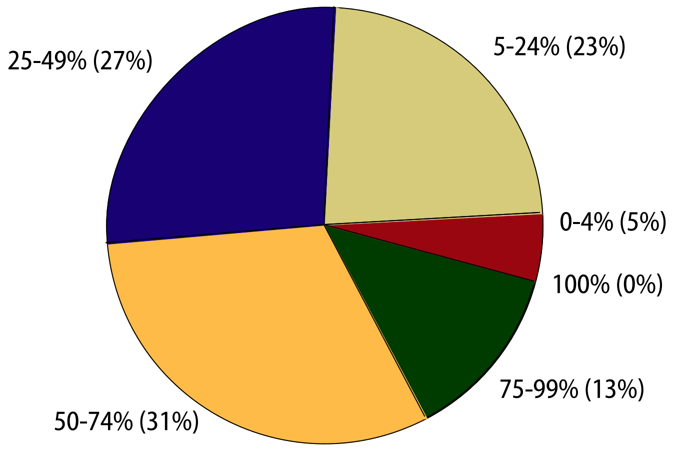
Ultimately the time spent searching and retrieving should be viewed via the return on time invested. In other words, are current efforts yielding valuable results to today's business users? In this regard, the survey responses uncovered another positive trend. While no one reported that they discover what they were looking for 100% of the time, an overwhelming majority of the respondents stated that they are successful more than half the time. Indeed, a full 46% reported success in 75% - 99% of the time. Thus, given enough time and willingness to endure "pain" and frustration, the "right" content is retrieved. This is perhaps one impetus to the respondents' more positive perspective on retrieval over the last two years. This finding points our investigation more towards the speed and ease of use of retrieval environments and less to the effectiveness, as the primary point of pain amongst today's business people. Here issues such as the availability of a taxonomy, its intuitive nature, and accurate relevancy ranking become more critical focal points for an information architecture strategy. Other

"What percent of the time do you find the information you are looking for?"



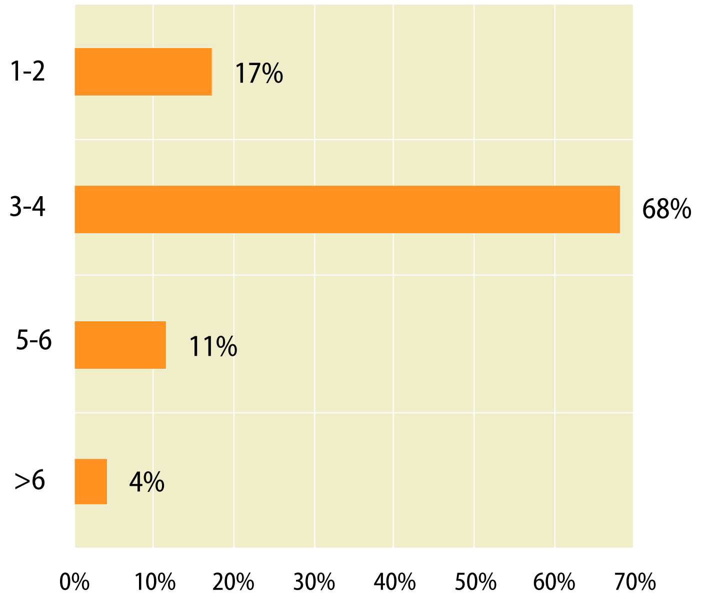
©2004 Delphi Group

"What percent of the time do you find the information you are looking for on the FIRST PAGE of suggested search results?"



©2004 Delphi Group

Levels of Taxonomy Navigated



©2004 Delphi Group

survey findings that follow support this conclusion.

While most respondents reported finding what they needed most of the time, it is often at the cost of seemingly unnecessary digging - as opposed to efficient finding. Survey respondents indicated that they rarely find what they need on the first page of retrieved content. While this is applicable to relevancy ranking in search, it also bears strong indication for the need for well-constructed taxonomies, or navigable paths to content. Designing and engineering the order or hierarchy into which taxonomies classify content is a core competency for the taxonomy practice.

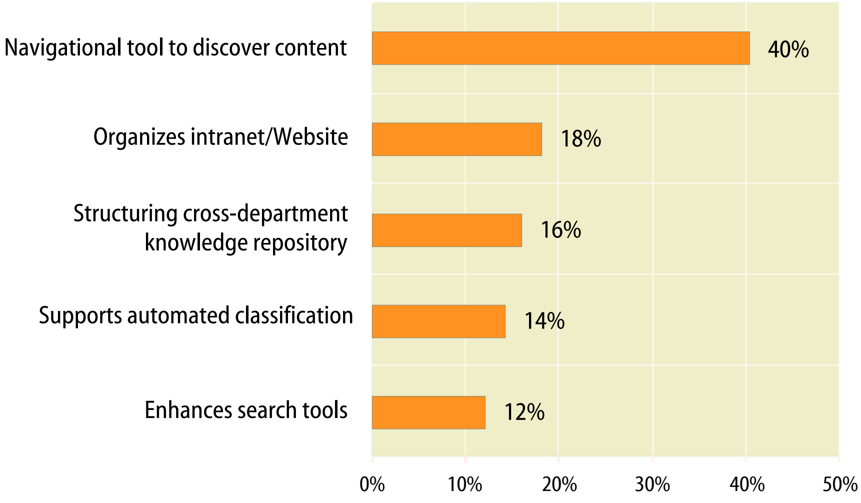
How far are users willing to keep searching through layers of content? Whether traversing a taxonomy to navigate a site, or in response to a query, or in organizing query results users exhibit a tolerance threshold. For years, we have purported a general rule of thumb in taxonomy design - the 10 x 4 rule. The premise is that good taxonomy design will never present more than 10 topics (choices) at any single level of the taxonomy, and that no path should navigate further than 4 levels. While generally accepted amongst many designers of taxonomies and web sites, these guidelines appear to resonate with surveyed respondents as well. Among respondents, the greatest number of them (68%) indicated a tolerance level for navigation of 3-4 levels. This speaks loud and clear, a well designed/useful business taxonomy should not provide more than 4 levels typically because user will not go beyond that. This can be a challenge for highly detailed and content-rich collections or web properties. Understanding the tolerance of the user or customer communities become key to a successful presentation or interface design.

Defining Taxonomy and Its Role in an Information Architecture

Our research focus now turns to user opinion regarding what a taxonomy should provide, and the nature of its role within an overall information architecture in the organization.

One of the stronger responses to come from the survey was in response to the inquiry asking for the role of the taxonomy in an organization. A solid 40% focused in on the primary value driver for taxonomy—that being a navigational tool for the discovery of content. 18% stated taxonomy would organize web sites, another 16% a means to organize content across multiple repositories—and thus might be viewed as a less strategic—yet similar role of the number one choice. 12% saw the taxonomy as primarily not an independent tool but as an infrastructure technology integrated into search to enhance the accuracy of results.

Primary Contribution of Taxonomy

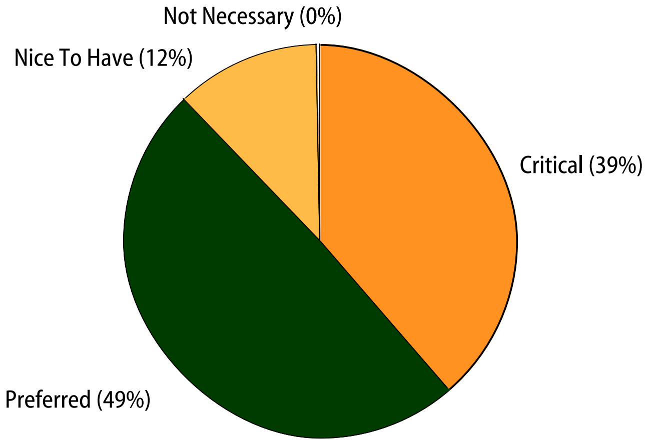


©2004 Delphi Group

14% of respondents identified the role of taxonomy as providing a means of automated classification, a feature of many software-based taxonomy applications. The criticality of this particular role is further investigated in the discussion on the following page.

The increased interest and position of taxonomy is verified by the views of the survey respondents with regards to the importance of taxonomy features in an information architecture and to the business. strategy. 39% of survey respondents indicate that the availability of a browsable taxonomy (as a form of navigation) was critical to their information architecture strategy. Another 49% indicated the browsable taxonomy was a preferred feature of the information architecture. No one indicted that it was not necessary.

Browsable Taxonomy: Importance to IA Strategy

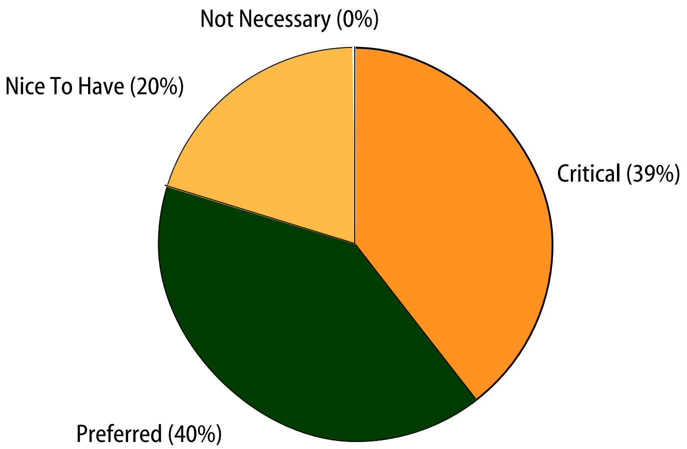


©2004 Delphi Group

When the survey focused on the process of constructing a taxonomy, respondents indicated a similarly strong opinion of the critical nature of taxonomy construction tools in an information architecture. 39% indicted that this function was critical to their information architecture. Another 40% see it as preferred, and again no one views it as not necessary. For those that might consider taxonomy creation a process that is best handled in a completely manual manner, that approach should be re-evaluated in light of this finding.

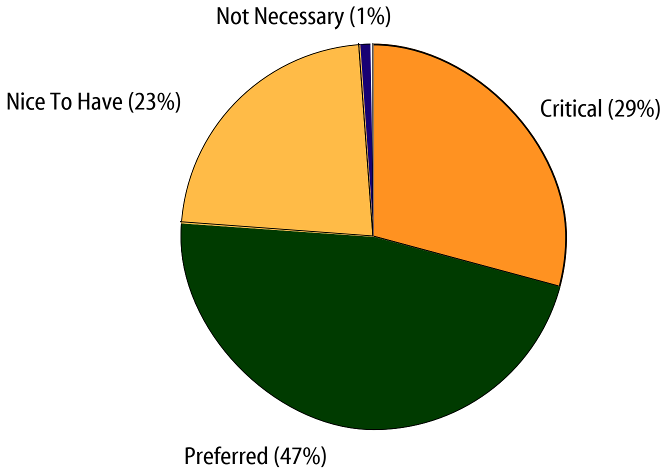
Similar observations are made with regards to an automated approach to classification of content into a

Taxonomy Construction Tool: Importance to IA Strategy



©2004 Delphi Group

Automated Classification: Importance to IA Strategy



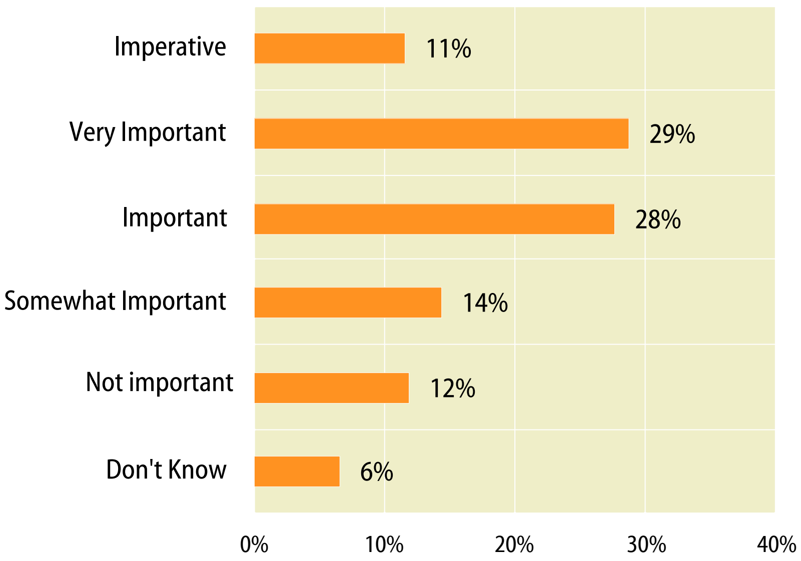
©2004 Delphi Group

taxonomy. While the percentage of respondents that ranked this feature as critical is not quite as strong as that for taxonomy construction tools (29%), when combined with those that rank it a preferred component to information architecture, a clear majority of 76% rank this functionality as strategic. Only 1% see automated classification as not necessary. Perhaps for those respondents that have not been through extensive classification exercises in tha past, the initial steps of creation can seem daunting and thus there is a strong perspective on construction assistance. The reality for those that have lived through construction is that this is only half the battle. Getting the business professional to tag/classify content is an equally daunting (perhaps even greater) task. Whatever assistance in this task that automation can provide should be seriously considered.

Perhaps the most significant finding with respect to the criticality of taxonomy within an organization is with respect to user opinion regarding its importance to a business strategy.

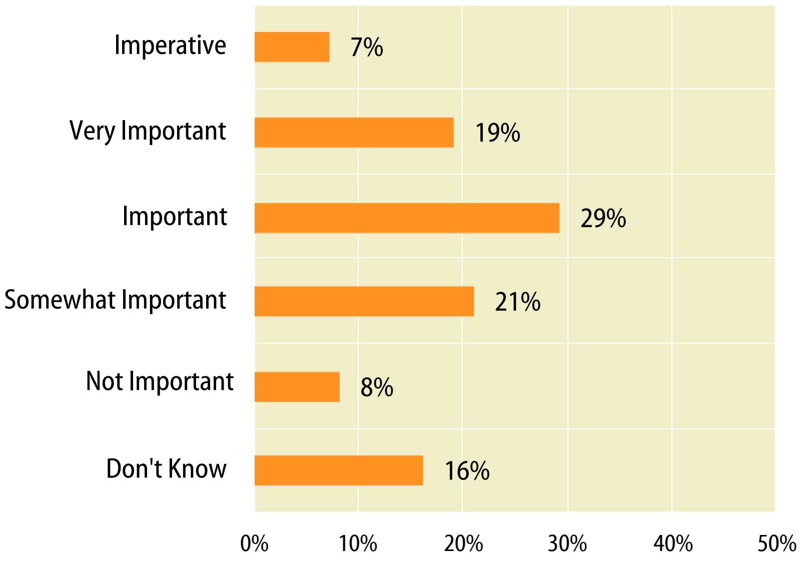
As taxonomy and classification technology mature and become more visible, as portals move into their second and third generation deployments in organizations, as web sites become more even popular vehicles for e-business, the role of taxonomy to the basic business strategy grows in importance. In this respect, we see a marked change from opinion two years ago. In 2002, 7% of survey respondents saw taxonomy

Importance of Taxonomy to Business Strategy



©2004 Delphi Group

Importance of Taxonomy to Business Strategy: 2002 Survey



©2004 Delphi Group

as imperative to their business strategy. 2004 respondents increased this opinion by 4 percentage points (11% ranked taxonomy as imperative to the business strategy), nearly a 60% increase from the earlier survey.

The largest gain in user opinion however, came from those that rank the importance of taxonomy to the business strategy as very important. Whereas in 2002, 19% of respondents ranked taxonomy as "very important," 29% of respondents in 2004 ranked taxonomy as very important to business strategy (a gain of 10 percentage points or a 53% shift in sentiment in this category).

Far fewer 2004 respondents viewed the taxonomy as only somewhat important to business strategy compared to 2002. But equally significant, those who reported that they simply did not know the importance (understand) the impact of taxonomy dropped from 16% in 2002 to a mere 6% in 2004. Clearly as business professionals get more educated and accustomed to the challenges of e-business, portals, web sites, and the growing complexity of related content, the role of a taxonomy becomes not only more clear, but more critical.

Current Practices

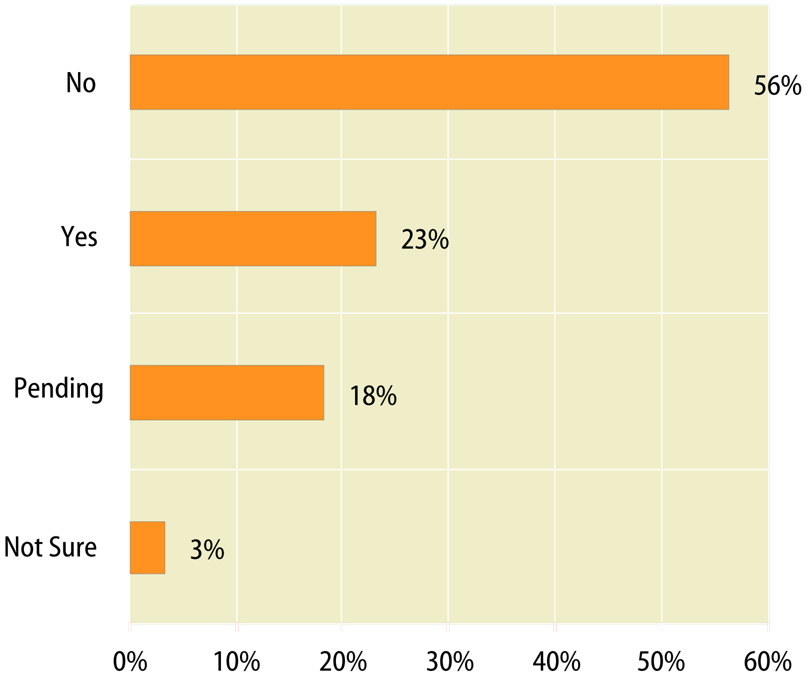
In spite of the raised awareness ofthe need for a taxonomy and classification of content, most organizations still do not have a formal approach to content classification. A little more than half—56%—of respondents indicated that they do not have access to a classification system, let alone use one. This is perhaps due to the fact that in many organizations the need is seen by the business user but the solution is believed to be (solely) a technical one.

When asked who develops the organization's taxonomy strategy, 57% of respondents indicated some form of IT (30% CIO and 27% IT not the CIO). While the taxonomy is a critical part of an overall information architecture, and could quite possibly utilize technology approaches, the strategy for the taxonomy should rest in the hands of those that own and use the content that is classified. Yet, Line of Business Managers and Librarians were indicated by only 20% (14% and 6% respectively) of the respondents as being the drivers of the taxonomy strategy.

Similar findings were reported with regards to the overall information architecture for the organization. At this higher level of strategy, greater input from IT, especially at the CIO level is to be expected. Best practices would dictate a hybrid approach with strong input from Line of Business Managers and Cxx management as well.

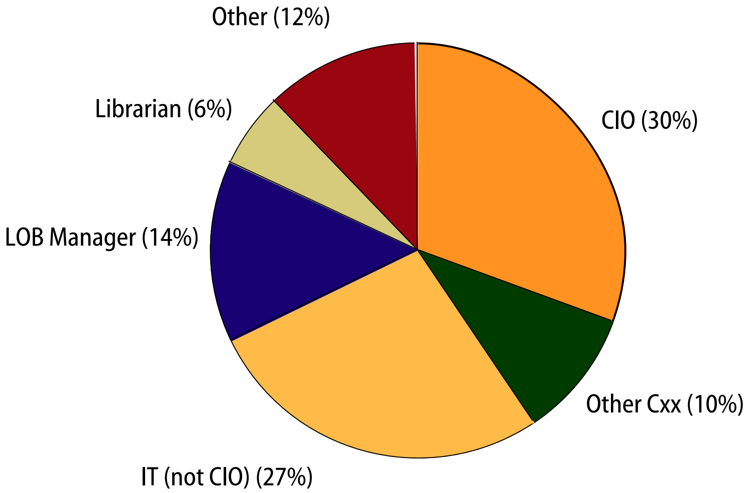
On the other hand, when the taxonomy is implemented in software, our respondents indicted that the IT group is pri-

User Access to Classification Facility



©2004 Delphi Group

Who Drives Taxonomy Strategy



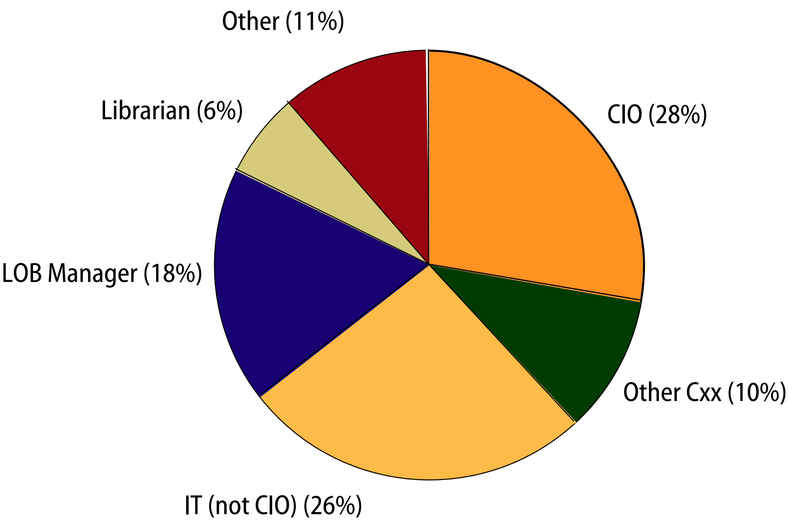
©2004 Delphi Group

marily tasked with the maintenance of the software. This is in direct contrast to maintenance of the taxonomy itself, which should be the domain of subject matter experts, librarians and content owners. Indeed, despite a strong focus on IT as the owners and strategists of the taxonomy, most organizations nonetheless (and in this case rightfully so), place the usage of the taxonomy and classification systems on the shoulders of the users.

Among those organizations that had some approach to taxonomy and classification, the most popular approach was a hybrid system (software and manual approaches), which is to be expected, as no software can completely automate this process.

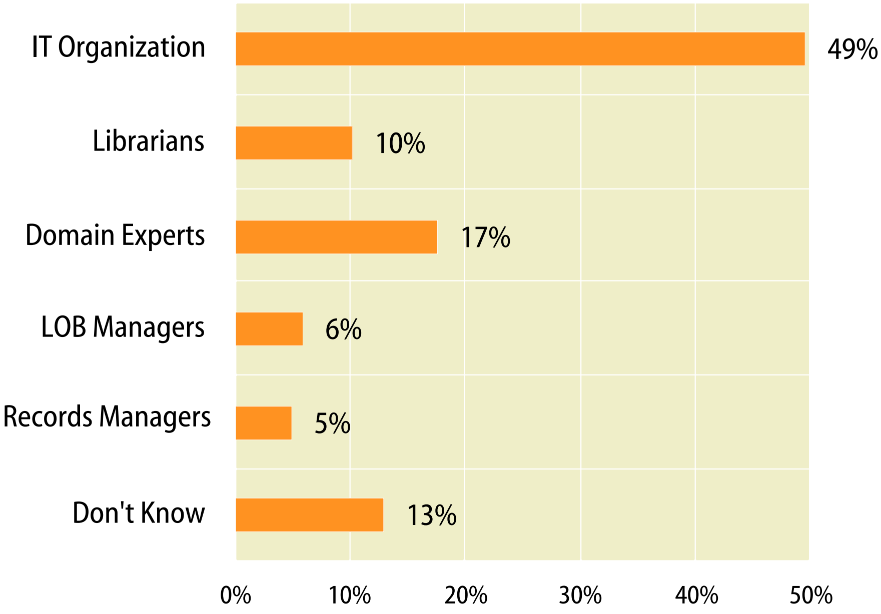
Indeed, despite a perhaps misdirected predominate focus on IT as the owners of the taxonomy strategy and maintenance, the individual user (i.e. author) is overwhelmingly targeted as the responsible party for actually executing the classification process at a local level.

Who Drives IA Strategy



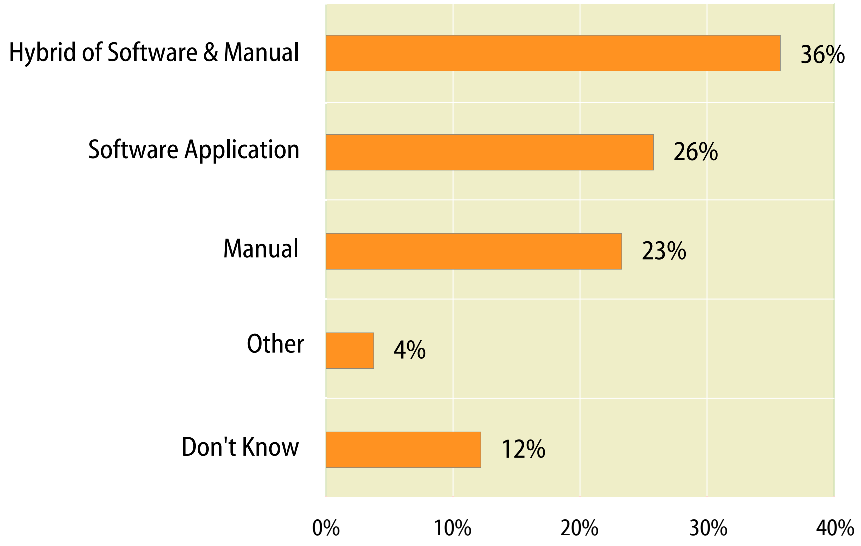
©2004 Delphi Group

Responsibility for Taxonomy Software Maintenance



©2004 Delphi Group

Taxonomy Implementation Strategy

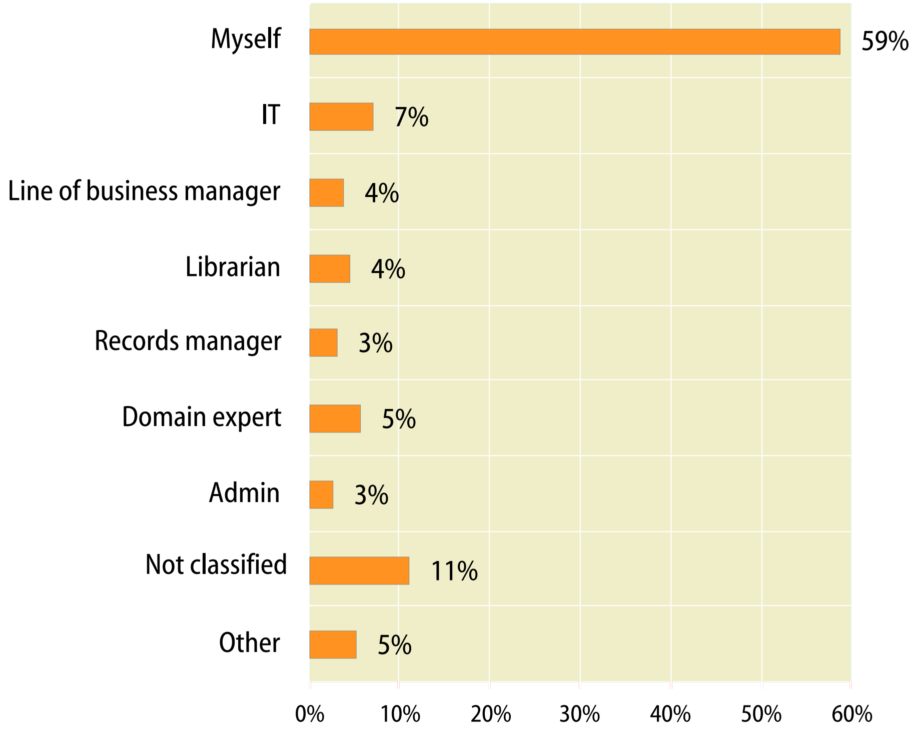


©2004 Delphi Group

Preferred Approach to Implementation and Cost Justification

Survey respondents were asked to identify the preferred approach to deploying taxonomy software. Here, respondents reinforced their sentiment that the taxonomy is a critical component of an information architecture and business strategy. Only 15% see the taxonomy as a stand-alone function. The largest single response category—29%—are those that prefer the taxonomy system as a toolkit that can be readily integrated with other business applications. Another 17% indicated a preference for the taxonomy software to

Who Classifies My Information?



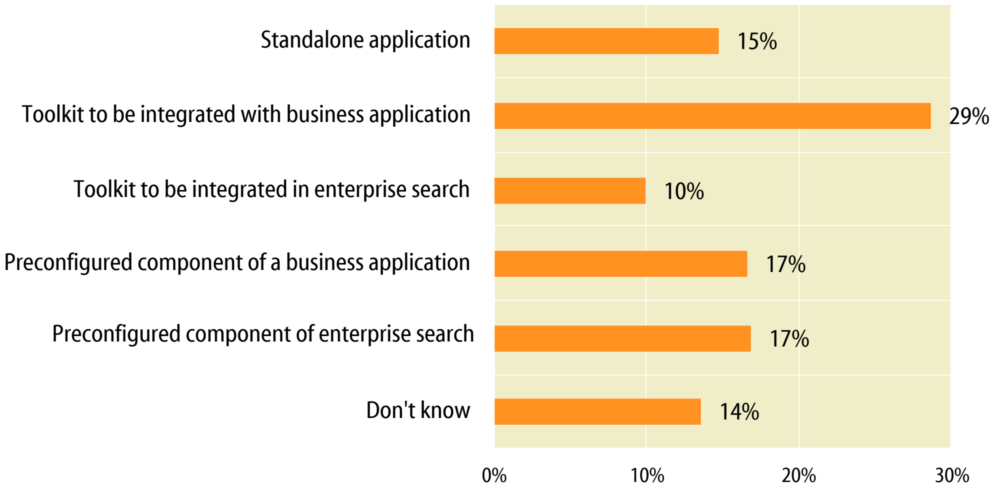
©2004 Delphi Group

be delivered fully integrated with another business application, and the same percentage indicated a preference for taxonomy integrated with a search tool.

We conclude this market analysis on a sobering note. Despite the increased appreciation for taxonomy and classification, higher awareness for its need and more strategic placement within business and information architecture strategy, justification of taxonomy and classification still lies predominately in soft dollar areas.

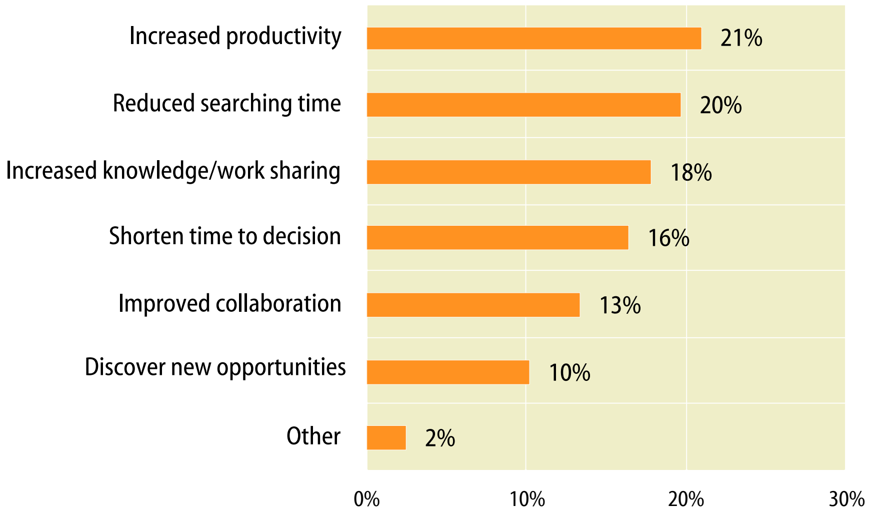
The benefits achieved and/or anticipated from the deployment of taxonomy and classification are predominately focused on reduced search time and increased sharing of knowledge and work product. While the number one response is "increased productivity" (by a small margin), this is often attributed to the reduction in search time, which in reality does not always directly translate into increased productivity.

Taxonomy Package & Deployment Preference



©2004 Delphi Group

Expected Benefits of IR & Taxonomy Software



©2004 Delphi Group

Taxonomy and Classification Technology

In this section of the report we take a detailed look at the state of the technology which has been developed to facilitate the creation and maintenance of taxonomies and the classification of content.

When investigating the applicability of technology to the taxonomy and classification process, it is helpful to break the process down into four distinct stages. The four distinct stages of taxonomy business practice each represent a level of functionality that must be incorporated in a solution definition, and thus can potentially be addressed via a technology implementation. These four stages are:

1. Developing of the taxonomy structure
2. Categorizing the content and placing the pointers to the documents in the hierarchical structure
3. Presenting the information (or building the interface that helps users find the information)
4. Incorporating and analyzing new content and maintaining the taxonomy structure

Taxonomy software has been developed over the past several years to increase the speed and efficiency of each of these stages. It is important to note the distinction between taxonomy design/construction and the classification of content sourcesand objects into that taxonomy. These are two separate and distinct operations and one should not assume that these processes are happening simultaneously or concurrently. These are serial steps. While these functions are closely related, technology solutions do not always provide both levels of functionality. More often, technology solutions focus more on the classification process. The taxonomy software's fundamental challenge is to understand the concepts and ideas that group like documents together and separate unlike documents. Taxonomy software provides automated and semi-automated approaches to defining these subject hierarchies and classifying submitted content into this organizational structure.

Methodology Algorithms

There are many approaches to tackling the problem of building automatic or semi-automatic taxonomies and automating the classification process. Technology vendors' solutions provide a variety of approaches which combine these functions in distinct configurations.

These underlying technologies are based on a number of different development approches and algorithms, the most common being:

- Rules-based
- Bayesian
- Linguistic and Semantic
- Support Vector Machine
- Pattern Matching and Other Statistical Algorithms
- Neural Networks

Rule-Based

The rules-based approach is perhaps the most straightforward and user-controllable approach. The rules-based approach requires experts to create and maintain a set of rules for a document to be included in any given category of a taxonomy. Thus, the rules-based approach focuses more on the classification process than the construction and definition of a taxonomy.

Experts define "If-Then" rules that can support complex operations and decision trees. Rule-based systems can precisely define the criteria by which a document is classified. The rule measures how well a given document meets the criteria for membership in a particular topic.

For example, a rule could be that all documents that include the terms "San Francisco," "Chicago," "New Orleans" be listed in a category called "Cities, USA." Such rules often break down when ambiguous values like "Cambridge" arise. Is this Cambridge in Massachusetts or in England? Rules must be carefully articulated and made as unambiguous as possible.

Besides the content of documents, rules can be applied to metadata and even business policies. For instance, a rule might specify that only PDF documents created since January 2000 should be included in a particular category. Thus rules are a powerful and flexible means for automatically classifying content based on not only content itself but also the metadata that describes the content's business context (for example: author, date, or keyword data can be used in rules).The downside of rule-based systems is that expensive human domain experts have to write and maintain the rules.

Statistical Analysis

This approach supports both the creation of a taxonomy (or a strawman/ first draft) and the subsequent classification of content into that taxonomy. The approach measures word frequency, placement and grouping, as well as the distance between words in a document.

Typically, the statistical approach to taxonomy definition and constructions (as opposed to subsequent classification) requires some form of preliminary training. This could take the form of a basic taxonomy, defined by a human expert. But the breadth and validity of the structure, and subsequent classification rules can be automated through application of a training set of documents into the design process. In this approach, subsets of documents are identified manually and presented to the software as "exemplary" to a given topic or node of the taxonomy. The provided sample content is analyzed and from this the taxonomy is further refined and the rules of classification established. These rules are then used to automate the analysis of new documents and their classification into the taxonomy. This approach is also referred to as "machine learning."

Limitations of the example-based taxonomy method include problems that arise from the fact that the resulting classification is totally dependent on the breadth and precision of the training set, and the training set still must be identified manually. In any case, the statistical analysis that is performed can deploy one or more approaches: Baysean Probability, Neural Networks and Support Vector Machines.

Bayesian Probability

The Bayesian approach attempts a concept-based analysis by learning the probabilities of words being related in a given category. The Bayesian algorithm sorts documents by examining the electronic patterns contained in the text or content contained therein. Bayesian probability uses statistical models from words in training sets, and uses pattern analysis to assign the probability of correlation. This is one of the more common methods applied to building categories and taxonomy structures.

An example of Bayesian probability would be that if a given document contains the words "apples" and "oranges" it is more than likely this document is about fruit, which leads to the assumption that other fruit nouns such as "grapes" or "tangerines" will occur.

Neural Networks

Neural Networks create a matrix of computational nodes. These nodes track and compare topic similarity. A neural network utilizes artificial intelligence to build an interconnected system of processing elements, each with a limited number of inputs and outputs. Rather than being programmed, these systems learn to recognize patterns. Neural networks are an information processing technique based on the way biological nervous systems, such as the brain, process information. Composed of a large number of highly interconnected processing elements, a neural network system uses the technique of learning by example to resolve problems. The neural network is configured for a specific application, such as data classification or pattern recognition, through a learning process called "training."

Support Vector Machine

Support Vector Machine (SVM) is a refinement of taxonomy-by-example (i.e. it requires a training set). These algorithms are derived from statistical learning theory. SVM's calculate the maximum "separation," in multiple dimensions of one document from another. Each document—essentially a collection of words and phrases that together have meaning—are represented as a vector. The direction of the vector is determined by the words (dimension) it spans. The magnitude of the vector is determined by how many times each word occurs in the document (distance traveled in each dimension).As this iterative method continuously analyses documents, it separates them into either the "relevant" space or the "irrelevant" space. By repeating the process it categorizes those documents that are "relevant" into like categories, but more importantly learns how they are different from other categories.

Semantic and Linguistic Clustering

Semantic analysis and clustering supports both the creation of a taxonomy and the categorizing of content. This approach is typically language dependent. Documents are clustered or grouped depending on meaning of words using thesauri, custom dictionaries (e.g. a dictionary of abbreviations), parts-of-speech analyzers, rule-based and probabilistic grammar, recognition of idioms, verb chain recognition, and noun phrase identifiers (e.g. "business unit manager").

Linguistic software also analyzes the structure of sentences, identifying the subject, verb, and objects. The sentence structure analysis is applied to extract the meaning. Stemming (reducing a word to its root) also helps linguistic or semantic clustering. Clustering is a technique for partitioning documents/words into subsets of similar documents/words based on the identification of common elements between the documents/words.

Combining Methodologies

Of course, no single taxonomy and classification methodology is superior to another for every possible application. The trend by taxonomy software companies is to combine multiple methods to categorize the corpus of documents to increase accuracy and the relevancy of groupings. Each approach and combination of approaches has pros and cons associated with them, and their use depends on a design perspective and performance characteristics desired. The bottom line is to understand how these differences affect system performance in the only environment that matters—your unique data environment.

Determining Your Taxonomy and Classification Requirements

It is therefore critical to precede any investigation of technology solutions for taxonomy and classification with a determination of what your needs are, the types of content that will be managed, and the best approach to exposing the taxonomy to the user.

It is important to understand the approaches available to taxonomy and classification presentation and integration. Technology tools are available as stand-alone applications and as components to integrated knowledge management and document management systems.

Presentation and Integration

When investing the technology alternatives pay attention to the user interface. Does the tool come with a predefined interface? Is this interface conducive to the way users think about their content? Nested file folders that mimic the popular MS Windows GUI are available, as are nested tree structures, alphabetical listings of topics and sub-topics, tab-based interfaces, heat maps, hyperbolic trees that resemble tinker-toy like connections of topics, and even voice recognition interfaces.

A number of vendors view this technology as a fundamental component of the information infrastructure. Just as relational databases are a fundamental infrastructure component of applications such as accounting, CRM, and other enterprise applications, taxonomy software can be the infrastructure component that correlates unstructured data. This design philosophy positions taxonomy software as a core module in an architecture that works on the unstructured data within the organization.

But whether using a vendor-provided interface or creating one, the technology should provide some approach to integration. Minimally, it is anticipated that the taxonomy functionality will be integrated with search tools—it is likely that your solution will extend even further.

Design Considerations

Manual vs. Automatic

There is often much debate about manual taxonomy, versus automatic, versus a hybrid of the two. This is not really a relevant issue, because in order to make the taxonomy relevant to the users, it must match their needs and unique rules for relevancy. It is important to re-emphasize that there are two distinct steps in constructing a useful taxonomy. The first is the design of the structure of the taxonomy. The second is populating the structure, that is classifying content into the structure. The universe of taxonomy software providers presents a wide range of functionality configurations.

It is critical to bear in mind that the end result of any taxonomy initiative is a *human* interface: concepts and ideas are inherently relative, personal and subject to change. Consequently, virtually all taxonomy vendors supply some type of tool to customize, rename, or dynamically create the nodes of the taxonomy structure to suit individual needs. The difference between these applications and tool sets is an important feature for evaluation.

There is often much debate about manual taxonomy, versus automatic, versus a hybrid of the two. This is not really a relevant issue, because in order to make the taxonomy relevant to the users, it must match their needs and unique rules for

One of the key advantages of an automatic system vs. a manual system is consistency. An important perspective here is to consider whether the people doing the classifying have the same criteria for assigning categories as do the users. Categorizing the same concepts into the same place is what automatic systems do well. If the automatic systems misunderstand a concept, they will at least mis-categorize all related documents and not scatter them in multiple categories.

The decision to adopt a manual, automatic or hybrid approach is a complex one. Whatever your choice, your organization should commit sufficient budget and human resources to maintaining an accurate, up-to-date and relevant taxonomy system. The trend in the industry is to combine machine methods and human processing to develop and maintain taxonomies, as was indicated by the recent research findings.

Maintenance and Dynamic Information.

As you become more involved in the process of designing and deploying a taxonomy, you will soon realize that this is not a one-time effort. Taxonomy and classification are ongoing processes that require a long-term investment—business priorities, technology, language, and human interest are in a constant state of flux. The more volatile the information, the more the need for a systematic process to keep the information categorized and relevant.

New content is continuously added to repositories, while new versions of existing content are released, and out-of-date content is/should be removed from circulation. Changing strategies, emerging new interests and foci, evolving products, and advancing technologies all drive the need for changes to a taxonomy design and approaches to classification. Taxonomies may be industry- or even department-specific. Information today is by nature dynamic—consequently, categorization systems must be dynamic as well.

Granularity of the Taxonomy Structure

Although there are two distinct sides to this debate, your decision will place you somewhere along a continuum of alternatives. For the purpose of this discussion, taxonomies can be arbitrarily divided into three sizes by the number of nodes or headings and subheadings:

- Small - 1,000 or less
- Medium - 1,001 to 20,000
- Large - +20,000.

There are many large taxonomies. The proponents of large taxonomies say that more is better. Since the organization of information is hierarchical, the users can drill down to as much detail as they wish. Levels of hierarchy greater than 10 are not uncommon in implementations on this scale.

At the other end of spectrum, proponents of small taxonomies argue that more than four levels once again confront users with a kind of infoglut, burdening them with receiving too many hits on a search, too much irrelevant information, etc.

The third interpretation here is that individuals or work groups can develop their own relevant taxonomies as a subset of large taxonomies.

In applications where the taxonomy will be used as a navigable front-end specifically to assist and accelerate research and discovery, the key design principle should be grounded on the user research, which shows that users do not want to navigate more than 3-4 layers down in a taxonomy.

Larger vs. Smaller

The proponents of large taxonomies say that more is better. Since the organization of information is hierarchical, the users can drill down to as much detail as they wish. Levels of hierarchy greater than 10 are not uncommon in implementations on this scale.

At the other end of spectrum, proponents of small taxonomies argue that more than four levels once again confront users with a new kind of Infoglut, receiving too many hits on a search, too much irrelevant information.

Exposing the Taxonomy

While this report focuses predominately on the construction and maintenance of a taxonomy, this represents only the back end of the delivered application. In order to realize the anticipated benefits of the taxonomy, your solution design must include an approach to exposing the taxonomy through some type of interface.

The taxonomy could be provided as a series of tabs on a web site or portal (an approach popular on many web sites today). These tabs could be screen/context sensitive, so that as the user selects one tab, it invokes a new screen with a new series of tabs (i.e. nested categories). The taxonomy could be exposed as a series of nested folders, each representing a topic. This approach can be popular with users accustomed to the Windows folder interface. The taxonomy could be a navigable tree structure in which clicking on any node of the tree exposes the sub-topics and content contained in that node. On the cutting edge there are voice-activated front-ends, heat maps, navigable hyperbolic structures, and other visualization models.

In addition to the interface itself, taxonomy design should also consider the physical structure that will be used to manipulate the taxonomy. The most basic approach is that of an Uber Tree—a nested hierarchy in its purest sense. Alternatively, looped tree structures allow for a networked approach, in which any one node or topic could be referenced by multiple other topics. A faceted deployment functions similarly to a series of linked but independent trees/hierarchies. In this approach, each tree is based on a different top level category (or facet) to the subject area. Each tree is cross linked to the others providing multiple paths to the same content. For example, a user could locate car products by color (initially) and then price range, followed by year, or any combination thereof. The matrix model links two separate taxonomies in a grid, providing cross selling capabilities. As an example, a subject matter taxonomy could be matrixed with a services and products taxonomy. A customer exploring a given service for a particular topic could traverse the matrix and discover how this service applied to all available topics.

A **faceted** deployment functions similarly to a series of linked but independent trees/hierarchies. In this approach, each tree is based on a different top level category (or facet) to the subject area. Each tree is cross linked to the others providing multiple paths to the same content. For example, a user could locate car products by color (initially) and then price range, followed by year, or any combination thereof.

The concept of a "stackonomy" is being explored by forward-thinking designers. In a stackonomy, multiple taxonomies are actually layered one on top of the other, and, where applicable, links between them are created and supported. In this approach the user can migrate from one facet or domain to another in a seamless manner.

Precedent - Vertical Taxonomies, Existing Thesauri, Records Mangers, and Librarians

When constructing a taxonomy, the availability of legacy knowledge and structures that exist in the enterprise should be explored. Their availability may provide a jump-start and save hours of rationalization and design consideration.

Some taxonomy providers develop and provide pre-built taxonomies geared toward a particular vertical market. These taxonomies are often tied to that vendor's software and commercial model. Typically, vendors supplying pre-built taxonomies do allow customizing of the organization structure and the naming of the nodes for each of the categories.

Alternatively, thesauri developed by third-parties can be purchased. Based on an industry file standard (see the Standards section on the following page for more detail), these structured lexicons can be plug-compatible with many search engines and also be used as powerful points of initiation to building a related taxonomy.

Many organizations have a records management function. The classification of records based on a records schedule can be very similar to a subject taxonomy, in a business setting. If such a records schedule exists, it behooves the taxonomy developer to investigate the schedule as a potential starting point to taxonomy design.

No organization should overlook the intellectual capital and domain expertise represented by a corporate librarian. Library skills can be invaluable to the taxonomy design and maintenance process. Librarians are skilled and schooled on establishing systems to aid in information tracking and retrieval. The on-line/automated taxonomy should not be viewed as a replacement to the librarian, but an extension of the librarian. No design effort should ignore the approaches to classification used by the librarian to date. As the issues of taxonomy maintenance and ownership come into focus, library staff should be considered as part of the ongoing taxonomy practice team.

The needs of individual users represent a major aspect to examine. Will one comprehensive enterprise taxonomy address everyone's needs, or will you need departmental taxonomies as well? Or will individual workers require their own unique taxonomies? Or will your environment require a blend of all of the above?

Users Needs and Personalized Taxonomies

The needs of individual users represent another major aspect to examine. Will one comprehensive enterprise taxonomy address everyone's needs, or will you need departmental taxonomies as well? Or will individual workers require their own unique taxonomies? Or will your environment require a blend of all of the above? As you investigate different taxonomy products, be sure to investigate how flexible the products are for generating multiple taxonomies and linking them as appropriate.

Standards

The structure and deployment of a business taxonomy does not itself have any specific standards that the designer can use to guide the design process or simplify the integration of taxonomies into other applications. However, there are a number of standards that reside on the periphery of taxonomy and classification construction that should be considered as possible facilitators to the design of the taxonomy and classification approaches.

ANSI Z39.19 - Thesaurus Construction

Thesaurus construction has been standardized by ANSI for nearly two decades. The availability of this standard has made commercially available plug-compatible thesauri available. It provides a series of standard operators that allow for the definition of a lexicon in a hierarchical fashion. There is support for broader meanings and narrower meanings, as well as synonyms. The "preferred term" operator, used to denote the preferred term amongst synonyms is typically a candidate for a taxonomy node label. The standard allows the user to decide between a pure hierarchy or a networked approach to word relationships. Availability of the "Scope Note" operator allows for intelligence about the language and specific words and phrases to be captured as well.

Dublin Core

This standard provides an approach to meta tagging unstructured content. Dublin Core came from the on-line library sharing area. It describes a standard set of meta tags that are used to track objects in a library collection, making the collections objects searchable in a standard manner. For organizations struggling with defining an internal standard for meta tagging, Dublin Core can provide a starting point. In practice, the standard tags are typically modified or customized for internal use.

RDF (Resource Description Framework)

RDF is a product of the semantic web and the W^{3}C project. RDF specifies a syntax and schema for describing web content in XML. While focused on providing a standard means to exchange content and processes across web sites, the standard could be utilized as a way to construct a taxonomy. The operators provided enable the construction of an intelligent taxonomy, meaning the relationships between topics can be self described.

DAML (DARPA Agent Markup Language)

This standard is also a product of the semantic web. It provides an extension to RDF. With wider support for self-describing tags, ontologies can be defined in DAML. DAML allows the designer to specify not only the relationship of the link between two topics (e.g.. parent-child, "was born in," "is the author of" links), but also to specify rules about the nature of the values in the link. For example, "disjointed" classes specify that a member of one node cannot also be the member of another node.

ISO/IEC 13250 Topic Maps

Topic maps are a standard XML schema, XTM, which utilizes meta tags to build a self-describing ontology, thesaurus or taxonomy. Topic Maps can be integrated to search features to provide intelligent search and/or to enable web site-to-web site automated communication and exchange. The components of a topic map are:

Topics - A declared “subject” (or more generally any “thing”) associated with the topic is the name. There are also distinct types of names – base names, display names, and sort names.

Occurrences - declarative links to web sites, other topics, or other forms of on-line content that represent (e.g. are examples of) the topic.

Associations - An association is a link element that asserts a relationship between two or more topics, e.g. ‘Grapes of Wrath’ (a topic type book) is linked to topic “John Steinbeck” via the “was written by” association.

Mind Share Ranking Methodology

For a view of the market's perception of the relative strength of contribution to taxonomy technology and practice achieved by competitive taxonomy software suppliers, the 2004 Delphi Research survey included a vehicle to gauge the relative "mindshare" of the current enterprise competitors in terms of brand awateness.

In the survey of over 300 professionals, we asked respondents the following open-ended question:

"Which technology companies do you perceive to be leading in the development of taxonomy?"

Respondents were provided three blank spaces in the survey form and asked to type in the correct name of their choices for the market leader, the second ranking and the third ranking firm respectively—an unprompted "open ballot" process.

The results tabulated in the accompanying chart are based on respondents input in the market leader field on the ballot form.

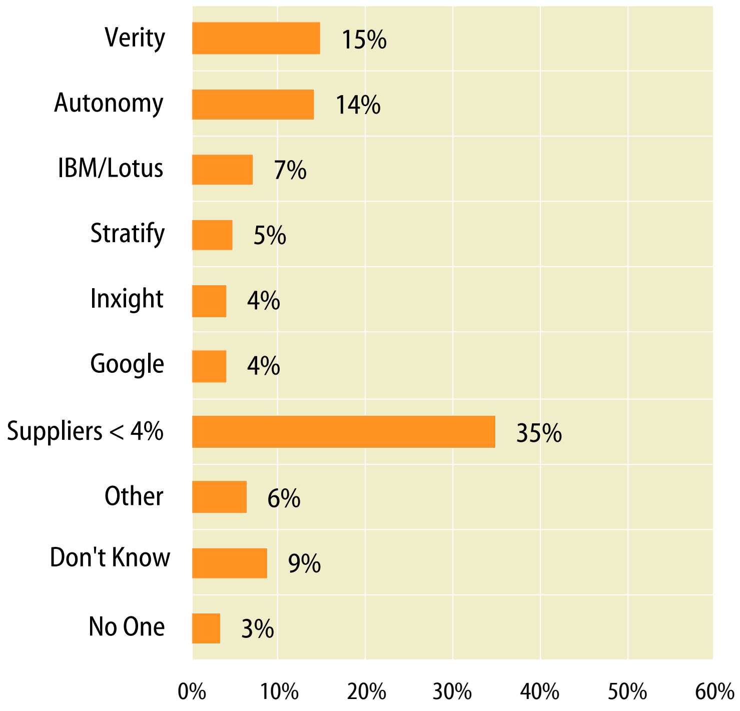
Approximately 50% of the respondents elected to write in their choices. Clearly inaccurate or frivolous responses were thrown out of the analysis. Cross check analytics were performed to eliminate "stuffing" of ballots.

Taxonomy Software Suppliers—Market Mindshare

In the 2004 market survey, Delphi Research included a vehicle to gauge consumer sentiment about the relative strength of brand recognition and/or perceived strength of industry contrubution among the competing taxonomy software suppliers. In the survey of over 300 professionals, we asked respondents the following open-ended question:

"Which technology companies do you perceive to be leading in the development of taxonomy?"

Taxonomy Software Suppliers – Market Mind Share Ranking



©2004 Delphi Group

Respondents were provided three blank spaces in the survey form and asked to type in the correct name of their choices for the market leader, the second ranking and the third ranking firm respectively—an unprompted "open ballot" process.

The results of this poll, based on reponses written in to the market leader field, are presented in the chart above.

From these responses, one could draw the conclusion that there is no clear leader among suppliers with regard to market mind share. On the other hand, Autonomy and Verity clearly each have developed a strong reputation for leadership in the taxonomy and classification area, with respondents naming the two companies virtually equally in the poll. With this situation of "split" market leadership, however, neither company's overall percentage of votes was very impressive—with results ranging in the low teens.

Clearly one factor contributing to the relatively low overall position of the mind share leaders is the large number of suppliers named only a few times each—the group of vendors polling less than 4%. This group, gathering 35% of the "Market Leader" write-ins, is by far the largest, indicating that this market is still highly fragmented on the supply side. While typical of the patterns we see in emergent software markets, we expect to see the process of consolidation accelerate in 2004 and 2005 along with a gradual increase in IT investment levels and an the arrival of more expansive integrated solution configurations from larger players in the space.

Vendors tallying fewer than 4% of the write-in votes for taxonomy market leadership included both specialist firms in the broad area of search technology and larger firms whose technology suites include some categorization or search capabilities. The table below lists the vendors who received mentions in the survey.

In the group of four suppliers identified relatively frequently behind Auonomy and Verity, IBM/Lotus received marginally higher results to gain the third place ranking overall, while Stratify, Inxight, and Google gathered very similar numbers of market leadership mentions. Stratify and Inxight have both focused technology efforts as well as market positioning around taxonomy, and those efforts have evidently been relatively well-received by the market. We conclude that Google's appearance among these leaders is based heavily on brand awareness for its internet-based search property, since taxonomy and classification is not a strong feature of the company's enterprise-focused Google Search Appliance offering.

Vendors Identified at Less Than 4%

Clear Forest
Convera
Documentum
Endeca
Entrieva (Semio)
Factiva
FAST
Hummingbird
InMagic
Interwoven
Lexis Nexis
Microsoft
Mondeca
Nstein
Open Text
Oracle
Plumtree
Recommind
Schemalogic
SER
Software AG
Teragram
Vignette

Appendix A – Research Population Statistics

The charts in this section of the report present the findings from the 2004 survey with responses from over 300 individual users and evaluators within firms worldwide. Approximately 60% of responses are from the U.S.

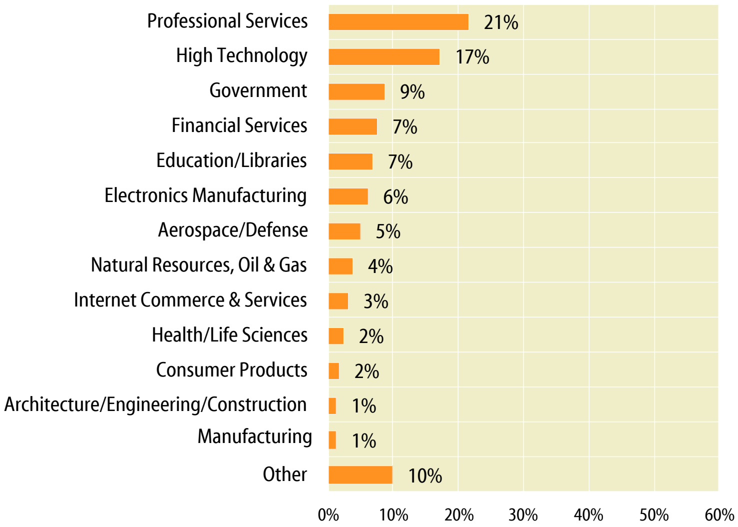
Respondents represent viewpoints from many vertical industries. The appeal of taxonomy and need for an information architecture appears to cut across horizontal industries. The same can also be said with regards to the size of the company in terms of both annual revenue and number of employees.

Perhaps more importantly. the views expressed here span across user types. The roles represented in this survey population range from senior management in both IT and business functions to line of business managers, project managers and end users.

Respondents also varied with regards to their current or anticipated role in the taxonomy project.

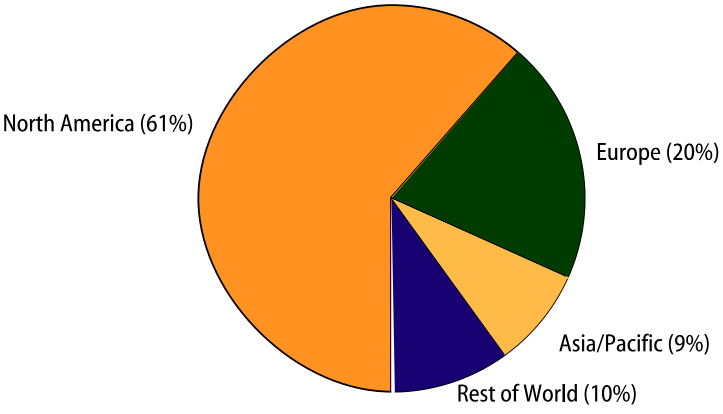
The 2004 survey follows a similar 2002 Delphi Research initiative. In some instances, questions were intentionally repeated from the 2002 survey to demonstrate and validate the evolving opinions, attitudes and behaviors within the taxonomy market. Where this was done, charts and data are presented with both years' findings juxtaposed to demonstrate contrast.

Respondents by Industry



©2004 Delphi Group

Respondents by Geography

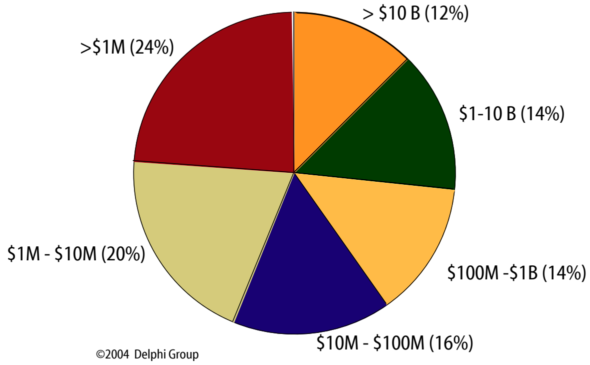


©2004 Delphi Group

Appendix A – Research Population

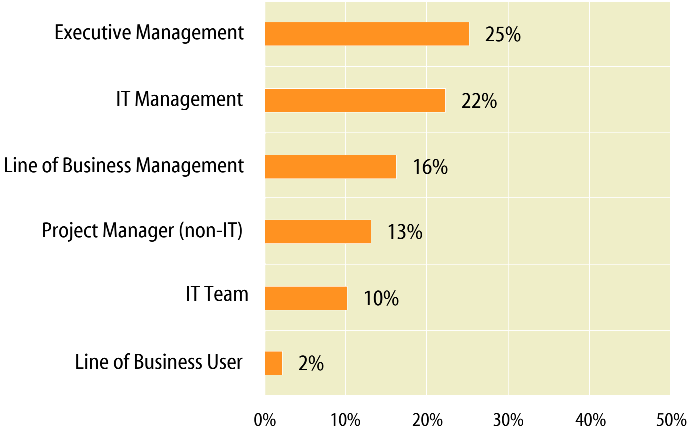
Statistics (con.)

Respondents by Company Revenue



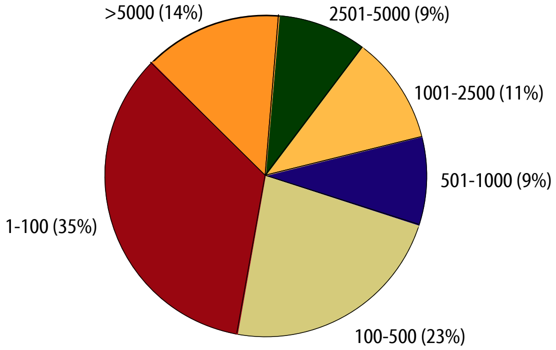
©2004 Delphi Group

Respondents by Organizational Role



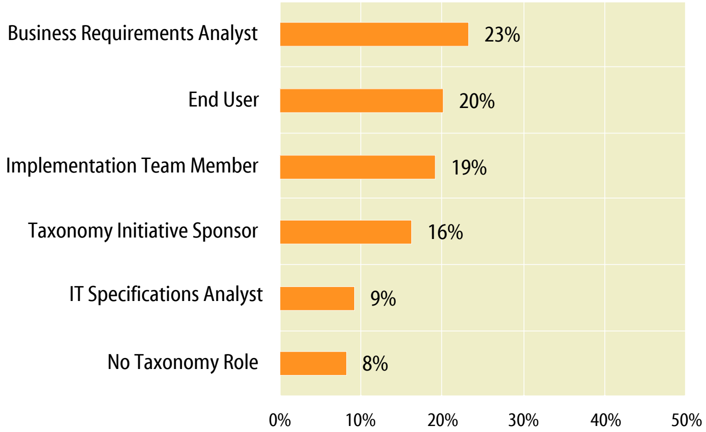
©2004 Delphi Group

Respondents by Location Employees



©2004 Delphi Group

Respondents by Taxonomy Role



©2004 Delphi Group

Technology Profiles and Use Cases

Publication and distribution of this report is partially underwritten by the technology suppliers who have engaged with Delphi Group in the Spring, 2004 Taxonomy Research & Education Program. This program is part of an ongoing series of Delphi Group studies, publications, educational programs, and conferences which seek to establish thought leadership perspectives and best practices in the area of information intelligence.

In this section of the report we provide an overview on six vendors and their product offerings that address the taxonomy and classification process. In each profile we include a profile of the company, the technology approach, the company's products, primary vertical markets if applicable, and use cases. The use cases help to illustrate how these technologies are deployed in real world settings.

Just as we have seen that there is no clear market domination from the point of view of mind share, it is important to consider that there is no universally accepted standard for evaluating the various algorithms or software configurations in regard to speed, accuracy, and scalability of taxonomy technology products.

When your organization is in the final stages of evaluation and has developed its short list of vendors, test the different solutions against a significant portion of your unstructured data, letting your users verify that the documents are categorized quickly and accurately and on a scale that meets your needs.

Autonomy – Autonomy IDOL Server

Autonomy

One Market Street
19th Floor
San Francisco, CA 94105

(415) 243-9955 phone
(415) 243-9984 fax

E-mail: info@us.autonomy.com
http://www.autonomy.com

Introduction

In its IDOL (Intelligent Data Operating Layer) Classification Server software, Autonomy offers a suite of capabilities to support automated categorization applications. The product line is fundamentally based on the statistical calculations associated with Bayesian inference, which Autonomy introduced to the commercial market in the late 1990s. With the core of its technology approach in the area of pattern recognition, in recent years, Autonomy has expanded its overall offering to support analytics and application development across the full range of unstructured information beyond text, specifically audio and video, in an automatic fashion.

Founded in 1996, Autonomy has been the most quickly growing business among the four publicly-traded firms in the industry over the past five year period. While one of the newest of these firms, Autonomy was the first company to focus on the comprehensive automation of the categorization process and the first to position its technology as a key infrastructure platform for realizing value from enterprise unstructured information resources.

Autonomy's customer base has grown to over 1,000, including enterprise corporate customers across the spectrum of industries and geographies, as well as significant implementations with governments and intelligence operations in a number of countries.

In 2003, Autonomy was the first firm in the market to introduce "free standing" business-focused information intelligence applications: one focused on more effective use of information in the call center context (Audentify), and one focused on aiding the implementation of compliance practices around electronic information in large enterprises (Aungate). These applications combine previously separate pieces of technology—voice recognition, concept search, expert identification, visualization, and automatic linking, for example—in a software configuration specifically tailored to target business operations or task domains.

Technology Approach

Autonomy's strength lies in a combination of technologies that employ advanced pattern matching techniques (nonlinear adaptive digital signal processing), utilizing Bayesian Inference and Claude Shannon's principles of information theory. Bayesian inference is a way of inferring an idea from a set of evidence, and Shannon's theory allows you to rank the importance of ideas. In the case of text, for example, the words 'black', 'white', 'sea', bird', 'flightless' being present provide strong evidence of the idea 'penguin' being expressed. Each word on its own is a fairly weak indicator of the idea to the extent that a single word is ambiguous (even the word 'penguin' is ambiguous – English chocolate bar, Batman character, Italian air conditioner, book publisher, bird).

The strength of this method is that the presence or absence of individual words doesn't significantly change the probability of the idea being expressed. This makes matching in the idea space more reliable and accurate than in the word space.

Autonomy does not require a mutually exclusive decision between automation and manual control. Combining automatic processing with a variety of human controllable overrides removes Autonomy from the purely automatic category of taxonomy vendors. Users can add rules, rename or modify taxonomy structures to suit their needs.

Because it does not rely on key words, it can work with any language independent of slang or regional variations. It treats words as abstract symbols of meaning, deriving its understanding through the context of their occurrence rather than a grammatical or semantic analysis.

Autonomy's Classification Server is part of a wider infrastructure, the Intelligent Data Operating Layer (IDOL). IDOL provides an integrated platform that integrates content through an understanding of it. Because Autonomy's technology understands the content it can do much more than simply classifying it. IDOL allows content to be cross-referenced, retrieved, classified and compared with other content.

The technology can profile users and their interests or expertise based on an understanding of the content they are working with. These profiles of users are technically identical to profiles of documents or taxonomy subjects, meaning that anything that can be done with content can also be done with users' profiles (e.g. a taxonomy of users and their interests can be created as well as one geared to content collections alone).

The IDOL infrastructure does three fundamental things:

- aggregates content from all unstructured sources across the enterprise or application domain
- creates a subject- or concept-level understanding of each piece of content which supports classification and other enhancement applications
- automatically associates content with defined categories or other custom application structures

Autonomy has engineered access to all unstructured repositories regardless of format or type (such as voice, video, text, XML, etc). IDOL builds a mathematical conceptual understanding of each piece of content for later use. It also allows each item to have an unlimited amount of metadata (source, date, author, subject heading, etc.) to assist in later manipulation.

A subject understanding can then be created by example from one or more pieces of content, or explicitly by sets of rules linked to metadata or keywords. Each of these would be a taxonomy element, or category, which is indexed into a server. As new information comes in, it is matched against the server's conceptual index (while still using all legacy boolean constructs and keywords a user may wish to define) as one operation (a highly scalable operation – each document in effect 'queries' the categories), and a decision is made based on the most closely matching categories. This enables the user to exercise whatever level of individual control they may want over a taxonomy component, while allowing them to make it a completely automatic process.

Product

Autonomy's IDOL Classification Server provides automated access to all of the information within an organization. As it aggregates the information from respective repositories, it respects all of the underlying security frameworks and rights definitions. Based on conceptual content, the Classification Server organizes it into a taxonomy, which is either manually defined on an application basis or automatically generated, or a blend of the two.

Autonomy's infrastructure technology deals with a static starting point of information as well as automatically dealing with ongoing changes in the information. In addition, it identifies new threads or themes of information as new content is added to the system.

The IDOL Classification Server offers a suite of processing modules:

- Automates categorization, cross-referencing, hyper-linking and presentation of information
- Dynamic updates: can be retrained on the fly eSummarize documents and recommend related article via hypertext
- Provides several styles of visualization interfaces to support navigation of highly complex information spaces.

After structuring the data into taxonomies, the Classification server tags the data automatically, adding relevant metadata tags based on an understanding of the content. Using these metadata tags a hierarchical tree can be generated pointing to the relevant documents.

Clustering

The Classification Server delivers the ability to automatically cluster information. Clustering is the process of taking a large repository of unstructured data, agents or profiles and automatically portioning the data so similar information is clustered together. Each cluster represents a concept area within the knowledge base and contains a set of items with common properties.

Autonomy has two types of automatic clustering:

1 Hot News Clustering – identifies the main topics of information present within the knowledge base

2 Breaking News Clustering – compares clusters from the previous periods and compares them to the current one. By identifying new clusters allows the automation of breaking news.

Cluster Mapping

The ability to visualize clusters is aided by two applets that identify the relationships between information clusters in one time period or between successive periods and sets of data. These are displayed as either a spectrograph or a 2D cluster map.

Primary Verticals

Autonomy's customers cover a broad range of industries and applications, including a large community of OEM relationships with other software ISVs. Publicc sector organizations, financial services, pharmaceutical/life sciences, and media companies are among the larger Autonomy customer groups.

Use Cases

Autonomy software is deployed in a large variety of applications. These include applications that support the information work of a range of business functions, including, for example: pharmaceutical and scientific research; government information management and intelligence analysis; call center service operations; enterprise video distribution and management; professional information publishing; and others.

The range of business uses for the technology is indicated in quotations from Autonomy customers quoted below.

"There is a lot of useful information out there, but the key to using it is to first make sense of it. This entails understanding the content, sending it to the right person and linking it to other pertinent information and people. Speed is another critical factor, thus all this should happen in a non-labour-intensive manner - that is, completely automatically. That is what Autonomy, uniquely, now enables us to do. It is the engine of our KnowledgeNavigator."

Jean-Pierre Krause, of Zurich Risk Engineering Group's Head Office

"Autonomy enables our software engineers and contractors to retrieve accurate and personalized information, which helps them design our earth- and star-observing platforms. The amount of information in our organization is akin to the number of stars in the universe – and we encounter more and more each day. Autonomy's technology automatically processes growing volumes of data and easily integrates with other products. The implementation has been smooth and straightforward throughout."

> Steve Nauss, associate head of NASA's Computing Environment &Technology Branch for Goddard Space Flight Center's Information System Center

"Novartis is a world leader in pharmaceuticals, healthcare, agribusiness and nutrition. That means we handle a vast amount of complex information daily, both from internal and external sources. Until now, that has cost us a great deal of time and money in manual processes, but Autonomy's technology can automate the whole process from start to finish."

John McCulloch, Manager, Executive Information System, Novartis

For more information on Autonomy use cases, visit: http://www.autonomy.com/c/content/customers/case_studies

Convera – Retrievalware 8

Categorization & Dynamic Classification Cartridge & Classification Workbench

CONVERA

Convera

1921 Gallows Road
Suite 200
Vienna, VA 22182

(800) 788-7758 toll-free
(703) 761-3700 phone
(703) 761-1990 fax

E-mail: info@convera.com
http://www.convera.com

Introduction

Convera is one of the pioneer technology companies focusing on the area of search, retrieval, and organization of information across the range of digital formats. One of the four major publicly-traded companies in this market, Convera is widely recognized for supplying innovative technology approaches to complex investigative requirements, particularly in the context of government and intelligence applications, but also in research-oriented commercial applications like those in pharmaceutical, life sciences, and financial services environments.

The company has recently engineered a new approach to categorization and classification that integrates these functions with Convera's broader platform of information retrieval technologies. The new offering, available as part of the Retrievalware 8 product platform, has focused on delivering what Convera refers to as dynamic classification, an advanced approach to support a user experience which enables users to utilize multiple taxonomic structures to help navigate and investigate electronic collections. The design centers on the idea of differentiating—both in technology and in presentation strategy—a large scale, underlying taxonomic structure from specific, dynamically-generated classifications which respond to specific user investigation priorities at query time. These dynamic classification structures can provide flexibility and navigational intelligence which allow the user to pursue the relevant categories in real time to guide his or her specific research requirements. This organizational dynamic is fully integrated with the search process.

Convera's categorization and classification software provides a new administrative module, or workbench, designed to give taxonomy managers and information architects an evaluation and control point for analyzing system behavior and effectiveness.

Convera's modular architecture offers implementers a suite of information technologies which share core functions like language awareness, security and provisioning frameworks, domain-specific semantic networks, and a variety of other processing modules that can be applied across text, image, voice, and multimedia data formats.

Technology Approach

Convera's categorization technology leverages established semantic models that mirror the way people associate different words and concepts with specific subject matters.

Convera utilizes stable authoritative taxonomies to automatically tag documents as they are acquired or input into the system. The software then provides the facilities to create dynamically-organized classifications for browsing and presentation based on user interest and user guidance of the discovery process.

The core design center is the concept of a "pool" of tags that allows the same set of documents can be re-organized on demand into any number of different specific classification schemes. This approach leaves behind the conventional ideas of identifying and maintaining a "universal" but static set of category definitions for the enterprise. In its place, Convera's technology offers users the opportunity to develop situationally-specific classifications based on the requirements of their current investigation.

Convera's categorization approach provides virtual domain expertise through the use of synchronized taxonomies and semantic networks in various languages. Convera's solution also provides categorization of documents across languages so that in multilingual environments users can intuitively navigate repositories of information that contain documents in multiple languages.

Convera's basic technology includes the following items:

- Categorization to support browsing and search
- Profiling (personal categories) to alert users to new relevant documents and changes in documents
- Industry-specific semantic networks that support concept-based categorization and search
- Pattern search through Convera's proprietary Adaptive Pattern Recognition Processing (APRP) for recall
- Synchronizers that access content in a wide range of repositories to find and update all the content in the enterprise lowered.
- A security model that supports secure access to those disparate repositories
- Multilingual support through a plug-in architecture for categorization across languages
- A set of pre-configured taxonomies which support a range of industry and professional domains
- A Cartridge and Classification Workbench for creating and/or customizing classification structures
- Software development kits (SDKs) and application programming interfaces (APIs) that allow customization and integration of categorization and search

Concept-Based Categorization

Concept-based categorization enhances "search" by addressing the problem of language ambiguity—where several different words can be used to express essentially the same concept, or where the same word can have multiple meanings associated with radically different subjects.

In the context of working with taxonomies, whether pre-configured or built on a customized basis, concept-based categorization enables the controlled expansion of terms within a taxonomy category rule set. For example, a search for "international commerce" will find documents that contain terms such as foreign trade, import, export global mercantilism and free trade. By using the concept-based rules engine, an administrator can take advantage of subject matter expertise to support a more accurate rule set that is automatically updated and improved with subsequent versions of the taxonomy and the semantic network.

Concept search clarifies the meaning of query words through analysis of surrounding words (e.g., the word tank when surrounded by words such as military and vehicle is more likely to be a fighting vehicle and less likely to be a container for holding fuel).

A key technology behind concept-based categorization is the RetrievalWare Semantic Network, a collection of approximately 500,000 English words that expands to over 1.6 million semantic relationships and idioms that are organized by concept. RetrievalWare's PowerSearch feature allows users to control word expansion of query terms in the Semantic Network. Users can also choose specific meanings for their query terms. For example, when using the term for banks in a category's rule set, an administrator can easily instruct RetrievalWare to use only those other meanings associated with financial institutions and not those dealing with the bank of a river or a turning aircraft. In addition it is possible to select individual terms within a meaning to sharpen the rules even more.

Support of Languages and Subject Domains

The RetrievalWare platform provides cross-lingual and cross-domain categorization and search. For example, when an English-speaking financial analyst wants to research the potential risk of making a loan to a German pharmaceutical company, there may be relevant documents created in other languages containing technical terms that could impact his decision. With cross-lingual and domain-specific categorization, the analyst may find articles about this company in English, German, Italian and French newspapers that discuss recent loan defaults and technical advances.

RetrievalWare includes modular support for more than 25 languages and many domain-specific semantic networks through it Cartridge architecture. This feature acts like a multilingual subject expert that not only understands the concepts and terms in a subject area, but also knows how they are used across languages.

Concept-based categorization that takes advantage of cross-lingual search and domain-specific semantic networks reduces the need for authors or administrators to tag documents, because this feature automatically provides these relationships. For example, if a category rule uses the term "central" meaning in or near a middle position, not only could documents containing middle, medial, midway and other relevant English terms be included, but also it could include documents in other languages using terms such as middelpunt (Dutch); milieu (French); Mittelpunkt (German); medio (Spanish) and mezzo (Italian). If the medical domain-specific term bronchus was used in a category rule, other English terms such as windpipe, respiratory tract and bronchial tubes could be automatically be included in the rules, along with terms from other languages such as bovenlip (Dutch); arrière-gorge (French); Atemwege (German); bocado de Adán (Spanish) and bocca (Italian).

Preconfigured Taxonomies

Convera offers preconfigured taxonomies in the areas of: Genetics, Finance & Business, General Enterprise, Technology and U.S. Government. These taxonomies contain industry standard thesaural content enhanced by Convera linguists and taxonomists.

The Convera Genetics Taxonomy covers gene ontology, cellular components, molecular functions and biological processes.

The Convera Finance & Business Taxonomy describes the world of finance, including corporate, personal, investment, regulations, accounting and banking and is recommended for any application involving a financial department, financial research, or compliance.

The Convera General Enterprise Taxonomy includes terminology for sales, marketing, management and human resources.

The Convera Technology Taxonomy includes technical vocabulary for many fields, including math, computer science, engineering and electronics. Civil engineering firms, pharmaceutical companies, high-technology manufacturing companies, medical instrumentation firms and software development companies are examples of enterprises that find this taxonomy useful.

The Convera U.S. Government Taxonomy is designed to organize information and help citizens who are not experts on government subjects find specific data on citizens' rights, public services, legislation and government structure. Implementers include government agencies responsible for providing public services as well as government departments responsible for making information available to citizens.

Convera taxonomies can be easily split into more targeted classifications to meet the specialized search and discovery needs of an organization. For example, the Convera Genetics Taxonomy can be split into cellular components, molecular functions and biological processes, each of which can be subsequently combined with any of the numerous other classifications Convera makes available.

This approach underlies Convera's capability to dynamically classify information based on a specific searcher's interests, by cross-referencing several classifications at search time. The taxonomies come into play in the user experience in which Convera's integrated search and categorization technology "slices and dices" results quickly into one or more selectable classification views, displaying information in easily understood table formats and incorporating information from completely disparate database or taxonomic locations. For users, this format helps to map relationships between data that may not have been known previously, and can enhance the possibility for serendipitous information discovery.

The new Convera taxonomies are enhanced versions of taxonomies that contain unique controlled vocabulary and classification structure developed by a leading taxonomy builder, ProQuest Information and Learning www.il.proquest.com. The ProQuest framework for each industry-specific taxonomy enables precise mapping of data and a more exact search of subject fields. Each taxonomy powered by ProQuest is regularly updated to capture the latest industry vocabulary and the most commonly used language for describing topics areas.

Convera Cartridge & Classification Workbench

The new Convera taxonomy workbench allows non-technical users—information architects, librarians or subject matter experts—to automatically import specialized taxonomies to organize disparate enterprise information. Utilizing a graphical interface, users can highlight important data for classification, identify taxonomies, and visually monitor the process of creating classifications and populating categories with data.

The Cartridge & Classification Workbench offers facilities in three major functional areas:

- Developing Completely New Taxonomies
- Importing Existing Taxonomies and Thesauri
- Testing, benchmarking, and tuning taxonomies and classifications

To leverage data organization models that are already in place, users can integrate and modify existing taxonomies or thesauri. Automated import, generation, combination and pruning wizards aid in accelerated taxonomy and classification development.

Testing, benchmarking, and tuning taxonomies and classifications is a prerequisite for successful applications, and Convera offers a suite of taxonomy quality measurement tools. These tools include static and dynamic benchmarking tools. All categorization processes, including normalization, latching, idiom detection, disambiguation and ranking can be evaluated through the workbench analytics.

Security

RetrievalWare provides a security infrastructure that respects the native security of the disparate document repositories accessed by RetrievalWare through a single log-on. RetrievalWare addresses the issues of authentication, distributed, cross-repository and document/library level security for categorization and search. RetrievalWare's cross repository, library, and document-level security ensures that only the documents that a given user should see are displayed, thus preserving the intended goals of the enterprise security framework.

Products

Convera RetrievalWare is an enterprise-class cross-lingual multimedia categorization and retrieval platform. RetrievalWare can index, organize and securely retrieve contents from a wide range of repositories and document types, including file systems, over 200 document types, groupware systems, Web pages, XML, relational databases, leading document management systems and video and scanned paper documents. RetrievalWare allows users to use natural language queries as a basis for categorization and retrieval without concern for search syntax, word usage and other details for content in over 50 languages.

The platform's "cartridge" architecture allows the system's functionality to be tailored to best address use cases. Cartridges are made of up three elements: specific industry domain definitions, modules to support unique localized language processing and retrieval requirements, and pre-configured or customized taxonomies.

Tools for integration and modification of RetrievalWare

RetrievalWare offers a series of toolkits that enable RetrievalWare to be used by System Integrators, OEMs, corporate developers, and Convera's Integration Services Group to modify, extend, or embed RetrievalWare for more custom solutions. The RetrievalWare SDK contains toolkits to enable the integration of RetrievalWare servers as well as the creation of custom client interfaces or modifications to out-of-the-box interface clients.

Convera's RetrievalWare runs on most major operating platforms (Microsoft windows based and UNIX) and uses industry standard Web servers, RDBMS and browsers. Although almost any standard business computer fulfills the minimum system requirements, each customer's requirements and computing environment are unique and therefore may need additional resources to achieve the desired level of performance and functionality.

Primary Verticals

Convera's customers cover a broad range of industries and applications. The company over the years has built a particularly strong set of implementations with the US government, particularly in the intelligence and security areas. Convera also has customers among financial services, pharmaceutical/life sciences and media/publishing firms.

Use Cases

First generation knowledge management solutions focused on the low-hanging fruit of explicit knowledge—knowledge that is easily codified and conveyed to others. Explicit knowledge is much less valuable, but far easier to "harvest," than tacit knowledge, that elusive, personal, experiential knowledge residing "in people's heads." The technologies associated with document, data, and content management are aimed at meeting the challenges associated with managing explicit knowledge while the next generation of knowledge management solutions incorporates an additional focus on realizing the value of tacit knowledge—a major challenge for most enterprises.

The continuing turnover within the rank-and-file of every organization, from federal agencies to the Fortune 500, hastens the need for an effective means of capturing and leveraging the "know how" and collective wisdom of personnel. Federal agencies are at particular risk of losing this wisdom, resulting from employee turnover and the "graying" of tacit knowledge.

Agencies are experiencing the actual loss of years of accumulated knowledge as workers leave or retire, taking their experience and tacit knowledge with them. Examples of these challenges have occurred recently at NASA, where years of attrition among its scientists ultimately meant NASA no longer had the expertise to send a man to the moon. Other examples abound: Sandia National Labs, where it was reported that younger employees may lack the capability of safely handling or disarming older nuclear weapons, and USAMRID where the heightened threat of biological weapons has shined new light on whether they still have the tacit knowledge gleaned from the 1960s experiences with variola and anthrax.

NASA uses RetrievalWare to ingest, archive and manage Space Shuttle and International Space Station video content in order to provide real-time and post mission analysis and re-purposing of the video over the NASA intranet.

The Naval Research Laboratory (NRL) online digital library, TORPEDO, illustrates the sheer magnitude of the problem. Powered by RetrievalWare, TORPEDO provides the NRL's 3,000 employees with access to the full content of 6,000 technical journals, 2,000 NRL press releases dating back to 1968, 10,000 NRL-authored articles and conference papers dating back to 1944, and nearly one million articles from 800+ journals. NRL users can use full text concept searching or fielded searching, optionally restricting searches to specific collections, and browse the results whether in the office, at home, or traveling.

The U.S. Air Force Research Laboratory's (AFRL) technical library is an information center for space and missile related and directed-energy weapons technology. AFRL researchers require access to a myriad of information residing in repositories including other scientific and technical databases, relational databases, LAN-based information repositories, other digital sources, and paper-based digital archives. These researchers rely on RetrievalWare to discover and retrieve accurate, relevant information from these many sources and across many forms such as structured and unstructured text, static text, images, and video.

Among commercial enterprises, there are equally challenging information management challenges being addressed by Convera software. In financial services, for example, organizations face a series of issues that new compliance regulation has brought to the forefront of management attention. These include: vast quantities of e-mail data, increased scrutiny in reporting requests by regulatory agencies, need to monitor, detect and eliminate non-compliant activity, the need to protect intellectual property from unauthorized dissemination, the need to protect enterprise communication systems from employee misuse, rapidly growing message traffic across different repositories, geographically distributed locations and diverse data-types, maintaining organizational barriers between investment bankers and financial analysts, and proactively monitoring and halting non-compliant activities, rather than waiting to discover and report problems after the fact. Convera software has application in these areas of compliance and threat detection.

In the pharmaceutical and life sciences areas, the importance of the core research function suggests the use of the most effective information discovery technologies. Convera's taxonomic and semantic resources make it possible to uncover potentially hidden relationships between entities in the research practice. For example, a researcher in a pharmaceutical company may attempt to identify which tractable targets have already been studied or tested with marketed drugs or compounds by exploring numerous scientific publications and abstracts. In a Convera application this can be facilitated by combining taxonomic content containing disease descriptors, the dendrogram of a given protein family (based on sequence similarity), CAS registry numbers, etc. By using the discovery platform and dynamic classification approaches, researchers can win back time to focus on their core work such as examining the classification of abstracts or text from publications looking for correlations between compounds, mechanisms of action, cellular efficacy, animal efficacy, etc.

Indraweb – ProcinQ Server

Introduction

Indraweb was founded in 1999. Their ProcinQ™ Server is approaching the third major revision (slated for mid-Q2 2004). The primary differentiator for Indraweb is their stance as a provider of pre-built taxonomies, which they call "Concept Indices". These Concept Indices are built upon the consumption of licensed content (what they term "Taxonomy Rights") from the largest reference work publishers in the world, and contain all of

Indraweb

15 Maple Avenue
Paoli, PA 19301

(610) 251-1076 phone
(520) 843-5345 fax

E-mail: info@indraweb.com
http://www.indraweb.com

the uniquely relevant words and/or phrases (i.e. "concept signatures") for every node in the taxonomy.

Pre-built Concept Indices are available from Indraweb across a variety of areas, including: medicine, chemical processing, pharmaceuticals, legal, defense intelligence, religion, environmental, food processing, skills, manufacturing, research, business, security, and legal. They deliver the pre-built/licensed taxonomies as well as output of analysis of your own content in Topic Map format (XTM)—an emerging industry standard that "opens" their platform on both the consumption and output ends. At this point, standard interchange formats are not a requirement for most buyers of these solutions, but for "future-proofing" your investment, this is something to keep an eye on.

The most recent updates to their offerings are: a cross platform editor's desktop with taxonomy content management tools and visualizations for trending analysis/reporting—functionality that previously existed strictly as APIs requiring custom application development.

Technology Approach

The primary mode through which Indraweb provides taxonomies is the licensing of Concept Indices, as most professionally published reference material is already organized in a semi-structured format.

By consuming semi-structured content across a wide variety of publishers, defining a "topical space" (such as pharmaceuticals), consistent structures begin to emerge from that process, and in the opinion of Indraweb high-quality Concept Indices are the end result. This methodology is what Indraweb calls Orthogonal Corpus Indexing (OCI), which is an algorithmic process that uses trusted reference sources as a way to bound the topic space.

The ProcinQ system is the core repository for classification, taxonomy and thesaurus information in Indraweb's offerings, and all content flows through this basic platform.

Auto-classification is the ultimate aim of Indraweb's system, although as with most solutions in this area, they provide tools/interfaces that allow for manual classification (machine-assisted), or for refinement of the auto-generated taxonomy/classification schemes their platform can create.

Indraweb uses a combination of linguistic and statistical methods for classification scoring on top of the OCI algorithm. These scores are used in combination with Receiver Operating Characteristics (ROC—decision making science originally developed for signal processing in radar applications) to choose the best combination of scoring algorithms for a specific application, balancing the costs of false positives versus false negatives. ROC uses economic costing as a way to measure whether a concept should be included. Explicitly built into the system are variants on the K-Nearest Neighbors statistical analysis. They have an open interface for external classifiers and have had customers experiment with Bayesian classifiers as well as other methods, although those are not "native" to the offering.

They provide a Concept Query Language (CQL—a "SQL-like" language) for data/concept mining as method to explore their concept indices manually or via machine input through linkages to other systems via APIs, or using their own administration and management tools that use these same APIs as the connectors to manipulate and examine content within their system.

Other facilities beyond core taxonomy/classification administration and maintenance offerings include navigation style interfaces to the taxonomies, federated and complex searches (Indraweb's own "enterprise search") that searches the full-text space, concept space, as well as any tagged meta-data (such as titles, authors, pre-applied keywords), and alerting capabilities with built-in notification scheduling (alerts are built using CQL, mentioned earlier, either manually or through interfaces provided by Indraweb or custom application interfaces using their APIs).

Indraweb's platform supports "stackonomies"—multiple layered taxonomies (indices) that use terminology from various audience constituencies. Some examples: a research scientist and a marketing manager with varying needs to understand the content in the repository and communicate those results to others inside or outside the organization can view different hierarchies built on the same original content; or in the life sciences environment, researchers and doctors will have different preferences in using chemical taxonomies and/or medical taxonomies that describe the chemical components that make up a new drug and the symptoms or benefits in treatment.

Products

ProcinQ™ Server

ProcinQ Classification Engines

Taxonomy ToolBench

- Editor's Desktop
- Concept Query Language (CQL)
- ProcinQ Portal

Concept Alert System

Concept Index™ Subscriptions

Knowledge Harvesting Service

ProcinQ™ – Intelligence Solutions

The ProcinQ engine is accessed via APIs that return standard XML data structures for integration into application architectures. The engine is written in Java and is deployed in production on both Windows 2000 and Linux Servers. All concept data is stored in an underlying Oracle database. Several user interfaces are available for the server: for search and directory browsing, the ProcinQ Portal (web-based) is provided, implemented as Java servlets and running in production on both Jrun and Apache Environments. Indraweb also provides a full featured editors desktop (deployed as native Java GUI applications) for taxonomy building and data mining, with separate installers for Windows, Macintosh and Unix / Linux operating environments.

Concept Index™ licenses are based on yearly subscriptions, although the contents of the indices may be refreshed on a more frequent basis depending on the content area.

Primary Verticals

Research & Scientific (Biotechnology, Chemicals, Pharmaceuticals)

High-tech manufacturing (Aerospace, Automobile, Electronics)

Intelligence (Military, Counter Terrorism, Competitive Intelligence, News)

Use Case

DFI International, an integrator and consulting firm in the DC area specializing in National Security and Homeland Security issues, worked with an intelligence agency to build an "Intelligence Portal" on the Indraweb platform for the purpose of "open source exploitation" (academic sources, trade magazines, unclassified sources, public domain information rather than classified information already consumed via other means) which intelligence agencies use to round out gaps or verify and enhance the classified material they receive from other areas.

The client required a knowledge management (KM) application built upon a taxonomical-based solution rather than a more traditional/legacy "full-text search only" solution to be able to manipulate the information via various taxonomies to suit high level and very low-level, detailed vocabularies and content in chemical and biological weapons development, proliferation, and warfare. Once the content has been "enriched" via this solution, human intelligence analysts manage the content seen and perform further analysis and routing to appropriate areas.

During the course of the setup and integration effort, roughly 3,000 topics were developed manually by DFI and the client, with Indraweb providing another 80,000 topics to round out the topic space. Further revision of the taxonomy – ignoring sections of the taxonomy to narrow the depth of the topic space to the client's accuracy requirements, and removing superfluous topics to their very specific needs was necessary, but the ability to leverage a pre-built taxonomy was cited as a major accelerator to the project.

As of this writing, the use case is a self-contained pilot installation, with all components installed a Linux web server using an Oracle RDBMS for the content and taxonomy store. The client is using the pre-built general-purpose interface provided by Indraweb for general administrative duties, and an end-user interface that was custom built via Java Server Pages (JSP) by DFI. On the roadmap for this installation is to provide integration between agencies to share information per Homeland Security directives.

Prior to this pilot project, the client used a packaged application that is no longer available, and desired to expand the comprehensiveness of the solution itself and the ability to provide a small labor force with greater capability to process volumes of information. Primary goals were to keep up with the inbound information in a timely manner and to maximize the effectiveness and responsiveness of analysis by the intelligence analyst teams which then in turn enabled faster response time to intelligence discovered and correlated through this system.

Stratify – Stratify Discovery System

Introduction

Stratify supplies solutions for automatic taxonomy generation, automatic and machine-assisted classification (using multiple classification algorithms/techniques), analytics and visualization tools for exploring categorized/classified content, and out-of-the-box interfaces for individual use of their system for navigation or search, as well as web services, JAVA and C++ APIs for integration into other systems - focused on the problem of unstructured data management.

First offering their Discovery System platform in October 2001, the latest revision is version 3.0, released in August 2003. Recent additions to the

Stratify

701N. Shoreline Blvd.
Mountain View, CA 94043

(800) 988-2686 toll-free
(650) 988-2000 phone
(650) 988-2159 fax

E-mail: info@stratify.com
http://www.stratify.com

solution line include Stratify Analytics, which provides unstructured data-mining, reporting and visualization tools that allow individual analysts and business users to uncover hidden, indirect patterns and relationships between entities such as people, locations and organizations (n-degree mapping, otherwise known as non-obvious relationship analysis.)

The most recent solution addition, the Stratify Legal Discovery Service - an ASP solution for legal eDiscovery - provides electronic document discovery for document review and production, case assessment and case strategy development and analysis for attorneys and their support staff. The Legal Discovery Service is the first verticalized offering from Stratify, with further vertical developments in the production pipeline, based on the Discovery System Platform. Enhancements in the interface specific to Legal Discovery Service have not yet been folded back into the base platform, but are scheduled for their near term roadmap.

Technology Approach

Stratify's solutions are aimed towards multiple user groups (including research analysts, managers, general business users and technical integrators) and provide coverage over the entire lifecycle of taxonomy and classification needs - needs that evolve over time as content changes and users gain more facility with the solution. The system enables editors and/or administrators to manage the entire lifecycle to create, define, test, publish and refine taxonomies using a combination of automated techniques, reporting and human review at every step. Stratify's recognition that taxonomies and their classification definitions cannot remain static in the face of changing conditions, and their delivery of tools to automatically refine and optimize existing taxonomies, is a key component of Stratify's approach to the taxonomy and classification challenge.

Enterprise search capabilities are provided out-of-the-box via Fast Search & Transfer or the Windows Indexing Service. Other search engines can be integrated using the API. Stratify's taxonomy and classification capabilities are integrated with the search engine and can be delivered within traditional search or via their own web-based interface.

A number of engines are applied to the inbound content to prepare it for consumption and accurate categorization. Document converters (for over 225 document types) parse the documents, boilerplate text (repeating headers/footers for example, or repeating legalese) ignored, stop words and phrases are removed, "dense text" is filtered out (content from HTML documents for example may contain ads, menus or other "non-content" items), titles (explicitly tagged or implicitly derived) of documents are extracted, and near and exact document duplicates are detected and filtered. The system is Unicode compatible, and currently provides stemmers for all Indo-European languages and Arabic.

To jump start a taxonomy project, Stratify will license pre-built taxonomies by domain specific areas (vertical industries), or a bundle of their entire collection of taxonomies. This may be a good method to burst through the initial (and inevitable) political discussions regarding which categories/terms are to be used in your organization. Pointing to a “standard” scheme is one way to defuse that particular issue. The reality is that taxonomy projects frequently partially fail AND succeed the first time out – and version two of the project is when significant progress is made. Consider the pre-built taxonomies as a way to jump from ground zero to version two.

Entity extraction with Stratify is flexible, with pre-built extraction of people, locations, organizations and other entities already handled within their extraction engine. Entities can also be defined by users to handle specific “known entities” such as product codes, social security numbers or other entities that can be statistically codified.

Integration points within the Stratify platform and their pre-built solutions as well as to third-party applications (such as enterprise search), are accomplished via Web Services (SOAP and WSDL), XML, as well as Java and C++ APIs, allowing access to the Stratify pipeline at numerous connection points, depending on the system being integrated to/with, and the outcome that is desired.

The Stratify Legal Discovery Service is offered as an ASP (Application Service Provider)/hosted service specifically for electronic discovery applications for attorneys and General Counsel in the legal domain. This solution leverages the core technologies of the Discovery System platform in conjunction with specific legal workflow capabilities for this specialized vertical solution, bootstrapping electronic discovery by automatically organizing the documents into a taxonomy of relevant concepts. Review workflows to tag documents and/or assign them to attorneys for further analysis (who can utilize Stratify’s visual analysis tools) are key components to this browser-based solution. In addition, the system maintains complete audit trails to ensure chain of custody records.

The Stratify Analytics application is a visual data-mining tool for unstructured information with interfaces both for business users and more advanced power users. The application includes many capabilities for in-depth analysis above and beyond the default facilities of the Discovery System. The Analytics Home interface enables users to create Personal Watchlists based on concepts, people, organizations and locations, or other entity classes that are defined. Users also have direct access to Most Active and Most Popular entities for immediate perusal.

Visual Heat Maps and Network Graphs, provided by Java applets within a web-based interface, are two of the visual tools used to graphically represent the structure of complex document universes as well as relationships between various entities. The Heat Map view represents entities as rectangles, with size indicating the number of total documents per entity and color (varying from red to blue) how “hot” or active the entity has been within a given time frame. Stratify Analytics can proactively identify N-degree relationships between any entities in the system (beyond 1-degree, direct connections), which can then be presented within Network Graphs, in which edge lines between entities show the number of connections between them based on the content.

Products

Stratify Discovery System 3.0

Stratify Analytics

Stratify Classification Server

Stratify Notification Server

Stratify Legal Discovery Service

The Stratify Discovery System is the base platform of Stratify's offerings, and includes the Taxonomy Manager administrative interface and Crawlers. The Classification Server, Notification Server and Stratify Analytics are licensed separately. The Stratify Legal Discovery Service is offered as an ASP (Application Service Provider)/hosted service currently.

Stratify Discovery System 3.0 server software runs on Windows 2000 and 2003 server platforms. The Stratify Taxonomy Manager application runs on Windows 2000 Professional and XP Professional platforms and is the single point of access for all administration tasks. The Stratify Analytics interface is entirely browser-based, requiring Internet Explorer 6+.

Availability of Products

Stratify Discovery System 1.1 - October 2001

Stratify Discovery System 2.0 - October 2002

Stratify Discovery System 3.0 - August 2003

Stratify Analytics - October 2003

Stratify Legal Discovery Service - November 2003

Primary Verticals

Law
Media/News Aggregation Services
Government and Intelligence Community
Oil and Gas

Use Case

Thomson Dialog, in the Publishing and News Aggregation industry, provides through their on-line offerings, complete, timely and accurate business information to its customers from a wide variety of sources. Faced with an information explosion and the task of integrating disparate technologies and information repositories acquired through mergers, Dialog needed a taxonomy platform that could meet a set of disparate and often contradictory requirements.

The senior management at NewsEdge (soon to be acquired by Dialog) realized that no matter how many people they applied to manual classification, the information explosion would soon overwhelm the human capability to process the continuing rise in information. NewsEdge engaged in a comprehensive study that evaluated a number of solutions and approaches for dealing with this problem. Mergers and acquisitions (in the context of Dialog and Thomson) changed priorities and requirements. Dialog realized that it had to combine these various companies, products and technologies to deliver a unified and complete "superstore" of information services.

Stratify proposed that a single, complete and standard taxonomy be developed and maintained. This taxonomy was designated the "canonical taxonomy." The canonical taxonomy was very detailed (over 3,000 hierarchically related topics). From this core taxonomy other "presentation" taxonomies could be developed and deployed to suit the particular requirements and profile of a large variety of users and information products and services.

Using the Stratify Discovery System 3.0 solution, Dialog has been able to expand its revenues and customer satisfaction by offering a wider variety of information service solutions to them. Dialog's customer's new ability to choose from a variety of information repositories and services has increased Dialog's revenue. Customers are able to choose more services that meet their specific business intelligence requirements by discovering relevant information from a wide variety of sources. Dialog's customers, in effect, self cross-sell themselves additional Dialog services enabled by the core technology found in Stratify's Discovery System 3.0 solution.

Teragram Corporation – Teragram Categorizer

Teragram Categorizer
Teragram Taxonomy Manager

Introduction

If there is one idea to keep in mind with Teragram's solution, it is "neutrality." Due to many factors, primarily from the strong OEM background of the company, Teragram's solutions are componentized and can connect with many systems as a holistic offering, or as individual pieces, and at heart, they make no assumptions about the environments which they will be deployed within.

For example, some classification/taxonomy systems may insist on tagging individual documents directly, while others lean towards central "meta repositories" based on perhaps an Oracle database or commercial content management system – to Teragram, either is acceptable, and possible.

This same attention to an easy co-existence in a wide variety of portals, content management systems, enterprise search environments, and other integration efforts is central to their offering.

Technology Approach

The primary technological means to Teragram's taxonomical ends is through linguistic and semantic analysis, a deep understanding of the meaning and construction of words and phrases, independent of the specific language in question, as well as (recently added) statistical and rules-based algorithms, rather than any single approach.

While combining methods is not unique to Teragram, their historical strength in the linguistic/semantic arenas is a recognized differentiator from most competitors in this arena.

Teragram internally maintains dictionaries and taxonomies of about 5 million terms, derived via linguistics, statistics and rule-based means – which allows them to do completely automatic taxonomy creation, but also allows for manual override and clarification by human operators or "domain experts" as the case may be.

The components of Teragram's offerings for lifecycle taxonomy management and search and related functionality, involve the following modules:

Teragram Taxonomy Manager – client/server interface and underlying engine for navigational, directory-based views on the taxonomical structure.

TK240™ – the primary and overall administration interface for taxonomy management, entity extraction (authority list and rules) for everything except search functionality administration

Teragram Editor Workbench – thin-client, native web application, no computation done on client. Search is embedded in this interface, auto-summarization (which can be modified to be more specific by human editor[s]), shows suggested categories and related categories, machine-assisted with workflow functionality for approval cycles

Teragram Entities Extractor™ - entity extraction is a further component of parsing unstructured data, and Teragram supports the many variants of entities, such as identifying people, places, things, events, products, company names and so on. Entities may be identified through rules, grammar, or "Authority Lists" (such as publicly traded companies and their stock tickers, baseball player lists, competitive company lists, and so on). This module is bought and licensed separately

Teragram
10 Fawcett Street
Cambridge, MA 02138

(617) 576-6800 phone
(617) 576-6888 fax

E-mail: info@teragram.com
http://www.teragram.com

Teragram Real Time Alerts – an alerting engine tuned out of the box for e-mail alerts, this module ties into the categorization, extraction and other modules. A typical alert scenario might be: alerting on new documents containing specific topics from the taxonomy or authority lists – customers have used this functionality for both public and internal facing applications. The alerting engine first quantifies the alert event, personalizes it to the audience to be delivered to (person, team, company), and outputs the alert through a delivery engine – which is theoretically

Enterprise Search – With version 3, released in late 2003, this is now fully intertwined with the other components if so desired.

Teragram's solutions work with 25 languages , are Unicode compatible, handle double-byte languages (including Asian languages such as Chinese, Japanese, Korean), Arabic, all major European languages and Eastern European Languages.

Products

Teragram Taxonomy Manager

TK240™

Teragram Editor Workbench

Teragram Entities Extractor™

Teragram Real Time Alerts

Enterprise Search

Teragram's solutions run on Windows, Unix and Macintosh platforms, and access/integration points to the individual modules are via APIs accessible in C, Java, and in some circumstances, Perl or JavaScript for their "browser-embeddable" modules such as client-side spelling correction.

As with many taxonomy solution providers, Teragram provides automatic taxonomy generation, manual manipulation or creation of taxonomies, as well as pre-built and licensed taxonomies (purchased separately – in areas such as IPTC (International Press Telecommunications Council – for the news/publishing industry), MeSH (Medical Subject Headings) and other vertical/domain-specific areas.

Primary Verticals

Government and Intelligence Community

Human Resources

Large Financial Institutions

News/publishing – most major US papers

Yellow Pages

Pharmaceuticals

Use Case

Teragram customers include media publishers, newspapers, hardware manufacturers, financial institutions, pharmaceutical companies, U.S. government agencies, information providers (portal, search, yellow page, etc.) and other major companies where the accurate management of information is critical.

An example of a corporation using Teragram technologies is the World Bank. The World Bank selected Teragram Corporation, as they required a taxonomy solution to support multiple taxonomies and multiple languages across the World Bank's large number of document repositories and content management systems. The mission of the World Bank is "to fight poverty and improve the living standards of people in the developing world. It is a development Bank, which provides loans, policy advice, technical assistance and knowledge sharing services to low and middle income countries to reduce poverty" (http://www.worldbank.org -- about us). The World Bank collects economic reports from around the world. These reports are summarized and categorized into multiple taxonomies.

Verity – K2 Enterprise

Introduction

In its K2 Enterprise offering, Verity provides a full set of technology capabilities for organizations implementing taxonomy strategies. The component elements include: content discovery services, pre-configured taxonomy structures, taxonomy construction tools, theme discovery technology and topic mapping, automated classification, business rule-driven classification, multiple taxonomy relational deployment, dynamic taxonomy views, and taxonomy maintenance and management facilities.

Verity is the largest of the publicly-traded software suppliers in the information retrieval market, and has developed relationships with over 11,000 clients on an international basis over its 16-year history of marketing search and information analytic software. In addition to K2 Enterprise, the company's product line includes the Verity Ultraseek search engine, a widely installed product geared primarily to site search and Web content. Verity also recently acquired Cardiff, a content capture and business process automation software provider, whose products (branded Verity LiquidOffice, Verity LiquidCapture, Verity TeleForm, and Verity Medi-

Verity™

Verity

894 Ross Drive
Sunnyvale, CA 94089

(408) 541-1500 phone
(408) 541-1600 fax

E-mail: info@verity.com
http://www.verity.com

Claim) are now supported by the combined organization. In addition, Verity recently acquired the strategic assets of NativeMinds, which adds natural language technology and a customer-facing service application to the company's product portfolio.

Verity's categorization and classification software includes both market-tested modules with many production implementations and newly-engineered facilities designed to incorporate new research in automated classification algorithms, profiling, network linking, and recommendation approaches, and the use of taxonomic structures in information discovery processes. The system also provides a full-function administrative module designed to provide a collaborative environment for taxonomy managers and information architects to carry out system development, maintenance, and operations workflows.

K2 Enterprise's modular platform architecture offers implementers an interoperating collection of information technologies which share the core environmental functions (e.g. language awareness, security and distributed processing frameworks) while enabling the use of selected domain-specific taxonomies, automatic and/or manual categorization approaches, parametric or federated search, profiling and alerting, and a variety of other processing modules to mold applications which can be applied across text, image, voice, and multimedia data formats.

Technology Approach

Verity takes the view that content organization is a four step process.

1. Build the taxonomy

Verity recognizes that different businesses have achieved different levels of practice in the area of content organization. Some may already be operating with taxonomies which have been developed in house or with industry standard taxonomies, while others are looking to begin the process of organizing their information, creating categories and classification structures. To accommodate these variances in practice, Verity offers implementers the flexibility to implement taxonomies in a number of ways. The system has facilities to import existing corporate taxonomies, to input industry taxonomies, and to create completely new categories and taxonomies. Verity also supplies pre-configured taxonomies—called Verity Taxonomies—which offer structures developed by domain experts across a variety of fields and subject areas. In addition, imported taxonomies can be modified and used in conjunction with custom-developed structures, and/or the corpus of information can be analyzed and relevant concepts automatically extracted.

2. Build the software model

Before a taxonomy can be populated with documents, a model defining each category must be built. Companies with limited resources for design and ongoing maintenance of structures may consider a completely automatic approach that relies on software to discover and populate categories and taxonomies. Others may choose to combine automatic methods with human guidance and maintenance to more closely accommodate specific business requirements. Still others may opt to build software models entirely manually using extensive and detailed business rules to identify categories and their content.

The Verity software supports any of these approaches, or combinations of approaches. Models can be built using a number of methods: rules defining categories can be generated automatically, rules can be imported from existing taxonomy models or industry taxonomies, and/or domain experts can build new rules or modify imported and automatically generated rules.

3. Populate the Taxonomy

The classification process takes place in an integrated manner with the indexing analytics used to support search operations. Using the model which the organization has implemented utilizing the methods described above, Verity automatically populates the taxonomy structure and categories with documents during the indexing process.

4. Use the Taxonomy

Populated taxonomies can be used in a number of ways. End users can browse through them by drilling down through broader categories to more focused concepts and individual documents, or they can narrow search results by limiting queries to specific categories. Taxonomy models can also be used to enable content notification applications. For example, specific individuals may opt to be notified any time a new document is introduced into a specific category.

Verity has also introduced a capability they refer to as Relational Taxonomies to enable multiple taxonomic structures to be brought into play, under user control, to create dynamic filters across complex collections of documents. This approach supports intelligent navigation and discovery operations either in conjunction with search or as a complementary function to the search tool.

Products

Verity Intelligent Classifier is the module that supports the K2 Enterprise content organization capabilities. Intelligent Classifier is integrated in Verity K2 Enterprise and Verity K2 Catalog, and operates in conjunction with the Verity Profiling Engine for content routing and notification.

Using Verity K2 Enterprise, businesses can combine what Verity calls the ABCs of content organization:

A. Automatic Classification – Positive and negative exemplary documents can be used to automatically generate the rules that define categories. This employs Verity's Logistic Regression Classification (LRC) technology.

B. Business rules – Domain experts can manually create new rules, or modify imported rules and/or automatically created rules to enhance accuracy or meet specific business goals. The ability to easily modify and fine-tune the rules that make up the model of each category is essential to classifying mission critical information.

C. Concept extraction – Verity's Thematic Mapping can be used to analyze the entire corpus of documents to reveal themes and concepts. This can be used to generate entire taxonomies, break populated categories down into subcategories or to mine enterprise knowledge for new insights. Once Thematic Mapping has identified concepts, it labels them with names that are easy for people to make sense of.

Verity Taxonomies

Verity Taxonomies are pre-configured taxonomic structures that are available for license and designed to speed and facilitate the process of deploying content organization strategies. Built by experienced knowledge engineers using practices learned over hundreds of consulting engagements with clients, Verity Taxonomies let you rapidly deploy industry-standard taxonomies that can be combined with your corporate taxonomies or easily customized to meet company- and industry-specific requirements.

Each Verity Taxonomy is based on industry standards, and built using the same business rules that provide the highest level of accuracy available in our Verity K2 content organization solutions. All taxonomies are updated on a regular basis to ensure the categories and the business rules that define them are up-to-date.

Available Verity Taxonomies include:

- Pharmaceutical Taxonomy
- Defense Taxonomy
- Homeland Security Taxonomy
- Enterprise Taxonomies: Human Resources, Information Technology, and Sales & Marketing

The Verity Pharmaceutical Taxonomy is the National Library of Medicine's MeSH (Medical Subject Headings) taxonomy converted to a Verity taxonomy file and Verity business rules. This comprehensive taxonomy is the industry standard with over 300,000 terms and topics.

The Verity Defense Taxonomy is based on the Defense Technical Information Center (DTIC) thesaurus published by the U.S. Department of Defense (DoD). This provides a basic multidisciplinary vocabulary that includes close to 12,000 topics.

Verity's knowledge engineers have built an extensive Verity Taxonomy on Homeland Security issues. Based on direct experience solving content organization issues for intelligence agencies in the U.S. and around the globe, this comprehensive taxonomy includes over 700 categories.

Verity's Enterprise Taxonomies are based on standard business terminology and processes that apply across industries, for functions including: human resources, information technology, and sales & marketing.

LexisNexis Content Organizer

Verity and leading professional information service LexisNexis have partnered to make available LexisNexis taxonomic definitions directly to Verity customers. Available only from Verity, the LexisNexis Content Organizer lets Verity K2 Enterprise customers use LexisNexis taxonomies and concept definitions that have been built, tested and fine-tuned by LexisNexis classification professionals over 25 years. LexisNexis concept definitions cover standard topics of major industries, business actions, financial news, legal and regulatory activity, scientific discovery, health and disease, sports, weather, society and cultural issues.

Factiva/Verity Alliance

Verity and Factiva, a leading professional information service and provider of enterprise information portals, have partnered to make available Factiva taxonomic definitions directly to Verity customers. The Factiva alliance allows Verity K2 Enterprise customers to use Factiva taxonomies and concept definitions that have been built, tested and fine-tuned by Factiva classification professionals over 25 years. Factiva concept definitions cover standard topics of major industries, business actions, financial news, legal and regulatory activity, scientific discovery, health and disease, sports, weather, society and cultural issues.

Primary Verticals

Verity's customers cover a broad range of industries and applications, including a large community of OEM relationships with other software ISVs. Public sector organizations, financial services, pharmaceutical/life sciences, professional services firms, and publishing companies are among the larger Verity customer groups.

Use Case

Verity software is deployed in a large variety of applications. These include applications that support the information work of a range of business functions, including, for example: pharmaceutical and scientific research; government information management and intelligence analysis; commercial Web site catalog and navigation support; employee and customer service and self-service; HIPAA, GPEA, and Sarbanes-Oxley regulatory compliance implementations; knowledge management systems and repositories, professional information publishing; and others.

A profile of one of Verity's professional services industry customer applications follows.

KPMG International Application

KPMG International is a leading global network of professional services providers. Over 100,000 KPMG professionals in member firms collaborate across industry, service, and national boundaries to deliver audit, tax, and advisory services.

With over 24 offices and nearly 10,000 partners and staff, KPMG LLP (UK), the United Kingdom member firm, is implementing Verity K2 to raise the bar for information intelligence on the employee intranet.

The Business Requirement

Like many firms, KPMG (UK) faced a pressing need to provide the firm's professionals with a unified point of access to information across a wide variety of sources, including Lotus Notes databases, Microsoft Exchange servers, Web servers, file sytems, custom enterprise applications, content management repositories, and packaged ERP systems.

In addition, the popularity of the intranet as a publishing, content sharing, and collaboration medium was exploding at KPMG, with content totals doubling and no agreed upon discipline or oranized process to manage the new materials.

Because the ready availability of much of this information is key to the effectiveness and delivered value of KPMG's knowledge professionals, the firm determined to take on a program to make its accumulating warehouses of intellectual capital easily accessible to partners and pofessionals. The challenge included not only ease of access to documents across the disparate sources, but also the desire to be able to connect professionals with questions to the firm's real experts in relevant subject areas.

Verity Implementation

KPMG began to deploy Verity K2 Enterprise in 2002. The initial task was to use the extensive set of repository and system gateways built into K2 Enterprise to expose the content from the disparate mix of KPMG content to make it accessible from a single interface. The application also added news and external research sources to the internal information sources, eliminating the need for professionals to replicate searches across different environments. The application succeeded in eliminating unproductive "thrash" which had previously resulted from professionals using multiple search engines associated with individual sources or formats (Web vs. business document content vs. collaborative material and e-mail, for example).

KPMG also elected to implement the Verity classification technology suite, focusing initially on the capability to capture domain expertise from the various practice areas to drive business rules which deliver customized contextual accuracy in search results. This dramatically improved the quality of professionals' search experience while leveraging the organization's intellectual capital in an innovative and persistent framework.

The deployment at KPMG utilized the Verity technology in a diagnostic role to identify and evaluate the balooning quantities of intranet content. By using Verity to produce a catalog of sources, documents, subjects, and usage data across the various systems in the enterprise, KPMG could identify for the first time what content was redundant, what content was actually being used, and what content was out of date. This process allowed KPMG to reduce the size of its intranet by over 30% while raising the quality level of the remaining information.

Business Benefits

KPMG has realized business benefits from this implementation that fall into three areas, one "hard dollar" and two "soft dollar" areas. The hard dollar benefits are related to the reduction of costs of hardware (particularly storage and processors) and software licenses related the the reduction in the physical size of the intranet.

The first soft dollar benefit area is related to giving back time to KPMG's professional work force. This is accomplished by centralizing the search operations and eliminating the time-consuming thrash of manually manipulating multiple search engines across different systems. It is also accomplished by raising the quality of the underlying information and the accuracy of the browsing and searching environment.

The second soft dollar benefit area is difficult to quantify but extremely tangible to knowledge professional like those at KPMG. This is the area related to the level of confidence a professional has that the research or investigation she has undertaken has delivered all the important material related to the question at hand. When facing a client, having all the information about a question and having intelligent things to suggest is the baseline level of performance for a KPMG professional. Failing to have the complete picture, or being "surprised" by the client knowing or discovering information the professional had not taken into account in her analysis, amounts to malfeasance and very likely to an early exit for

KPMG from the account.

The strategic value of implementations like Verity K2 for knowledge-based businesses is that they can help reduce the risk of putting poorly informed professionals in the field, thereby improving performance for the client and maintaining the value of the KPMG brand.

End Note

A new set of practices to support the on-demand enterprise is in the process of being born. A critical infrastructure element of the information intelligence platform, the emerging classification and taxonomy practice will help the enterprise deliver dramatically simplified and impactful interactions with information in existing systems; better interoperability among previously incompatible systems; much more intelligent access and display of all kinds of content; and vastly improved flexibility in information-based processes.

These developments will translate into significant benefits to firms who adopt these practices, expressed through the organizations' fresh ability to leverage existing investments in information assets. Companies will be able to move beyond the boundaries and limitations of previous legacy application and search systems to mobilize composite, intelligent applications that deliver appropriate and tailored information across a range of online interaction opportunities.

The areas of investment return from information intelligence facilities will cross the range of the organization's practices and lines of business, but they will share a new ability to bring the organization's customer intelligence, process knowledge, and collaborative facility to bear on the process of creating value for internal and external customers.

About Delphi Group

Delphi Group, a Perot Systems company, has been providing technology management services to the Global 2000 since 1989.

Through focused market research, consulting and value-driven educational programs, Delphi Group consultants illuminate the emerging roles of technology in the digital enterprise.

To learn more, contact Delphi Client Services at 800-575-3367. Or email: client.services@delphigroup.com.

Publication and distribution of this report is partially underwritten by the technology suppliers who have engaged with Delphi Group in the Spring, 2004 Taxonomy Research & Education Program. This program is part of an ongoing series of Delphi Group studies, publications, educational programs, and conferences which seek to establish thought leadership perspectives and best practices in the area of information intelligence. Profiles of these suppliers' technologies are included in this report.

DELPHI
GROUP
111 Huntington Avenue, Suite 2750
Boston, MA 02199
V: (617) 247.1511

www.delphigroup.com