Taxonomy & Content Classification

Market Milestone Report

Featuring a Delphi Group Assessment of: LingoMotors

Do You Know What You Know?

LingoMotors
585 Massachusetts Avenue,
3rd Floor
Cambridge, MA 02139
(617) 492 7377
(617) 492 7399 fax
hello@lingomotors.com
www.lingomotors.com

Introduction

Cataloguing unstructured information is a chronic problem, that if not adequately addressed can be terminal for your organization. Today we have many easy-to-use and accessible tools to create and publish information electronically. Examples are: the ubiquitous suite of Microsoft office products like Word and PowerPoint, Adobe Portable Document Files (PDF), Web Pages (HTML files), e-mail, news feeds, and the like.

Lack of information is no longer the problem—but lack of time to correlate, categorize, analyze and act on the information is a crucial problem. The information is there, hidden in reports and e-mails and published on the corporate Web site. We are placed in the position of being unable to find applicable and pertinent information to make timely business decisions. This comes at a time when the agility to quickly make fast, informed decisions is increasingly critical to survival and prosperity. As the volume of opportunities increases, the duration of the time to act on each opportunity decreases. The information-based economy is in danger of drowning in a sea of irrelevant, unstructured data.

A new segment of software has emerged to help with the task of combating "infoglut." For example, there is software that enhances the performance of search engines, text mining, natural language search

applications, ontology, summarization and taxonomy. Taxonomy software correlates and groups unstructured information from a myriad of sources. Taxonomy software also helps us in the automation of this process. The software's fundamental challenge is to understand the concepts and ideas that group like documents together and separate unlike documents. Think of taxonomies as computer-generated card catalogs that allow us to locate, retrieve and cross-reference information in our digital libraries.

Taxonomy software can reduce our reaction time to make informed and timely business decisions based on knowledge and information contained within the unstructured data of an organization's digital documents. This software helps us form ideas from information we didn't know we had while revealing relationships and correlations that were submerged or lost in the depths of the ocean of information overload. This happens on an individual basis and a community basis. We work in groups and we must be able to communicate with the constituents of the group so we can have actionable knowledge. Productivity comes from seeing connections, evaluating importance, recognizing context, understanding the implications and understanding the correlations of data and information.

Although taxonomy software cannot stem the tide of "infoglut," it can help us find the information we need to survive and prosper in the new knowledge-based economy – to truly know what we know.

Irrelevant Information and Infoglut

An ancient Greek philosopher and ascetic, Diogenes searched for an honest man and never fulfilled his quest. You may feel like Diogenes when trying to search for and retrieve information from your enterprise portal or the Web to make a business decision.

Lets use the example of searching for information about chips. A search on the Web for "chips" (using the Google search engine) returned 2,430,000 references. Even if only 1% of these documents were relevant, that is over 24,000 documents, which is way beyond the capability of most of us to wade through. Some of the documents contained information about – chocolate chips, potato chips, buffalo chips, wood chips, poker chips, an old TV series (ChiPS), or integrated circuits. "Chips" is one of those words that have multiple ambiguous meanings like "java," "can," "branches" and "boot" (noun or a verb).

People can distinguish concepts from each other based on context and the specific meaning of a word as it should be applied to the situation. Computers cannot. Other methods must be used by computers to give us the results we want. But before we explore these methods, let's look at some examples of how people think versus how computers "think."

Context

The definition of context is that which surrounds, and gives meaning to, something else. People explore concepts; computers primarily search for key words. Relevancy is entirely 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. 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 people with the proper context can know that Reverend Billy Graham will perform the ceremony and not be the recipient of Bruce's ardor.

"Customers" may be a descriptor for an employee in an HR document about medical benefits, but "customers" may mean something different in the context of the sales department. If a given document mentions "customers" where should it be categorized?

In the above example about "chips," you might have been looking for 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, but 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 the English 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 "buy the farm" (drop dead) or did Joe violently place his foot in contact with a container of water. The context of the surrounding phrases in the rest of the document will clarify this. So the question becomes, how do you categorize the documents that has words like "chips" or "java" or phrases like "kick the bucket."

Browsing vs. Searching

To search effectively you must know the terms you want to use *before* you see what is in the collection of documents. Key word search however 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 25% of the day is involved in searching for information on the knowledge workers computer system. About 70% of that time is spent browsing for information[1]. 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" It is much 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.

Report Scope

The objectives of this report are to:

- Inform about the current state of the market for taxonomy software
- Explain with examples how these technologies are applied to real life business issues
- Report the findings of a market survey of over 450 organizations about taxonomy
- Understand how those organizations are planning to implement taxonomy technology
- Analyze trends of taxonomy software technology, development and implementation
- Describe the technological landscape of the taxonomy software market,
- Compare the technologies' similarities and differences

Taxonomies

Why Now?

The first question we have to ask is, "Why the current high interest about taxonomy?"

Delphi Group's research on user experiences with corporate Webs reveals that lack of organization of information is in fact the number one problem in the opinion of business professionals. These professionals may include the customer service representatives, the sales team, the financial services professionals, the R&D engineers and the health care professionals that work in our

[2] Delphi Group Research

organizations. If these professionals are spending 25% of their time or more looking for information, then this results in an opportunity cost and represents a runaway expense item in many organizations. Our ability to create information has substantially outpaced our ability to retrieve relevant information. Delphi calls this information explosion "digital sprawl." The impact of this situation has dramatically affected the way we work.

We now have many tools to aid in the creation of electronic information. It is easy to create an electronic report or presentation and convert those to PDF files for easy viewing and sharing. E-mails are the ubiquitous mechanism for communication within business. An estimated eight billion e-mail messages will flood U.S. corporations each day in 2002 say leading industry sources. Proposals, resumes and contracts represent other sources of unstructured documents. Portals, Web sites, visible and invisible[2], and intranets have also made it easy to place and share information. Web pages are increasing at the rate of 7 million pages per day.[3]

Since it is easy to create, share and store information, the rate of unstructured information is now growing exponentially and turning into what one taxonomy vendor calls knowledge asset "landfills." There are 250 megabytes of information for every man, woman, and child on earth.[4] The percentage of unstructured data to the total amount of data is estimated at 85% and is growing. 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 fix the problem it has helped us create.

Evolution of Taxonomy Technologies

People are natural categorizers. We tend to group similar documents into categories by conceptual subject matter. We have developed filing systems for our paper documents. Examples of these are found in every office. We have filing cabinet filled with paper documents that are grouped together by some system:

- Categorized by type e.g. contracts, marketing collateral, or invoices
- Alphabetized by company or individual's last name
- Arranged by date
- Grouped by department e.g. legal, HR, sales, accounting

We do this so that when we can find the relevant information to complete the actions or make the decisions – in short to run our business. We call this "actionable information."

We have structured databases to keep track of our transactions and customer lists. This data is in rows and columns that tell us who to contact, how much to pay, how many we sold, etc. But this structured information doesn't tell us what will likely happen or why. The "why" is usually buried in our "corporate landfill" of associated e-mails, reports, or presentations.

The first step was search and retrieval software. Unstructured documents were "scanned" and indexed and key words along with their frequency of occurrence were placed into databases. Search engines were developed so that we could scan the index and find the documents and create pointers or hyperlinks to the documents that contained the key words that we thought were relevant. The simple search engines and indexers soon developed into complex ones that could support Boolean logic such as "and,"

[2] Invisible refers to dynamically generated Web like Amazon.com catalog pages, Wall Street Journal Archives. – Invisible Web, Chris Sherman

[3] UC Berkeley

[4] UC Berkeley SIMS How much information.

"or," "not" type of qualifiers. These functions were expanded to include "wild card" searches and "fuzzy" searches like "near." When sources were relatively limited these tools helped you find the information you wanted.

Now, however, instead of returning too small a list of relevant documents, the search engines return too many choices, as in our previous example of searching for "chips" with 2,400,000 possible documents. These are brain-dead lists that insult our intelligence. When Boolean logic searches turned out to be too complicated for most people to use, natural language query software helped develop key word searches that reflected our vernacular and intentions when searching for information. The same problem soon developed – too much recall, and too many references, even when ranked by relevancy percentages.

Another mechanism soon evolved, that of metadata, or "data about data." With this approach, when a document is created the author is prompted to list the creation date, title, subject matter, synopsis and a few keywords. Since authors are very close to the document as the creator, the relevancy and the consistency of this information were questionable.

The next step in retrieval technology was a methodology called link ranking. The more the Web page in question is linked to other systems and searched by other systems, the more probable it is important to more people. Each time a link is clicked, the click is recorded in the database and the client is redirected to the proper site, creating a simple click thru monitor that records the popularity of links. The flaw here is that this becomes a popularity poll or self-fulfilling mechanism; the more your site is visited the more likely it is to be visited.

The limitations of these various approaches were recognized by a number of "information and language" experts and a new generation of technology has emerged. Taxonomy software was originally developed to help speed the process of creating hierarchical structures.

Is "Taxonomy" a Misnomer?

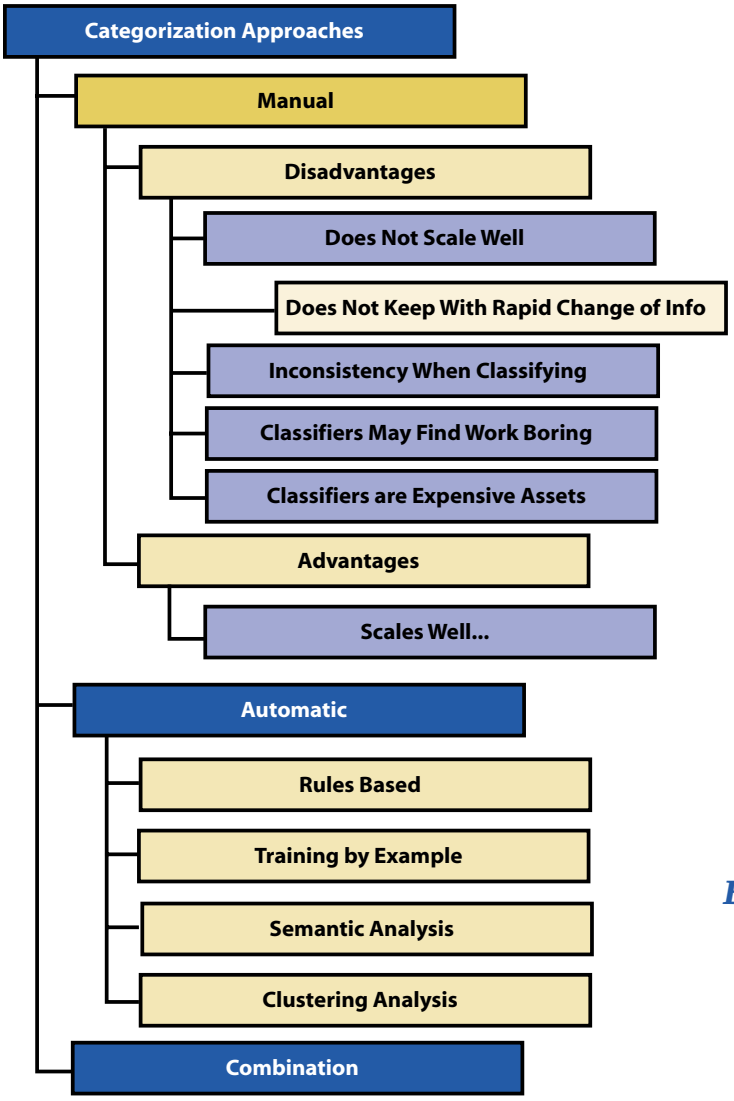
The word taxonomy evolved from the life sciences where a plant or animal is placed in a *single* spot describing its hierarchical relationship to other plants and animals. One of the interesting ironies that developed during the research phase of the project was the realization that the description of this category of technology was somewhat misleading. 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 digital information, taxonomy is a systematic classification of a conceptual space. When categorizing a document, it could and might be placed in multiple categories depending on the context.

For example at a pharmaceutical company the research scientists may place a particular document about a new drug in a category based on its chemical composition. The marketing and sales people may want to place that same document in a category called "competition." The same document can be classified into two or more categories by different groups or under different contexts or conditions. (You can see where the analogy breaks down: scientific taxonomies are fundamentally rigid, while categorization is personal and subjective and sometimes even arbitrary).

Using Categories in the Search Process

People search for information in two different ways. The first is the process you use 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.

Many times people don't know the answer they are looking for when they begin the search for information. In fact, as discussed above, they often don't know what they don't know yet. Let's use the example of doing research on categorization technologies. Here is an example of a possible hierarchy of categorization information. Each of these categories can also be called a node.



As you start the search process for categorization software you soon realize there is one alternative, the purely manual approach to categorization. There are several reasons, however, why an exclusively manual system may not be as effective: the chief drawbacks are that human classifiers are expensive, inconsistent, and obviously not scalable to the same degree as automated approaches. Although these considerations were probably not foremost in your mind when you started your search, they certainly can support your argument in favor of the purchase of a software system.

The methodologies used in the automatic process of categorization are likewise not something you would know about at the beginning of your search. 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 we are making here is that when you started your quest, you were not thinking about cost-justifying a software solution, nor were you aware of the various methodological approaches to the problem. This information was discovered either during or after your search—maybe. Wouldn't it be better to have such information at your disposal at the beginning of your project?

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

Browsing Process

As you browse for relevant information about your particular subject, you will find three more characteristics that describe the process: browsing is dynamic, interactive and iterative.

Browsing is dynamic. Information changes all the time. In today's world, virtually any search on a complex topic becomes a hunt for a moving target. For example, my search for "chips" today now yields 2,490,000 hits—60,000 more instances of information than just a few days previously. Versions will change, articles will be removed, information added, etc. Delphi's research shows that at least 10% of enterprise information changes on a monthly basis. This is a conservative finding—Delphi Group believes that the dynamic, volatile nature of information sources is the number one reason that knowledge workers have difficulty finding the information they are looking for.

Browsing is an interactive process. As you navigate a well-designed interface to information, you will automatically be directed to other relevant topics. If you search and browse through information about categorization software, for example, you will find reviews, analysis, white papers and commentaries with information about other technologies, companies or related topics of information that may be worth investigating.

Browsing is an iterative process. Repeating the process refines your focus while broadening your knowledge. Accessing relevant information and interrelated ideas and concepts supports a fundamental change in your activity—from simply searching, to finding and discovering.

Taxonomy and Search

One purpose of taxonomy is to aid in the retrieval of relevant information. 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 filters are employed. For example, if we had had a general category like "computers" in our search for "chips," we would not have wasted any time with false returns from the "recipes" category.

Benefits of Taxonomy

Finding relevant information quicker is the key benefit, especially when it provides immediate access to the right information that allows the user to take effective actions. Equipping enterprise knowledge workers with the tools to make faster and better-informed 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.

To paraphrase a quote, "to search via a computer without a taxonomy system is like trying to find your way around an unfamiliar country without a map." Taxonomy helps delineate the conceptual relationships that exist within and between various topics contained in the multitude of unstructured data within various enterprise documents.

The benefits are:

- Discovering information you didn't know you had
- Avoiding duplicate efforts within large organizations where independent groups "reinvent the wheel" over and over again
- Not repeating the same mistake
- Reports are better prepared if the author really expects to be read.[5]
- Provide overview as well as details about a subject
- Demonstrate relationships
- Reduce complexity

Taxonomy Software Integration With Other Applications

Taxonomy can impact many aspects of your organization. As organizations implement various software solutions to manage their knowledge assets, taxonomy can dramatically increase the effectiveness of such solutions. All the software features in the world won't matter if they don't facilitate "just-in-time" knowledge retrieval. Software applications such as portals, content management systems, knowledge management systems, search and retrieval software, personalization software, data extraction, and data mining can all benefit from taxonomy. Many taxonomy solutions are sold with Application Programming Interfaces to integrate into these existing applications.

[5] There is a product marketing manager that places a peach cobbler recipe in his market research reports. He places the recipe in a relatively arcane section of the report. At the end of the recipe, he offers to whomever finds the recipe a bottle of Don Perignon champagne. In his career, producing many reports, he never has had to buy the champagne.

Taxonomy Is Not Just For Enhancing Search

Let's apply taxonomy to other applications. Taxonomy technology is currently making search engines more efficient by, as we've seen above, limiting a search to a subset of relevant topics. Browsing topics categorized by hierarchical relationships will allow new insights and correlations when searching for information. Imagine the impact of applying this technology to other applications that must access and manipulate masses of unstructured data.

One taxonomy vendor's customer is a professional services organization. This organization is constantly developing proposals for its customers. Since the organization is national in scope, at any given time there are many offices developing proposals for many customers—often in the same vertical industry. The professional services organization has an existing-proposal generating application. They are now using a taxonomy classification tool to discover:

- Proposals covering similar industries that have been successfully presented to similar customers
- Sections of proposals that can be reused
- Research data from previous proposals that are pertinent to the new customer
- The correlation between different proposal patterns or approaches and the success or failure of the proposal
- Up-to-date info from Internet news feeds of particular relevance to prospective customers

In another example, imagine categorizing unstructured data for the benefit of your CRM application. You might readily identify patterns and correlations that bridge seemingly disparate occurrences that are in fact tied together by some previously unknown common factor

Finally, consider the case of a data mining application that discovers, for example, that sales of rain gear declines in the northeast while increasing in Arizona. Data mining is notoriously good at telling you the "what" but not the "why." By browsing a taxonomy-enhanced knowledge base, however, you may identify related e-mails or reports that point to the existence of an annual summer "monsoon" season in Arizona corresponding with an unusually dry summer in the Northeast. Now you have a "why" to substantiate a business decision to increase rain-related inventory in a desert region.

Determining Your Categorization Requirements

These questions will help you quantify how a taxonomy system can be applied to your organization. The primary strategic consideration for any firm is what kind of resources to commit to the process and in what proportions.

Suggested Questions

- Can employees, partners and customers find the information they want?
- How are you currently categorizing your information?
- What is the diversity of the documents?
- How many documents will you categorize?
- Are there isolated areas that have a minimal number of documents? (e.g. small intranets within the organization that don't have enough documents to develop categories for them)
- Will this number grow over time or remain constant?
- What rate of growth do you expect—linear or exponential?
- What file types and formats are the information stored in? (e.g., text, PDF, PowerPoint, HTML, etc)
- How volatile and dynamic is the information? Does it change hourly, daily, monthly?
- When does it change?
- What is your official document publishing process and policy?
- What are the life cycle parameters?
- Who categorizes information now?
- Who sets up the categories?
- How available are they for this task?
- How do categories get updated and expanded?

Market Survey: Results and Analysis

During the first week of February 2002, Delphi conducted an extensive survey of approximately 450 end user organizations on the subject of categorization and taxonomy management. Several dozen questions were asked in regards to the business issues surrounding the evaluation, planning and implementation of taxonomy technology.

The objectives of the survey include:

- Validate the extent of the unstructured data problem faced by knowledge workers in today's organization
- Determine the relative importance of the business issues surrounding the retrieval of information from unstructured data sources
- Understand the scope of the problem and the perceived impediments associated with job performance and unstructured data.
- Confirm the characteristics of unstructured information sources in terms of size, volatility, and language
- Verify if there are classification processes and policies in place today
- Ascertain if there are taxonomy software projects underway or pending, their relative importance, and the proposed budgets for implementation and maintenance of taxonomy
- Find out who will be responsible for defining and then maintaining the taxonomy software
- Clarify how the taxonomy software should be configured and deployed
- Discover to what extent the market recognizes the leading providers of taxonomy software

Survey Summary

Enterprises know they have a serious and growing problem with unstructured data, and that the problem is dramatically impacting their ability to make rapid and effective business decisions. Current systems are not adequate. Organizations planning on developing a taxonomy strategy remain unsure about how to do so. This presents a significant opportunity for the technology providers seeking to fill this need.

Profile of Respondents

The survey of 450-plus respondents represents a fair sampling of enterprise organizations, with over half the organizations having revenues of over $100 million. 73% are located in North America. The respondents were either executives, IT, or LOB or had project management responsibilities. Respondents' role in determining taxonomy software was primarily as a sponsor or project lead, or involved in defining need or specifications.

Survey Methodology

Individuals identified by Delphi's analyst team were contracted directly and asked to answer a series of structured survey questions. The survey format was primarily multiple-choice, with either single or multiple answers possible depending on the question. Respondents were also provided the opportunity to volunteer more detailed and otherwise restricted answers (i.e,. "fill-in-the-blank") which are quoted at various points throughout this report. The results from the survey are tabulated and graphed in the sections to follow. Listed below are the profiles of the respondents, the questions, the results and, finally, Delphi's analysis of this market survey.

Survey Limitations and Risks

The survey resulted in over 450 respondents. While this number is sufficient to develop quantitative and qualitative trends, variations may be found within individual deployments or taxonomy initiatives. While respondents represents a valid cross-section of enterprise-class organizations, the population's pre-established interest in this or similar technology may distinguish this group as more knowledgeable and aware than a similar group of randomly selected enterprise respondents.

Survey Results

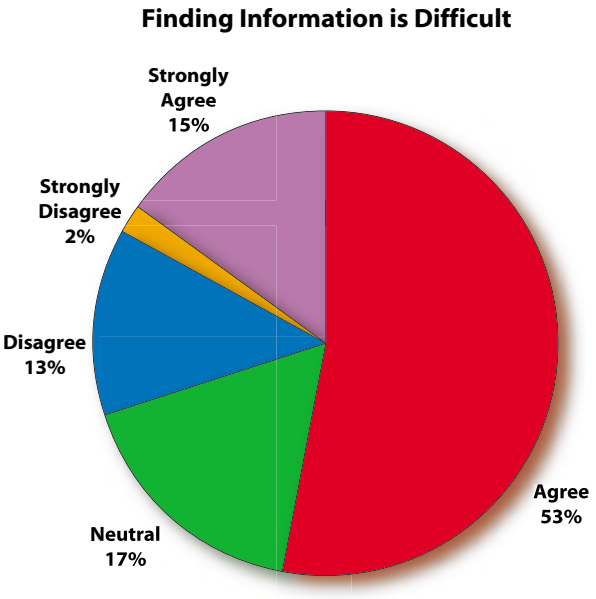
"Infoglut" and Knowledge Worker Efficiency

A number of questions were asked to validate the extent of infoglut faced by survey respondents and to verify the environmental causes.

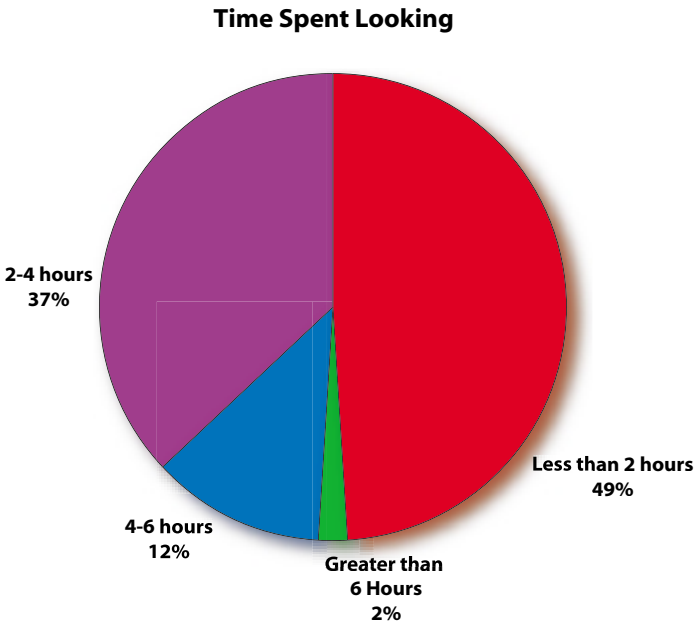
"Infoglut," the business problem of too much information, is real, growing and recognized as an important issue in today's enterprise. The exponential rise of unstructured data is having a major detrimental impact on the efficiency of enterprise organizations. Most respondents to this survey (with job descriptions such as executives, IT management, Line-Of-Business managers and project managers) spend more than two hours a day (25% or more of an 8-hour day) searching for information to perform their jobs. More than 60% agree that finding information was a difficult process, and much of the time they cannot find the information they need to do their jobs.

Question #1 – "Finding the information I need to do my job is difficult:

agree – disagree"



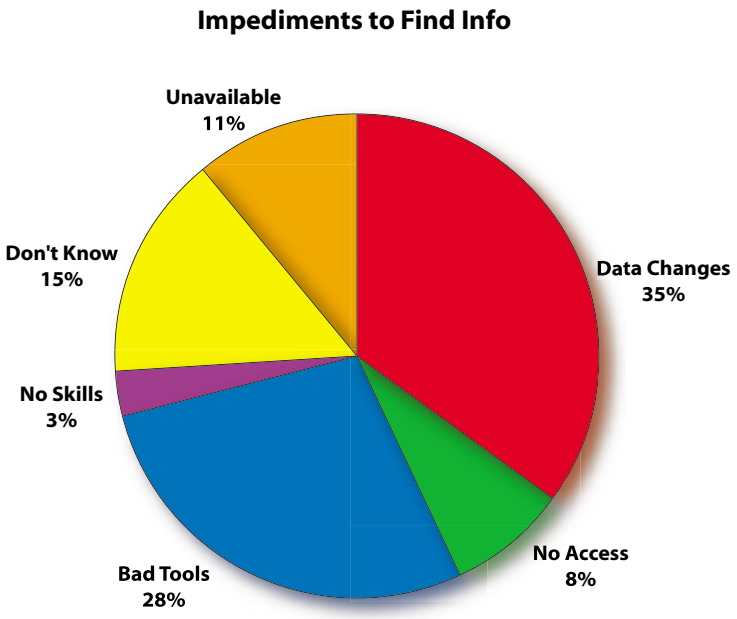
Question #2 – "How much time do you spend each day searching for information that is critical to your job performance?"



More than 60% of the respondents agreed or strongly agreed that finding information was a difficult process and over 50% of the respondents were spending 2 or more hours each day searching for information. This is consistent with Delphi direct client experience and various other surveys, which of shown knowledge workers typically spend 20% to 30% of their time searching for business-related information, the majority of which is stored electronically and should otherwise be easily identified.

The issue of "search time" is one of the fundamental symptoms of infoglut, and is at the heart of how most organizations measure the business impact of taxonomy management. While the positives impact information accessibly on productivity is at times overrated (i.e., simply because information is accessible is no guarantee that work is being done), there is no denying the negative impact resulting from the lack of information. Simply put, very few individual are able to generate business value through the act of searching for information – this frequently required task is what economists refer to as a "transaction cost" or otherwise a drain on productivity. And it is what the rest of us would call "a waste of time."

Question #3 – "The biggest impediment to finding the information I need to do my job is:"



As respondents looked and searched for information, 61% of the time they had a 75% chance or less of finding the information they needed. The two main impediments to finding the information they were seeking:

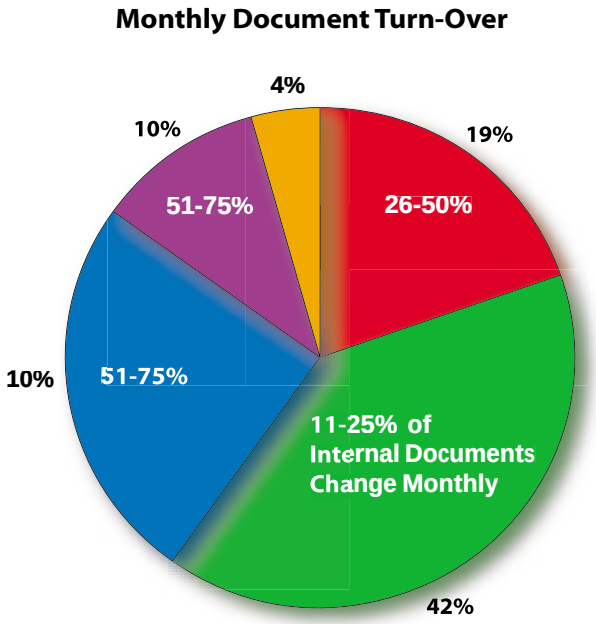
- bad tools 28% of the time
- volatility of the data 35% of the time.

The two most frequently cited reasons for this are "Bad Tools" and "Data Changes" or the concern that information is changing too fast. Even with the wide array of available search tools and content management solutions, knowledge workers either can't find the information they need to do their jobs, or they are spending an inordinate amount of time looking for that information.

Document Turn-Over and Data Change

Another area of focus for the survey was an examination of information volatility, consistent with the finding that "data change" was the most frequently sited impediment to finding information. Over 95% of the respondents reported that more than 10% of their documents change on monthly basis. In a separate series of questions, more than half (59%) stated that they have over 50,000 documents in the corpus of information contained in their environment. Compounding the problem, the rate of information volatility is certain to accelerate. This study shows that respondents have access to at least 50,000 documents within their immediate organization— with often 10,000 or more of those documents changing on at least a monthly basis.

Question #4 – "What percentage of unstructured data changes at least monthly?"



Current Software and Manual Systems Not Adequate

Enterprise organizations are looking to taxonomy software to help solve this problem. Most of the organizations surveyed had some type of system for classifying information within their organization, but most had no formal policy for tagging the documents and left the classifying up to the author. The survey's

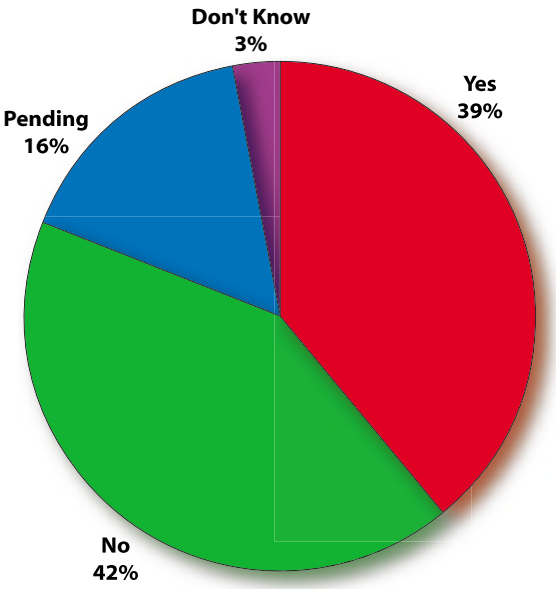
central message—that finding information is extraordinarily difficult and time-consuming – point to the fundamental inadequacy of current approaches.

It's hardly surprising, then, that even where there is some type of software in place to help with classifying unstructured data, most (over 55%) respondents felt it was necessary and important to put a taxonomy software system in place. Over 90% plan to have a taxonomy strategy in place within the next 24 months.

The objective of this set of questions is to understand if there are systems currently in place for categorizing documents and who within the organization is responsible.

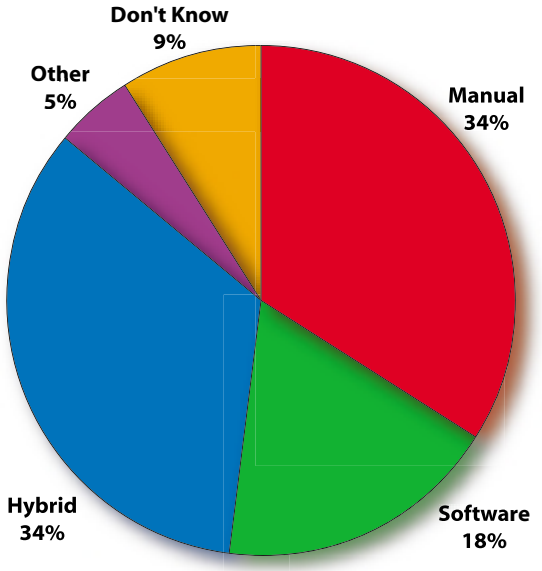
Question #5 – "Does your organization provide a system for classifying the information you work with?"

System for Classifying Documents



"Question #6 – What type of system for classification of documents?"

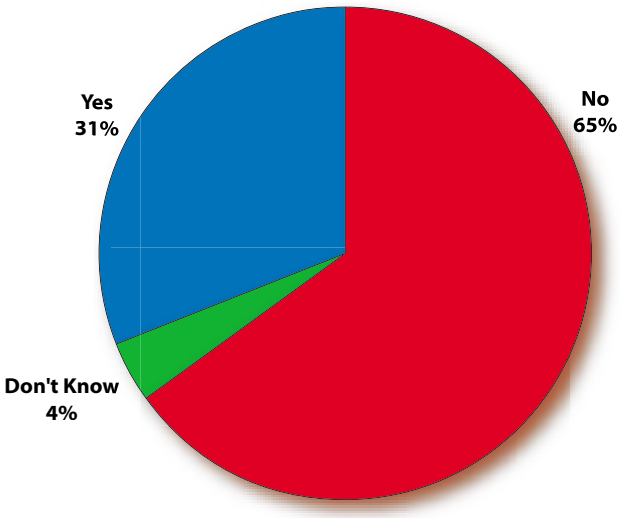
Type of System for Classifying



For a slight majority, 55%, there is some system for classifying documents used in the organization. Although not a true taxonomy system, over half were using some type of software or a combination of manual approaches and software for classifying documents.

In contrast, only a third of respondents had a system in place for tagging documents.

Tagging System for Documents

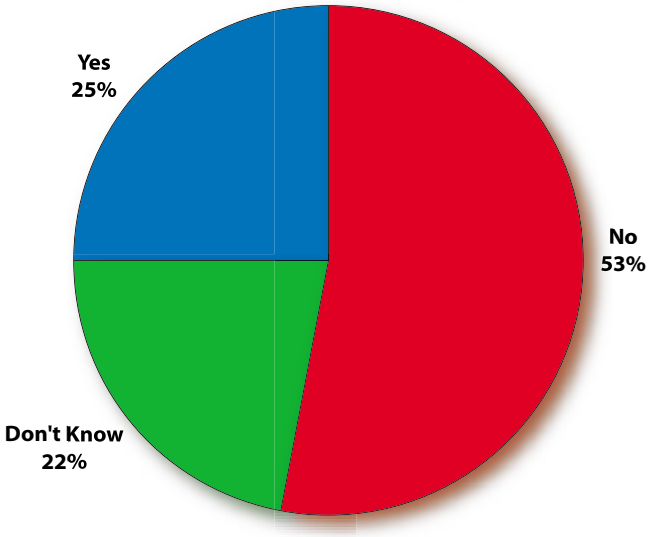


Enterprise Organizations Deomonstrate Interest; But Unclear on Timing and How To Implement Taxonomy

This next set of questions explores specific plans for implementing taxonomy software within the respondents' organizations.

Question #7 – "Is anyone in your organization working with taxonomy software today?"

Working With Taxonomy Software

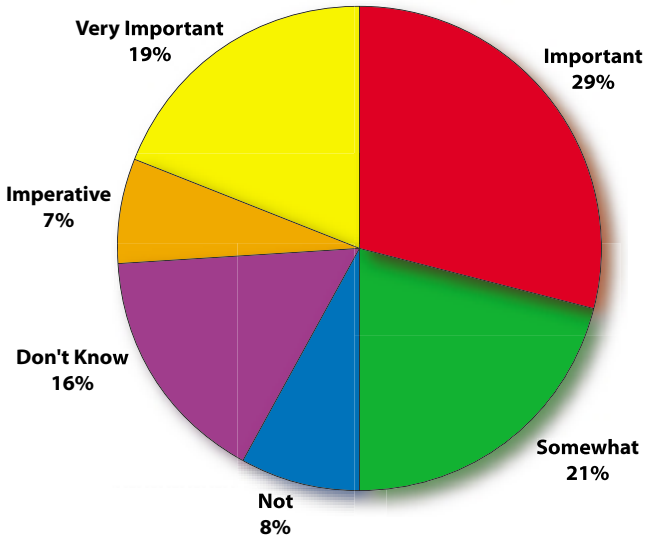


Over 75% of the respondents either don't have or don't know if they have taxonomy software being used or proposed within the organization. Yet 55%—a clear majority—feel it is important to have a taxonomy system.

The time frame for 90% of the respondents is to have a taxonomy strategy within the next 2 years or less, indicating the likelihood that many organization will follow-through on the priority placed on taxonomy, as indicated by Question #8. The delta between those who indicate someone working with taxonomy software and those who have an enterprise strategy in place (about 2:1 respect ively) indicate that the deployment of taxonomy management software is still very much in the experimental or pilot stages, a notion consistent with other finds throughout the study.

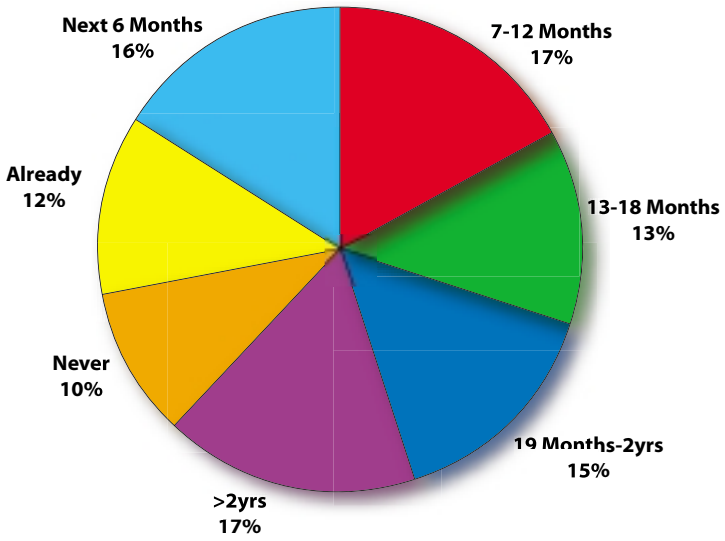
Question #8 – "How important is taxonomy software to your organization's business strategy?"

Taxonomy Important to Business Strategy



Question #9 – "When will your organization develop a taxonomy management strategy?"

Time Frame for Taxonomy Strategy



There are a number of taxonomy-related issues that the enterprise is unclear about. This is shown by both the diversity of answers and the preponderance of "I don't know." responses. The data from this survey shows that organizations are particularly uncertain about:

- Who will be responsible for determining enterprise strategy for choosing, implementing and maintaining taxonomy software.
- The costs of acquiring or maintaining the taxonomy software. (This is especially troublesome given that most respondents have a role in determining needs and/or specifications for the software.)
- The topology of the vendor landscape. No vendor is recognized as a leader in this technology segment. The vast majority of respondents could not answer that question nor could they identify *any* company in this space in any consistent way. This obviously translates to a significant opportunity for one or more vendor organizations to take on this leadership challenge.

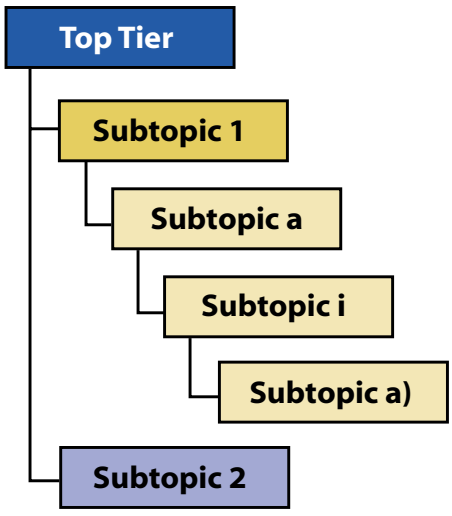
Definition of Terms Used in Discussions on Taxonomy

Recall, Relevancy and Precision

Recall is inversely proportional to precision. The more precise you are about defining what you want to look for, the fewer documents are recalled. Using our "chips" example, 2,430,000 documents found represents "recall," answering the basic question of "How much is out there?" The subset of 24,000 documents indicates relevancy. The 10 documents found regarding the newest Intel processor reflect precision. The ultimate measure of a taxonomy's success or failure, precision addresses whether or not you found the answer to your question.

Hierarchies

As discussed before, building hierarchical relationships is something that humans do inherently. Every piece of information we gather gets placed in a virtual cubbyhole in our brain. That information is not useful until we can relate it to something else. Since many of us are visually oriented, the process of building graphic representations of hierarchical relationships helps us build those relationships faster, allowing us to more readily envision how items are related to each other. Hierarchies also allow us to see both the big picture (with just the main topics), and the details and relationships contained in each "nested" subtopic. The number of nested levels varies greatly according to subject matter, user preference, and the built-in parameters of a given system. The question of how many



levels of information are either necessary or pragmatic is open to debate among systems designers.

Bottom-Up or Top-Down

Methods that are used to build categories vary. One approach is to take the details and start placing them in "buckets." This approach is termed "Bottom-Up." Returning to our "chips" example, a Bottom-Up approach would start by separating a subset of documents into either the integrated circuit category or the food category. If you run into too many documents falling into different *kinds* of food categories, you may, for example, need to divide that group into both potato chips and chocolate chips. The Bottom-Up approach thus expands categories iteratively as part of the classification process.

The inverse process is called "Top-Down," and is how people traditionally categorize. In this approach, domain experts analyze the subject matter at hand, and determine that the documents will fall into, for example, 10 general topics. Each document is then examined and placed them in its appropriate, pre-existing category.

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. Each document can be considered a "bag of words."; clustering essentially groups the similar words contained in each bag.

Pattern Matching

Pattern matching is the process of looking for groups of words that are often grouped together. One example is "Business Unit Manager" as a title. (Note: this is also a "noun phrase" that could be processed using semantic analysis.) Other recognized patterns include frequency of words used in a document, placement of words, proximity of words to each other, and clusters of related words. Pattern matching is inherently language-independent.

Controlled Vocabulary and Thesauri

A controlled vocabulary is a finite set of terms. The typical application for a controlled vocabulary is for lists of allowed values in structured record systems. A thesaurus is a particular type of controlled vocabulary that represents a formalized description containing a finite set of terms and relations between terms, and frequently also contains information on how the terms are to be applied. The following topics are included in a thesaurus:

- Preferred terms
- Non-preferred terms
- Semantic relations between terms
- How to apply terms

Thesauri are usually in vertical industries such as legal, medicine, and pharmaceutical. An example might be a music thesaurus: Soprano: Broader terms = vocalist, singer; Narrower terms = lyric soprano, coloratura soprano; Related terms = mezzo-soprano, treble.

Example-Based

Another approach to classifying unstructured data is to develop a subset of documents that pre-establish categories defined by a set of reference content. These "training sets" can be automatic or supervised. The software analyzes new documents in comparison to the training set and searches for similar concepts and ideas. This approach is also referred to as "machine learning."

A limitation of the example-based taxonomy method is that the resulting classification is totally dependent on the breadth and precision of the training set.

Neural Networks

Artificial Intelligence as applied to a computer system is modeled after the neurons (nerve cells) in a biological nervous system. A neural network is designed as 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 human-like 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."

Spiders

Spiders are automated processes used to feed pages to data extraction and parsing engines. It's called a spider because it "crawls" over the data. Another term for these programs is *crawler*.

Taxonomy Market Landscape

Segmenting the Market

The market for classification software is exceptionally dynamic. It is evolving and changing on a daily basis. An overview analysis like this one, therefore, can at best hope to take a snapshot in time. There are many companies supplying this software. The companies discussed in this section are market leaders and are fairly representative of the myriad of approaches and technologies offered within the market. The organizations featured in this section are listed alphabetically, and chose to participate in this project with Delphi.

Methodology Algorithms

This section will deal with the technologies underneath the technologies. As you will see there are many approaches to tackling the problem of building automatic or semiautomatic taxonomies. We will examine each of the technologies from an overview perspective and talk about them in more detail as we examine the individual products that utilize them. These technologies are based on various algorithms using statistical analysis, semantics and neural networks.

There are a number of algorithms that technology vendors customize, optimize, combine and patent in order to categorize digital documents. For a variety of reasons each vendor has chosen a particular algorithm method or combination of methods. This is a list of each of the methods that are discussed in more detail in the following sections:

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

To help understand these methodologies, consider the example of the various types of engines used by car manufacturers today. There are 4-, 6-, 8-, or 12- cylinder internal combustion gasoline engines. There are also diesel, electric, hydrogen or natural-gas powered engines. Increasingly, a car's power plant might be a hybrid of any of these. Some of these engines are more suitable for the race track than a commute to work, but all will get you to the grocery store and back. These engines have many pros and cons associated with them, and their use depends on a design perspective and what performance characteristics the engineers want to provide to the users.

Similarly, there are many approaches to building and populating a taxonomy. No one method or even a combination of methods seems to yield particularly superior results. Many of the companies profiled below are very technology-centric, and spend much of their marketing effort trying to convince us of the advantages of their approach or methodology. The bottom line is to understand how these differences affect system performance in the only environment that matters—your unique data environment.

Keyword

Let's start with a brief analysis of search and retrieval as we know it today (that is, without an accompanying classification solution). We will then have a basis for understanding how other technologies try to improve on its weaknesses.

Search engines look for keywords in the title, synopsis or abstract, body of a document, or the meta-tag (i.e., "data about data") section. A program "crawls" through the subject document, and each instance of a keyword is put into an indexing database that describes how many times it found that particular word and where that word was located within the paragraph, page or document. The "crawling" software looks for unique words but does not include prepositions and conjunctions such as "to," "in," "or," "and," etc.

The results are analogous to a book index alphabetizing *all* the words in the book. This is obviously not the most efficient way to locate pertinent information—you may as well just read the book itself.

Advanced functions of search engines, however, allow one to apply Boolean logic to the search process. For example, multiple words can be searched by looking for instances where all the words are present, or where some words are present and others are "not."

Some search engines also "stem" the word. "Stemming" is process of extracting the root of the word and ignoring plural versions or other modifications of the word. For example, the root of *jumped, jumping* and *jumps* is *jump*.

Accurate spelling and unambiguous terms (*not* like our "chips" example) are required for keyword searches to be successful. The correct spelling requirement applies to the document as well as the search criteria.

Because of the proliferation of unstructured data, the problem we now face is not finding information about a particular subject, but finding relevant information. In a recent Dartmouth College study, it was found that one in five or about 20% of all Web pages are less than 12 days old. When you consider that Internet spiders take 3 to 4 weeks to index information, the implication here is that you are automatically by default of the process missing 20% of the most current data.[6]

Keyword search systems are actually a subset of rules-based systems.

Rule-Based

An example of "rule-based" taxonomy could be that all documents that include the terms "San Francisco," "Chicago," "New Orleans" be listed in a category called "Cities, USA." The rules break down when an example like "Cambridge" is used. Is this Cambridge in Massachusetts or in England?

Taxonomies were developed to aid in the search for information. This has been done manually in libraries for hundreds of years. Various systems were developed to apply particular "rules" on how such taxonomies have traditionally been built. Rule-based taxonomy classifies documents based on the existence or absence of pre-above. Rule-based classification requires experts to create and maintain a rule for a document to be included in a given category. Experts organize concepts into categories using "If-Then" rules. These rules can support complex operation and decision trees and are very accurate. Rule-based systems have their supporters, then, because these 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 that topic.

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

[6] Mohomine
[7] Verity http://www.verity.com/techbuzz/content.html

a powerful and flexible means for automatically classifying content based on not just content itself but the metadata that describes the content's business context.[7] The down side of rule-based system is that expensive human domain experts have to write and maintain the rules. Other examples of rules are source of document, age, size and document type.

a) Statistical Text Analysis and Clustering

This technology observes and measures co-occurrences of words. For example, "Java" used in connection with Starbucks probably relates to a document about coffee instead of a programming language. Relative placement of words is important. Words in the first lines of a document are likely more important than information contained in the copyright section. Statistical analysis and clustering also look for word frequency, placement and grouping, as well as the distance between words in a document. Pattern analysis improves precision by resolving ambiguous or multiple meetings.

b) Bayesian Probability

The Bayesian approach attempts to learn the probabilities of words for a given category. An example of Bayesian probability applied is 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.

Applying a Bayesian algorithm sorts documents by examining the terms, words and phrases 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.

c) Semantic and Linguistic Clustering

Semantic analysis depends on a particular language and dialect. 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 the sentences identifying the subject, verbs and objects, like you did when you first studied grammar in grade school. Then sentence structure analysis is applied to extract the meaning. Stemming or reducing a word to its root also helps linguistic or semantic clustering.

d) Support Vector Machine

Support Vector Machine (SVM) is a refinement of taxonomy-by-example. 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—can be 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).[8] As this iterative method continuously analyses documents, it separates them into either the "relevant" side 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.

Combining Methodologies

Of course, no single taxonomy methodology, algorithm, or technology is superior to another for every possible application. The trend by more and more taxonomy software companies is to combine multiple methods to categorize the corpus of documents to increase the accuracy and the relevancy of grouping similar documents.

[8] Mohomine White Papers

Stages of Taxonomy

There are four stages of the taxonomy process. To help clarify this process, consider the example of a traditional book library.

1. Development of the taxonomy structure—the equivalent of the Dewey Decimal System

2. The process of categorizing the content and placing the pointers to the documents in the hierarchical structure—analogous to putting the books on the right shelves

3. Presentation of the information or the interface that helps you find the information—the card catalog would be the library equivalent

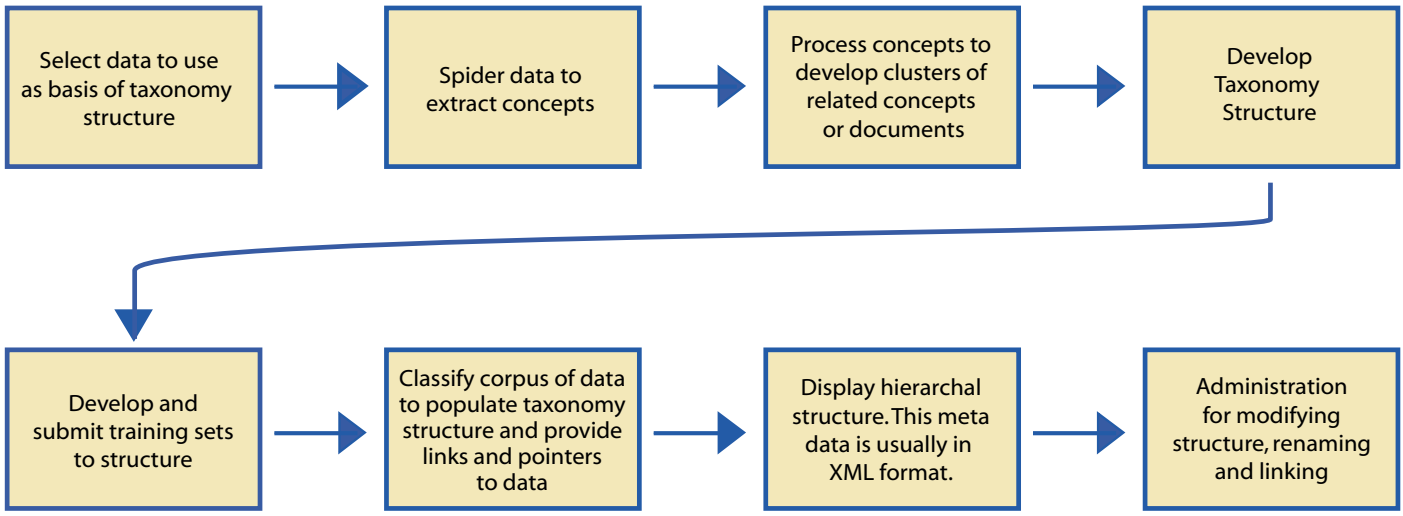
4. Monitoring new input and maintaining knowledge assets—analogous to the daily work of librarians; classifying new books, writing up new card catalogs, updating periodicals, and putting returned books back on the proper shelves.

Process

There are a number of subtleties that are not readily apparent when first reading through the descriptions of how these various taxonomy software solutions operate.

One major distinction often lost on the first time investigator is the differentiation between building the taxonomy structure and then populating the resulting hierarchical tree with pointers to the source of the information—in other words, "populating the tree." These are two separate and distinct operations and one should not assume that these processes are happening simultaneously or concurrently. Rather, these are serial steps.

Below is a diagram showing the general process or workflow of taxonomy software as it categorizes unstructured data. Vendors have their variations on this, but this schematic will give you an overview of how the software basically works. Some vendors leave out steps and other add steps in between these depending on the design approach and architecture of the taxonomy process.



API or Standalone Application

As with many new technologies, a subset of the vendors have developed standalone applications that come complete with end-user browser-based clients and MS Windows-based clients. Users and/or administrators can point the software at the body of documents to be classified on hard disks, servers, intranet sites, portals and Web sites. The taxonomy engines residing on servers perform the categorization process and usually populate a database of metadata.

Taxonomy software can be an enabling technology. It aids and enhances other applications that the user interacts with through a GUI. Most of the vendors support this concept and offer an API to integrate into other applications. Since many enterprise applications are custom-built this is an important consideration.

A number of vendors view this technology as eventually being an even more 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 will be the infrastructure component that correlate unstructured data. This design philosophy positions taxonomy software as a core module in a suite of products that work on the unstructured data within an organization.

Another series of vendors examined in the course of developing this report have added and tightly integrated the functionality of taxonomy into the search and retrieve application. Here are various ways we can group them:

- Application With GUI
- API Integration OEM
- Integrated Within Search Applications

The companies examined run the gamut from those which market just a development framework to the OEM market to those which market an application suite as a complete end-to-end solution for the enterprise. Most companies offer both approaches, and depending on their orientation emphasize either a strongly developed GUI as an application or a more extensive API for integration.

Company and Product Profile

The next section details one of the taxonomy software vendors that participated in this project with Delphi Group. This section is organized in the following way:

- Introduction – short description of the company and its origins
- Vision or customer study – a statement by the executives of the technology company relative to where they see the market heading or an example of a taxonomy implementation
- Technology and Products – a basic description of the underlying technology of the software and a description of the functions and features of the taxonomy product
- Assessment – a short evaluation of the technology and product approach by Delphi Group

Vendor Assessment Report: LingoMotors

Introduction

One of the principle challenges in creating and maintaining any taxonomy is that of ownership – specifically, ownership of the taxonomy by the end users. If this ownership is not expressed in the users direct involvement with building and diligently maintaining the taxonomy it will at worst never be considered reliable and at best fall quickly into a set of disrepair.

The challenge is a mechanical one. Users have neither the time nor the interest in becoming librarians. They will not spend time above and beyond the normal course of their responsibilities without an immediate payback on the time invested.

The history of desktop and personal productivity tools has been a constant testimonial to this phenomenon. The tools that succeed are almost always touted as far too simple for power users.

Consider early editions of Windows, Spreadsheets, desktop publishing software, and even the World Wide Web. Each was criticized for not having the apparent complexity necessary to deal with complex problems. But these criticisms often ignore the underlying power in these tools by discounting the appeal of a simple intuitive interface in the hands of motivated end users.

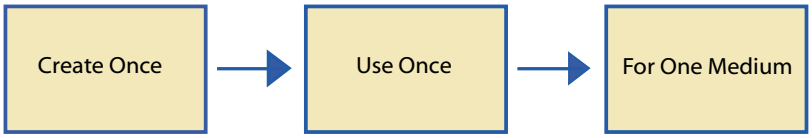
LingoMotors has applied the power of this sort of highly interactive and intuitive environment to taxonomy building. The result is a taxonomy tool set that is easily understood and accessed by end users with no formal training in taxonomy development, and little or no time to dedicate to taxonomy maintenance.

The core product, TurboCat™, interprets unstructured content such as published documents, web sites, and reports and creates highly granular subject categories – what LingoMotors calls Rich Information Objects (RIOs™) – which contain metadata that can be leveraged to build powerful information delivery systems.

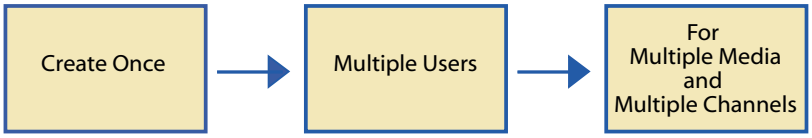
which are the sweet spot for LingoMotors, rely on new content. The better the content is labeled, the easier it is to structure, organize and locate it to increase process efficiencies and enable new revenue opportunities by enhancing the content lifecycle.

The integration of metadata solutions in the LingoMotors' approach enables media and publishing companies to efficiently repurpose information to create just-in-time ads, promotions and "event driven" information

The Typical Content Life Cycle Today



The Content Life Cycle Tomorrow



The Vision

At the heart of the LingoMotors approach is a simple philosophy manually inserting categories is expensive, slow and inconsistent. Therefore, automatic or semi-automatic categorization technologies, which assist companies in identifying highly granular subject topics and categories, are critical for managing and monetizing content. Historically automated categorization has been based on "black box" learning methods that required skilled knowledge engineers to implement and took significant investment of time to create.

LingoMotors core product, TurboCat, employs an innovative approach called seed-based learning, which enables ordinary users to define category profiles for a taxonomy on the fly while allowing flexible editing of categories as they change.

LingoMotors looks at the creation of taxonomies as more than a categorization exercise, preferring to regard it as an opportunity for additional revenue generation for information-rich businesses. Publishing and media businesses, across channels. This just-in-time information can be used to create new revenue through:

Analytics – performing detailed analysis of the content's use and make-up

Reselling – content that has already been created for one medium can be resold multiple times through multiple channels, such as secondary licensing and syndication

Advertising – providing ad contextualization through categorization that allows far greater flexibility

In addition much of the value of automating categorization comes from reducing manual labor and increasing classification consistency – two often-contrary objectives. Any product attempting to replace a classification expert has to deal with the challenges of human cognition and language. In particular, it has to address the following problems:

- **Diverse Terminology**
  There are often many different concepts and terms that refer to the same category and it is

difficult to make sure that the majority of these terms are indeed acquired for the profile. If important terms are missed or misplaced for any of these category profiles, the category will be inadequate.

- **Language Ambiguity**
  Language terms are often ambiguous. A term that is relevant for a profile may have one meaning that is relevant to that category as well as other unrelated meanings. An accurate categorization system must properly disambiguate the occurrences of such terms in order to categorize correctly.
- **Noisy Contexts**
  Document content usually focuses on a limited number of core topics as well as secondary, marginal or anecdotal topics. An accurate classification algorithm should properly distinguish between primary and secondary topics, understanding which are categories and which are just contextual noise.
- **Diversity of Document Styles**
  Documents of different types exhibit a huge variety in their style, language type, length and similar aspects. This variety poses a challenge to the classification algorithm, which must focus on the generic properties of categorization and should not be sensitive to document style.

These inherent complexities of human language and content pose major challenges for accurate categorization.

To address this TurboCat employs a mix of categorization approaches: automatic and manual. Humans are ultimately smarter than machines; therefore, some use of human knowledge is crucial. But highly trained knowledge experts are expensive.

The Technology and Products

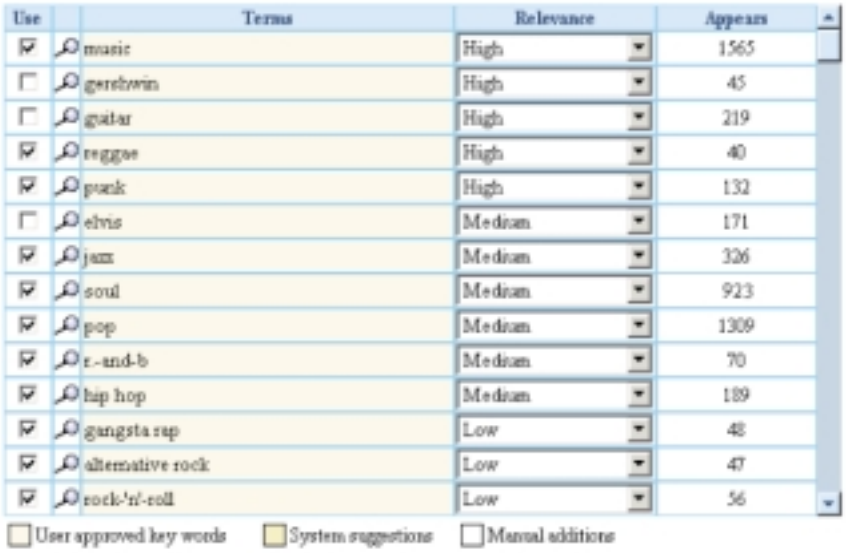
The TurboCat product suite includes the following products:

- Taxonomy Builder™ allows subject experts to build taxonomies and contextually rich categories that relate directly to your organization. Its seed-based learning approach lets you accomplish this quickly, without complicated rules, training examples or specialized technical staff.
- Taxonomy Mapper™ automatically converts ordinary content into RIOs by rapidly and accurately tagging the content based on the taxonomy and category definitions.
- Categorization Assistant™ provides an interactive environment for reviewing, managing and enhancing RIOs and the underlying taxonomy.
- Category Reporter™ analyzes the patterns of use and the make-up of a content collection and generates reports that enable businesses to enhance and increase the value of their content.
- TurboSearch™ provides fast, accurate access to categories and the underlying content, with category and directory browsing, enhanced natural-language search and a broad set of programming interfaces for integration into other applications.

LingoMotors products can be used individually and in combination with other categorization and digital asset management products. They can also be combined to form custom, end-to-end solutions.

Taxonomy Builder™ provides the capability for a subject expert to build taxonomies and contextually rich category profiles that relate directly to a specific organization. The profile is initiated with the use of one or more seed terms and phrases. The system is asked to "suggest" additional terms and phrases that are relevant to the category by performing numerous passes of the corpus to identify relevant terms. Terms are identified by the proximity to the initial seed term and phrases, the location of the words within the document, and the frequency and the syntax of the words. The subject expert can then decide which terms are to be used in the category profile and what their relevancy will be by choosing from a number of settings, like High, Medium, Low, Negative or Irrelevant. Taxonomy Builder does not require rules or training sets to be created, nor does it use a lexicon or thesaurus.

Building The Profile For The Category: MUSIC STYLES



The above figure shows how Taxonomy Builder's user interface enables subject experts to build category profiles. The category "Musical Styles" is being built from the seed term "music" – all of the additional words have been extracted from the existing corpus. The editor can interact with the system to choose which terms should be included in the profile and then rank the relevance of the terms.

The Categorization Assistant Reveals the Terms that Result in a Category Assignment

The above figure shows how the Categorization Assistant allows subject experts to participate in the content's categorization and make recommendations on changes. Here, the editor has clicked on the "Music" Category to find out which terms in the document led to it being categorized in "Music."

Assessment

LingoMotors targets organizations with complex publishing requirements. Its approach to taxonomy uses "seed-based learning" to expedite the processes of creating and maintaining taxonomies. Significant emphasis is put on the ability provided for end users to access the taxonomy building and maintenance process through a highly visual and interactive interface.

By combining automated taxonomy creation with human judgment LingoMotors is able to deliver a hybrid approach that not only speeds the creation of taxonomies but also increases the accuracy, consistency, and predictability of taxonomy. The hybrid approach is especially valuable for the creation of multiple personalized taxonomies where direct and unfettered involvement of end users is the only practical way to create taxonomies with adequate utility and precision.

LingoMotors also considers the exercise of building a taxonomy not only a means to cut manual categorization costs but also as a revenue enhancing opportunity by extending the content lifecycle to multiple users, channels, and media.

The core strength of the LingoMotors product set is its ability to combine powerful taxonomy modeling with a rapid taxonomy development environment. Fundamental to this approach is LingoMotors' use of RIOs – Rich Information Objects – which can be used to associate metadata with categories that are then reusable as business objects. This building block approach leverages the effort used to create taxonomy and also provides an asset base for long term enterprise or value chain taxonomy creation.

Controversies and Pitfalls

Taxonomy Strategy: The Time is Now!

As an organization you need to start organizing your data now. The longer you wait, the more unwieldy and overwhelming the task will be. Although manual classification works well for "small" volumes of data, how small is small and how big is big, depends on your resources, requirements and expectations. And due to the inherently ubiquitous nature of Infoglut, even heretofore "small" classification projects are becoming evermore unmanageable.

Our recommendation is that you start with an expert on taxonomy and classification – people with extensive training in the library sciences. Then, interview your internal domain experts and start them setting up agreed-upon categories.

Then decide how big a problem this is. How volatile is the information? How much time do your people spend looking for information? How much do you lose in opportunity costs because employees can't make an informed decision quickly?

Manual vs. Automatic

There seems to be a lot of "controversy" about manual taxonomy, versus automatic, versus a hybrid of the two. Delphi believes this is a tempest in a teapot and 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. This is a good time 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 (e.g. the Dewey Decimal System). The second is populating the structure. The range of taxonomy software providers is a continuum of approaches. This range is exemplified from Mohomine representing a completely automatic, strictly parent-child hierarchy to Wherewithal's and Entopia's approach, which entail collaboration of participants regarding both the construction of the taxonomy and actually populating the resulting hierarchical structure.
It's 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 and rename the nodes of the taxonomy structure to suit individual needs. The difference between these applications and tool sets is one of degree.

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, though, 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 machines and human processing to develop and maintain taxonomies.

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 onetime effort. Taxonomy is an ongoing process that requires 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 documents are constantly being added to repositories, while new versions of old

Manual		Automatic	
Pros	**Cons**	**Pros**	**Cons**
Accurate, Logical	Inefficient	Efficient	Limited Accuracy
Controllable	Does not scale	Scalable	Lacks Control
Inconsistent	Resource Intensive		Difficult to Train

documents are released, and out-of-date documents are removed from circulation. Changing strategies, evolving products, and advancing technologies all drive changes. Taxonomies may be industry- or even department-specific. Information today is by nature dynamic—consequently, categorization systems must be dynamic as well.

Directory Building vs. Hierarchical Categories

Another argument in this field emphasizes the fact that directories are about stored things as opposed to related concepts. Directories are virtual bins and do not necessarily reflect a hierarchical relationship. Proponents of this view would simply ask you to look at the directory structure on your personal computer. Carrying this argument forward supports the idea of customizing directories to help reflect your unique view of the world. The alternative approach is to implement a strict hierarchy with topics and subtopics organized in a strict grandparent-parent-child structure. Delphi believes that each approach has advantages and disadvantages. The choice is one of working style more than substance.

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 t0 20,000
- Large - +20,000

There are many very 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 five levels once again confront users with a kind of Infoglut, 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.

Librarians

Librarians are, first and foremost, people who help you find information. Corporate librarians are experts on how various categorization schema are designed. They know how to find information—that "needle in a haystack." The idea of an automatic taxonomy system may at first seem threatening to them. The reality is that, because of the dynamic nature of information, these experts will become more and more valuable as the amount of information expands exponentially.

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.

Speed, Accuracy, Robustness and Scalability

There is no universally accepted standard for evaluating the various algorithms or software configurations in regard to speed, accuracy, and scalability. When your organization is in the final stages of evaluation and has developed its short list of vendors, Delphi Group recommends testing 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.

Future Trends

Topic Maps

"Topic maps are a new ISO standard for describing knowledge structures and associating them with information resources. As such they constitute an enabling technology for knowledge management. Dubbed 'the GPS of the information universe,' topic maps are also

destined to provide powerful new ways of navigating large and interconnected corpora."[9]

The best way to describe how topic maps are different from taxonomies is to go back to our "chips" example. We may find that when we develop a topic map around chips, there are recipes for chocolate chip cookies. A taxonomy would list the various recipes and supply the pointers to them. A topic map would "associate" a white chocolate chip recipe with dark chocolate recipes and milk chocolate chip recipes. Then a topic map is formed between "chips," recipes for cookies, and various types of chocolate chips.

The components of a topic map are:

Topics - A "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 - These are topics linked to one or more information sources.

Associations - An association is a link element that asserts a relationship between two or more topics, e.g. 'Tosca' was written by 'Puccini,' 'Tosca' takes place in Rome, Rome is in Italy.

For more information on this subject visit http://www.topicmap.com/, a site dedicated to the use of topic maps.

Personalization

"The right information, at the right time, for the right person" has been the mantra of knowledge management software developers for some time. Taxonomies are by their very nature volatile, contextually relevant and personal. An individual's shifting priorities and goals can at any given time change the way they want to view information categories. If an individual is in an education mode, for example, they may want to discover information about a particular process or a new product. If that same individual is doing a competitive analysis, then they want to look at information from that perspective and context. A number of companies have added personalization functionality to their taxonomy products. Delphi Group expects this capability to be incorporated in more and more products in the future.

Vertical Taxonomies

Some taxonomy providers are developing and marketing vertical taxonomies geared toward a particular vertical market such as pharmaceuticals, financial, etc. The proponents of such an approach say this will "jump start" the process for your organization. Detractors, however, maintain that each organization has its own culture and its own way of categorizing. Trying to use someone else's taxonomy can be like wearing someone else's shoes—shoes that are already molded to that person's feet and will not fit another person even if the size is the same. Despite this, each of the vendors supplying pre-built taxonomies do allow customization of organization and the naming of the nodes for each of the categories.

Some of these currently available taxonomies are:

- Hardware/Software
- Healthcare/Pharmaceuticals
- Telecom
- HR
- Financial/Investment Banking
- Legal
- Sales & Marketing
- Geography

Other vertical taxonomies are: petrochemical, sports, entertainment, religion, education, sciences, Internet, insurance, construction, paper products, food, and government.

Taxonomy Integrated With Applications

Delphi Group believes taxonomy software is an enabling technology. The long-term evolution of the market for taxonomy will be its integration into applications such as Enterprise Portals and Content Management. The first integration of taxonomy technology will be with search and retrieval software. We are already seeing this trend as companies develop aggressive OEM programs.

Expect to see taxonomy software increasingly integrated with applications like these:

- Search & Retrieval
- Internet & Intranet
- Portals

- Content & Document Management
- Supply Chain
- CRM & Business Intelligence

Security

Security issues are among the key topics on the minds of information managers. Most taxonomy applications follow the standard security model for the server they run on within the enterprise system. This approach can be summarized by this simple rule: if you have access to the area where the document is stored, you have security clearance to see the document. Delphi Group believes this type of security implementation is the minimum level allowable. Security issues relative to unstructured data in the future will involve at least three different aspects:

1. Security attributes of each individual document. When the document is created, various properties will be assigned as to which individuals, which set of clearances, which departments, etc., will be permitted to view or edit the documents.
2. Security attributes of the individual. This type of security will be based on clearance levels, membership in particular departments or assigned role(s) in the organization.
3. Security issues regarding the overall operating environment. Examples of these considerations will be factors such as what time of day access is allowed, access from inside or outside the firewall, number of documents accessed, etc.

For instance, you probably do not want all your employees to see the list of personnel files that are assigned to specific categories of diseases. If you look up oncology you certainly don't want to see John Doe's personnel file linked to that topic. Even if you don't have access to John Doe's personnel file, the fact of its potential association with sensitive topics such as an employee's medical condition is of obvious concern. A number of technology providers are expanding their security functionality.

Ontologies

An Enterprise Ontology is a collection of terms and definitions relevant to business enterprises.[10] An ontology is more than a taxonomy or classification of terms. Although taxonomy contributes to the semantics of a term in a vocabulary, ontologies include richer relationships between terms. It is these rich relationships that enable the expression of domain-specific knowledge, without the need to include domain-specific terms.[11]

An ontology is more than an agreed-upon vocabulary, however. The terms in an ontology are selected with great care, ensuring that the most basic (abstract) foundational concepts and distinctions are defined and specified. The terms chosen form a complete set, whose relationship one to another is defined using formal techniques. It is these formally defined relationships that provide the semantic basis for the terminology chosen.

In the context of knowledge sharing, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents.

An ontology often takes the form of an extremely large database of words and phrases, their meanings and their conceptual relationships. Examples of conceptual relationships are: a "commissioner" is a member of a "commission"; "good" is an antonym to "bad"; and "lumber" has substance, i.e. "wood."

Taxonomies in effect are simplified ontologies. Where taxonomies generally classify categories in "broader" or "narrower" terms, ontologies can include more descriptive classifiers such as "located in" or "part of."

Beyond Text

A few companies in today's marketplace, whose directives have been mentioned in this report, are expanding the concept of taxonomy to embrace types of data other than text. Although somewhat governed by the limitation of accurate conversion of multimedia files (consisting of video and audio) into text, the classification

[9] http://www.gca.org/papers/xmleurope2000/papers/s11-01.html
[10] http://www.aiai.ed.ac.uk/~entprise/enterprise/ontology.html
[11] http://www.ontology.org/main/papers/faq.html

About This Document:

The product-specific information contained in this document is intended to provide an overview of a specific product and vendor at the date of publishing. Facts presented have been verified to the best of our ability with the vendor and actual users of the product where indicated, however, Delphi cannot insure the accuracy of this information since products, vendors, and market conditions change rapidly. Delphi Group makes no implied or explicit warranties, endorsements, or recommendations in this report nor should such warranties be inferred from its contents. A complete assessment of your specific application, the method of implementation for a given product or technology, and the current state of that product must be considered in order for a recommendation to be made on any product's suitability for your purpose, needs and requirements.

Delphi Group is a leading provider of business and technology advisory services to Global 2000 organizations. With offices established around the world, Delphi has assisted professionals across disciplines and industries at nearly every major national and global organization and branch of government. Its clients and subscribers include more than half of the Global 2000.

---

engines can still perform adequately to categorize these files by their content. The normal requirements of accurate conversion from audio content to text are not as stringent when pattern-matching analysis is applied. For example, one might view a news clip about former president Clinton and understand its theme without accurately knowing each word that was spoken.

End Note

"Knowing what you know" —that is, enjoying ready access to actionable information— is one of the major factors determining success in today's knowledge-driven economy. The ability to correlate the "whats" from relational structured databases to the "whys" embedded in the unstructured data of e-mails, reports, contracts, presentations, and Web pages will be a primary driver of innovation and competitive advantage. As in the past, when just-in-time inventory was critical to success, so today is "just in time knowledge" the critical factor.

Delphi Group's research shows that the unproductive time spent looking for information within the digital repositories of the enterprise is growing, and affecting more and more of the expensive personnel we have managing our organizations. Executives, managers and middle-office knowledge workers alike require sophisticated new tools for the delivery of relevant information.

Taxonomy software can reduce our reaction time to make informed and timely business decisions based on the knowledge and information contained within the unstructured data of an organization's digital documents. This software helps us form ideas from information we didn't know we had, while revealing relationships and correlations that would otherwise be lost in the ocean of information overload. Productivity and innovation come from seeing connections, evaluating importance, recognizing context, and understanding the implications of these correlations.

Although taxonomy software cannot completely stem the tide of "Infoglut," it can help us find the information we need to survive and prosper in the new knowledge-based economy—to truly "know what we know."

DELPHI® GROUP

A Delphi Group White Paper

Ten Post Office Square
Boston, MA 02109-4603
v (617) 247.1511
f (617) 247.4957
www.delphigroup.com