Chapter 2

USE OF ONTOLOGIES FOR ORGANIZATIONAL KNOWLEDGE MANAGEMENT AND KNOWLEDGE MANAGEMENT SYSTEMS

Vasudeva Varma
International Institute of Information Technology, Hyderabad, India

Abstract: This chapter describes the role of ontologies and corporate taxonomies in managing the content and knowledge within organizations. Managing content in a reusable and effective manner is becoming increasingly important in knowledge centric organizations as the amount of content generated, both text based and rich media, is growing exponentially. Search, categorization and document characterization, content staging and content delivery are the key technology challenges in knowledge management systems. This chapter describes how corporate taxonomies and ontologies can help in making sense of huge amount of content that gets generated across the locations in different languages and formats Different information silos can be connected and workflow and collaboration can be achieved using ontologies. As the KM solutions are moving from a centralized approach to a distributed approach, a framework where multiple taxonomies and ontologies can co-exist with uniform interfaces is needed.

Key words: Knowledge Management; Knowledge Management Systems (KMS); corporate taxonomy; categorization; document classification

1. INTRODUCTION

In this era of knowledge economy, every organization is producing a lot more content than before, resulting in a situation where we need to deal with the problem of information overload. As documents of structured and unstructured nature are growing exponentially; we have to find most relevant document(s) in the least possible time. Hence, obtaining very high precision

and recall in information retrieval systems is very important. In addition, mergers and acquisitions are major hurdles faced by the architects of information technology. As a number of organizations are being merged or acquired, making sure that the content of organizations can also be merged seamlessly is very important.

Recent studies in enterprise content management [Venkata, 2002] [Winkle, 2004] have estimated that 85% of the corporate content is in the form of unstructured data that doesn't fit neatly into relational database tables. Considering the effort and money that goes into creating such volumes of data, there is a compelling need for the organizations competing in today's economy to leverage unstructured data. Product development, sales and marketing, as well as executive planning and decision-making all depend upon information that resides within corporate documents. It is, hence, a challenge to manage the critical information that is scattered amongst various kinds of documents originating from various sources such as emails and web pages, various document-authoring applications, file systems, document management systems. In many corporations, it is a well known fact that decision makers are unable to leverage unstructured data to gain valuable business insights as these systems cannot easily exchange information and as a result users cannot easily explore and navigate documents from multiple sources.

The latest University of California at Berkeley study [Berkeley] into information growth estimates that 5 exabytes of recorded information were created worldwide in 2002 (equivalent to 800 Mb for each person on the planet). If access to these volumes of information is to be a benefit rather than a burden, then order and control become prerequisites. Information management techniques must be improved if we are to gain more control over these information flows, and taxonomies should be a key part of it.

To address this major challenge, companies need a platform to establish a shared vocabulary across disparate sources of unstructured information. If a company cannot provide a transparent view of its unstructured data, employees will neither be able to consistently locate nor share documents, thereby significantly hindering their ability to act effectively. The shared vocabulary is the backbone of the entire content and knowledge management infrastructure. This well-crafted vocabulary resides in the corporate taxonomy[1] or ontology. Corporate taxonomy is the key to success for building effective content and knowledge management systems. Content management system is an important sub-system of any corporate knowledge management initiatives.

[1] A simple definition of taxonomy is that it is a hierarchy of categories used to classify documents and other information. A corporate taxonomy is a way of representing the information available within an enterprise.

A classical taxonomy assumes that each element can only belong to one branch of the hierarchical tree. However, in a corporate environment, such formal ordering is neither feasible nor desirable. For example, a document on a competitor's product may be of interest to different departments in the organization for different reasons--forcing it into a single predefined category may be neater, but it also reduces its usefulness. Corporate taxonomies need to be flexible and pragmatic as well as consistent.

In the context of corporate intranet and knowledge organization, I would like to make note of two important characteristics of ontologies and taxonomies.

- **Ontology is more than an agreed vocabulary:** Ontology provides a set of well-founded constructs that can be leveraged to build meaningful higher level knowledge. The terms in taxonomies and ontologies 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 relationships to each other are defined using formal techniques. It is these formally defined relationships that provide a semantic basis for the terminology chosen.
- **Ontology is more than a 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.

Taxonomy-based knowledge management solutions are well known and widely practiced in the industry today. However, the limitations of corporate taxonomies are the entry points for ontology-based approaches. This issue will be discussed in detail later in the chapter. However, it is important to note that the organizational content management systems and knowledge-management systems make use of ontologies and taxonomies at several functional points that include: document categorization, indexing, document retrieval (whole or partial), user query expansion, query matching, and result verification. Since rich media documents are also becoming pervasive and important (perhaps more important than the textual documents) there is an emphasis on extending the ontologies work for multimedia documents as well.

In this chapter we first take a general look at the knowledge management (KM, henceforth) problems and issues, and the role of technology in KM in section two. The importance of categorization in KM arena is discussed in section three, where the limits of categorization and how taxonomy improves on categorization is our main emphasis. In section four, I will discuss the role of ontology in knowledge management systems and discuss how

ontologies can help where taxonomies expose their limitations. A framework called as UTON, Uniform Taxonomy and Ontology Network, where several ontologies and taxonomies can co-exist and accessible through uniform interface is also described here. Section five discusses future trends in using ontologies in knowledge management applications and presents our conclusions.

2. KNOWLEDGE MANAGEMENT TECHNOLOGIES

In the last decade, knowledge management (KM) has developed into an important success factor for organizations. KM has matured to provide a substantial insight into some of the fundamental challenges facing modern organizations, including the management of intellectual and social capital, the promotion of innovation and support for new forms of collaborative working [Woods, 2004]. As organizations could no longer draw sustainable competitive advantage by further improvements in operational effectiveness, knowledge has turned into a crucial competitive factor in business. Increasing complexity of products and services, globalization, virtual organizations and customer orientation are the major developments that demand more thorough management of knowledge – within and between organizations.

2.1 Knowledge Management Challenges

While it is easy to win the intellectual argument for KM, arguing for any kind of technology poses certain challenges. A chief knowledge officer or anyone else who is responsible for organizational KM initiatives will have a tough time while answering questions like “How KM can help company manage the knowledge that it needs in order to increase profits?” At times when business pressure is growing on KM groups to make costs and benefits transparent, it is of utmost importance to come up with a cogent elevator-ride answer to such a simple question.

Even though we will focus mostly on the technology aspects of KM in this chapter, I would like to emphasize that KM is more of a culture than a technology, a fact that is well documented. KM requires a strong commitment from all the stake holders, specially the top management, to information sharing, collaboration in order to deliver on its potential. KM can be defined simply as “The process of turning information into useful knowledge that is accessible when needed.”

The goal of KM is to enable all the individual knowledge workers and their communities to gain instantaneous access wherever, whenever and however needed to most relevant information so that they can make better decisions. It should also help in improving efficiency and in saving money. Organizations are putting an increasing amount of investment into their overall information architecture to provide a framework for managing data, information and knowledge. The growing value being placed on corporate taxonomies, metadata management and XML-based standards for information description are all part of that trend. But these solutions need also to be implemented in the light of a deep and rigorous understanding of the information needs of an organization. It means that knowledge management methodologies and practices have a vital role to play in guiding the evolution of a new generation of information architectures.

Various enterprise applications and content databases are designed keeping the specific user needs and business functions as major drivers. Although this helped to make these systems valuable and indispensable, this narrow focus has also created information silos that are difficult to integrate and harness. Some of the major issues with fragmented enterprise application are: proprietary data formats, non-standard interfaces, non-extensible and application-specific search and information access tools. This phenomenon across the organizations worldwide has given birth to disciplines such as Enterprise Application Integration (EAI) and many middleware software tools. Organizations spend a lot of money and effort in integrating enterprise wide applications.

2.2 Technology in Knowledge Management

In the past we have seen that much of technology has been abused in Knowledge Management. Especially during the early days of the KM adoption in organizations, technology was projected as a panacea for all knowledge needs. One major reason for this is the vendors of information management and CRM products have projected their products as solution to KM problems by hastily trying to re-label their offerings as KM solutions. After realizing the inadequacy of these solutions, the second generation KM implementers made the mistake of substituting enterprise application integration (EAI) and business intelligence projects for KM solutions. The current generation KM solutions are riding the wave of intranets and portals by externalizing internal information and by providing access to so called “corporate knowledge” to knowledge workers any where, any how and any time. We can also notice that the process and people aspects of KM, such as

workflow applications, collaborative work spaces, project management tools etc. are also appearing in the KM implementation agenda.

The current KM implementations illustrate a strong disillusionment that followed the misuse of many of these technologies in the past. It is important to use technologies as tools that help in implementing KM solutions. Most KM architects typically put together the technology infrastructure after studying the organizational need for knowledge management and identifying right KM solutions. The typical technology infrastructure is built using KM technology enablers that are listed in the following subsection.

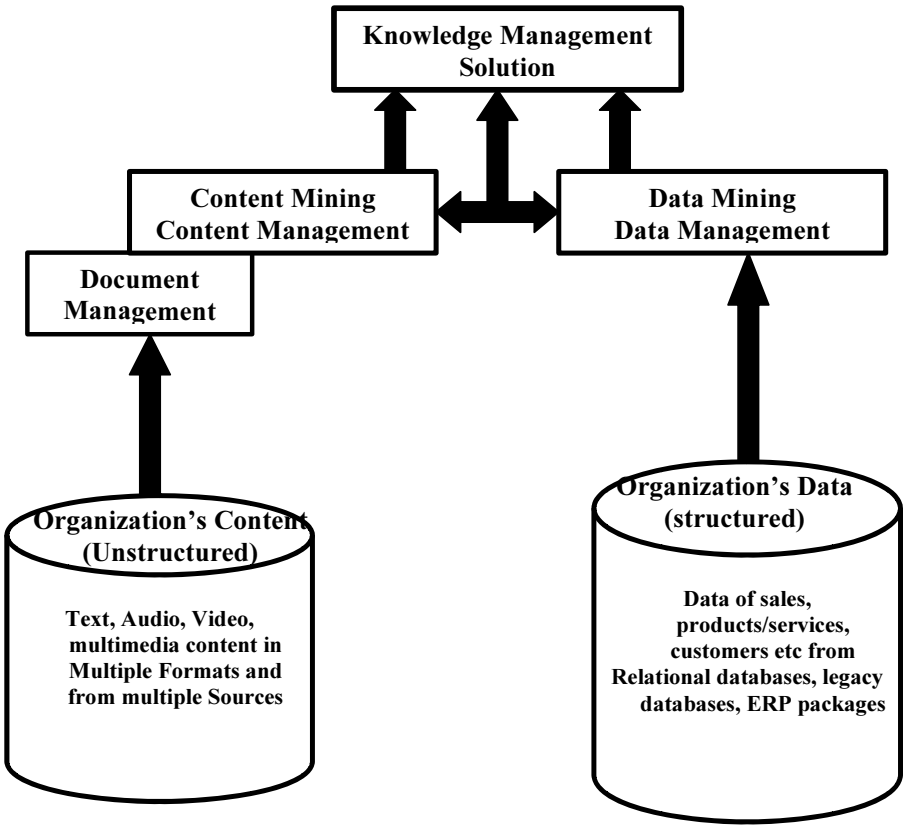

Figure 2-1. Knowledge Management Solution

A typical knowledge management solution will consist of document management, content management (including content analysis, content categorization, and search technologies), together with data mining and data warehousing sub systems that collaborate with each other to manage both

tacit and explicit organizational knowledge. The above diagram shows how a knowledge management solution can be seen as a confluence of data and content management solutions.

2.3 Technology Infrastructure for KM

The two most important and widely used components of the technology infrastructure of KM systems are search and categorization. These two provide an alternative to expensive content-integration projects. These information access and discovery tools must harness the information residing in disparate enterprise applications and information repositories.

Full-featured advanced search technologies are required to access, manage, and organize information stored in many and varied sources throughout the enterprise. Intuitive yet powerful search capabilities that enable users to look for mission-critical information and have it presented in a variety of formats to suit their particular need or preference is essential. Superior search and retrieval tools are capable of indexing and accessing information stored in a wide range of business systems, e-mail packages, document management systems, file systems and other repositories, regardless of whether the information is structured or unstructured.

It is important to find the right digital content in the shortest interaction time and in a very intuitive manner. We need to employ techniques such as "pearl growing" (improving and building upon the initial query). Ability to combine keyword or text-based approach with sample image or image parameter based approach. An example query would look some thing like: "show me all the Toyotas which are shaped like this [insert or select an image] and are in black color and registered in New Delhi".

The system should be able to navigate through vast digital content with ease and efficiency. Users should be able to interact with the digital content servers in a very meaningful manner using the latest technologies such as conversational systems to arrive at the right content in the fastest possible manner.

Categorization tools are the lifeblood of knowledge management initiatives. They are a key building block in that they add context to content. These tools leverage and enrich an existing corporate taxonomy. Solid categorization engines develop an intuitive, precise "table of contents" that enables users to find the information they require faster by providing them with a contextual map of search results—organizing related information by similar theme or concept.

There is another aspect to categorization that is called as media specific categorization. Once the document is parsed and various media objects are

extracted from the document, we need to index the sub-documents (or media objects) based on the media type. For example, video objects need to be parsed and indexed by the video indexers, similarly, textual objects need to be indexed by text indexers, image objects need to be indexed by the image indexers. There may be specializations and more than one indexer may be applicable for any specific media type. The content management system architecture needs to be extensible to add new indexing engines.

Besides the search and categorization, some knowledge management systems are equipped with the following tools and technologies:

- **Organizational knowledge network:** One of the main objectives of KM is to connect people and create communities that interact with each other for mutual benefit. To achieve this objective, KM architects build a soft infrastructure with the help of information technology, which includes: Email, creating user groups, building extranet bridge, creating corporate portals, online events (such as Virtual meeting interfaces, the corporate showcase) and virtual discussion forums (such as chat or instant messaging, peer-to-peer networking, video, file sharing, conversation, seminars and conferences). Some advanced KM systems also feature external knowledge hang-outs including activities such as touring customer hangouts and creating customer hangouts.
- **Crawlers and Agents:** Another key feature of a knowledge management solution is the provision of "intelligent agents" capable of pushing the required information to users. Agents allow users to define criteria, alerts or changes to documents, Web site content, or new information from other sources. Crawlers are enabling technologies that provide for Internet content and other external information sources to be included in user- and agent-based searches. This can also involve "brokered" searches whereby a search and retrieval solution brokers out searches to Internet-based search engines and then organizes those results as part of its own search. It is important to capture the digital assets from various sources and systems including that of file storage, web, databases and storage networks. The multi-media content can originate from any sources. The content crawlers and document filters can automatically 'grab' content from content databases, video sources, voice sources, XML Sources, emails etc. The content from various sources is being captured by crawlers. Depending on the document type and document format, various filters can be used to separate the media specific content.
- **Document Summarization:** Providing contextual summaries of documents, offering a "preview" format of the related result. This enables readers to see the document in a minimized form. Summarization is important for textual documents as well as rich media documents. For

example, video summarization is possible using the techniques that include video skimming or fast flipping of select frames and video information collages. For audio and text media we will use the summarization techniques developed in natural language processing and linguistics research. For images we can create thumb nails.

- **Personalization:** Customization of the content for an individual requires mixing and matching content elements. Personalization has become very important in web content management systems and this area has proven to be highly specialized and complex. A new trend of personalization of results obtained by search engines is gaining popularity within search community. Personalization takes into account various parameters such as past and present user behavior, the context in which the search is being made, and the predicted information need of the user.
- **Multiple Language Support:** In today's global economy, the ability to search and return result sets across a variety of not only major European languages but also Asian languages is essential.
- **Application "Hooks":** The ability of knowledge management tools to access and categorize enterprise business systems is critical. Hooks, or activators, that enable knowledge management technologies to index, categorize, retrieve, and display comprehensive, flexible result sets from packages such as Siebel, SAP, and J.D. Edwards are extremely valuable to organizations looking to ensure that the entire range of business content is available to knowledge workers conducting information-based activities.
- **Application Programming Interface (API):** The ability of organizations to tailor knowledge management tools, including information search and retrieval and categorization tools is essential. From an information search and retrieval perspective, this equates to enabling organizations to develop custom interfaces, leverage a variety of advanced features, and include natural language capabilities. From a categorization standpoint, API enables organizations to develop, manage, and modify business taxonomy, provide a variety of knowledge agents for users, and initiate supervised or unsupervised categorization; or a combination of the two to monitor and fine-tune the contextualization of enterprise content.

In sum, knowledge management tools are rapidly emerging as the primary means of leveraging business information. Typical KM implementations combine these tools and techniques with the benefits and capabilities of an enterprise portal and organizations can begin truly realizing and capitalizing on the wealth of information available to them.

I would like to end this section with a word of caution. KM is a controversial subject with no clearly defined boundaries. We can be more comfortable discussing the technical aspects that can aid the KM architecture rather than discuss about feasibility, strategy or planning of knowledge management. The technical approaches to managing knowledge have some known limitations. Information professionals are or should be involved in the creation and maintenance of the intellectual infrastructure of their organization. While technology and organizational infrastructures have received more attention and resources, some of the imbalance could be corrected through the intelligent utilization and integration of new software, new methods of working with both content providers and content consumers, and new ways of presenting information.

3. CATEGORIZATION AND TAXONOMY IN KM SYSTEMS

An important problem in content and knowledge management systems is that of organizing a large set of documents either in response to a query or during the staging phase or characterization phase of the document. Text categorization and text clustering are two techniques that are applied to achieve this task. Text categorization refers to an algorithm or procedure that assigns a category(s) to a given document and clustering refers to an algorithm that sorts documents into a finite set of groups based on associations among the intra-document features.

Categorization is a supervised process where as clustering is unsupervised. Various studies as mentioned in [Hearst, 1999] explored the advantages and disadvantages of both these approaches. In this section, it is not our main focus to distinguish between the two but to treat them as text classification methods that play a major role in knowledge management systems in organizing documents.

There are also a very large number of companies offering their version of this new software and, of course, most claim that their approach is the best, the fastest, and the smartest (an informal survey in 2004, puts the number of categorization companies at nearly fifty). Further, more and more search and content management companies are scrambling to incorporate categorization into their products. The reason why search engine companies now are taking categorization very seriously is that when users can't find anything, the categorized content enables a browse or search/browse functionality. Often users prefer browsing, which is more successful than simple keyword search and facilitates knowledge discovery.

3.1 Document Categorization

As we saw before, one of the most central pieces of KM infrastructure is categorization. Moreover, like all infrastructure activities, integration with other components is essential. Categorization needs to be incorporated into content creation through content management tools and at the same time incorporated into content consumption through search and collaboration software. Categorization software classifies documents using taxonomy and a classification model. There are several algorithms that can be used in a classification model. Among them are:

- Statistical classifiers, which provide accurate classification for broad and abstract concepts based on robust training sets;
- Keyword classifiers, which excel at defining granular topics within broad statistically generated topics;
- Source classifiers, which leverage pre-existing definitions from sources; for example, documents from a Web site or news service that are already categorized;
- Boolean classifiers, which can accommodate complex classification rules, as well as many legacy information systems.

Each of these classifiers provides optimum results for different use cases, which occur in a normal document corpus and taxonomy. Often it makes sense to create rules between classifiers as to when and how they should operate. In many situations the optimal use of keyword and Boolean classifiers is to provide more granular topic distinctions under broad-based, statistically derived topics. In this case, it can be necessary not only to create a hierarchical relationship between topics, but also to create an explicit rule regarding how and when a classifier should operate based upon an earlier classification or condition being satisfied. Some of the most recent research in text classification indicates that a parallel classification architecture utilizing multiple classifiers provides the best results.

As regard to all these above-mentioned classification techniques, it should be noted that none of them categorizes the way humans do, which is both strength and a weakness. Typically, each category is represented as a multi-nominal distribution of sets of terms. These parameters are estimated from the frequency of certain sets of words and phrases in the training sets of documents. The probability that a document with such a distribution of terms would belong to a given category is given be *P(document | Category)*. The category chosen for the document is the one with maximum *P(Document | Category)* [Gutherie et al, 1999].

Text categorization has been mostly influenced by the statistical information retrieval models. But, some text categorization systems have been built that exhibit strong performance using hand-coded knowledge

bases [Hayes and Weinstein, 1991] and [Goodman, 1991]. There is also a middle path as suggested by [Riloff and Lorenzen, 1999] that benefits from a strong domain knowledge but acquires this domain knowledge in an automated fashion.

Categorization software can very quickly scan every word in a document and analyze the frequencies of patterns of words and based on a comparison with an existing taxonomy and assign the document to a particular category in the taxonomy. Some other things that are being done with this software are clustering or taxonomy building in which the software is simply pointed at a collection of say 10,000 to 100,000 documents, and search through all the combinations of words to find clumps or clusters of documents that appear to belong together. It's not as successful as the first use, but it can be an interesting way of aiding a human in creating or refining a taxonomy.

Another feature of categorization software is metadata generation. The idea is that the software categorizes the document and then searches for keywords that are related to the category. This can be useful even if the suggested metadata isn't simply taken, since authors or editors work better selecting from an existing set of keywords than when starting fresh with a blank field.

One reason for preferring auto-categorization on intranet is the sheer amount of unstructured but very valuable content that resides on corporate intranets. However, because of the factors noted above, it requires a different mix of automatic and manual categorization and also calls for a better auto-categorization than has been adequate for news feeds. The "browse and post" feature, which is guided by the corporate taxonomy, aids human editors to manually categorize the content.

3.2 Applications of Categorization

There are various areas where categorization software worked wonders. For example, news and content provider arena depends on auto categorization. The reason is clear, it is an environment in which you have tens of thousands of documents a day to categorize and so obviously it has to be done in an automated fashion. Various characteristics such as that the material is written by professionals who know how to write good titles and opening paragraphs helps the auto categorization software perform better. A related market is in sites that categorize web sites on the Internet into browse taxonomy. Though the pioneer in this area, Yahoo, started with all human editors, it uses auto categorization to make editors more productive by supporting, not replacing them.

A new and intriguing market is in the intelligence industry. They, like the publishing industry, have huge volumes of content to categorize. However there are two features of the intelligence industry that are different; they need a finer granularity of categorization and not coincidentally, the material is not designated for a community of readers but is routed to one or a few experts. Not only does the intelligence industry need more specific categories, they also need to categorize content, not just at the document level, but at the paragraph level. This requires a level of sophistication beyond the early simple Bayesian statistics.

Finally, the corporate intranet and organizational knowledge management is a very important area where categorization has a major role to play. It is difficult because all the things that make news feeds work very well when pushed through the auto-categorizer are missing on almost all corporate intranets.

3.3 Role of Taxonomy in Categorization

There are several challenges we face while dealing with content available on the internet or even on an organization's intranet. The content is written by a really wild mix of writers. Some of the content may be pure literature, which unfortunately may sits next to an accounting document which may be next to a scientific research paper. Some of the content may have good titles, some may have very bad titles and some may have every page on the site with the same title. Some of the documents may be a single page of HTML, some may be book-length PDF documents and some may have about a paragraph of content but links to all sorts of other pages or sites.

Given this situation, there is every reason to believe that the corporate information organization systems such as content and knowledge management systems will see the most lucrative employment of auto-categorization software. Certainly the need is great and there are a very large number of corporate intranets which makes the challenge worthwhile.

Taxonomy enables "disambiguation" based on word meanings–for example, to distinguish the meanings of "java" as coffee, a country or a programming language. One advantage of using Taxonomy for categorization is that the system does not need to be "trained" with representative documents, as systems based on statistical techniques do. Furthermore, having an existing knowledge base enables accurate categorization for very short documents as well as longer ones.

The most important challenge is to create a corporate taxonomy rather than to categorize thousands of documents. Another challenge is to create either a very broad taxonomy or else integrate a number of taxonomies.

Finally, there is a need for both general categorization to support browsing and a very deep, specific taxonomy that supports quickly finding a particular document or even a paragraph.

3.4 Role of Taxonomy in KM systems

Software support for the corporate taxonomy is not limited to the provision of automatic classification tools. More effort is now being invested in software that can increase the usability of taxonomies for both corporate users and consumers and in tools to support taxonomy design and management.

The experience gained in building intranet and e-commerce sites is driving the development of more flexible technologies for the definition, management and use of taxonomies. The goal is to combine:

- The need for control and order in information management,
- An understanding of how users navigate through large volumes of information, and
- The realities of corporate and e-commerce information models.

Anyone trying to implement a corporate taxonomy would know that the world does not fit easily into neatly labeled boxes.

The enterprise search industry realized that it cannot provide very relevant results without making use of taxonomy. In the early stages of KM solutions, knowledge workers spent more time looking for information than using it. A solution to this problem that has become very popular is to combine search with browse functionality. When the user knows exactly what she wants, she will use search functionality and when she is not sure, a combination of search and browse is more effective. The backbone of browse functionality is the corporate taxonomy. Browse and search works better than search. When the entire document collection is held together by taxonomy, the information access becomes easy for the user. Any organization that needs to make significant volumes of information available in an efficient and consistent way to its customers, partners or employees needs to understand the value of a serious approach to taxonomy design and management.

Without an up-to-date Enterprise Content Taxonomy database to understand the inter-relationships, controls, and potential liabilities of structured and unstructured content, we suspect regulations such as Sarbanes-Oxley, and concerns over possible shareholder and other litigation, would result in a gridlock over any future document destruction.

Most organizations skip formal taxonomy development and rush into deployment by building their classification system around index keywords for retrieval. This is not a complete taxonomy and causes many problems

later with their document management system, as well as missing many reengineering and process improvement opportunities. Taxonomy incorporates the practice of characterizing the context and relationships among documents and their retrieval aspects.

In the following, we list some scenarios resulting from the lack of a consistent taxonomy across all areas of the enterprise:

- Isolated silos of information systems and processes expand, causing significant cost, duplication of effort, and liabilities to the organization. How many times do you have to give your personal and insurance information to different departments of the same hospital?
- One group calls a form or report one name, another group another name. Both groups file them. Retention rules are applied to the form under one name, but not the other. When litigation occurs and a discovery action results, the information which was properly destroyed in one system is discovered in another and becomes a liability
- One department creates a business form with many data fields. Another form already exists with the same fields in a different layout. A third version exists as an electronic web form.
- Different courts, social services agencies, and prison systems file the same paper documents. Most of this paper is created by other government agencies. When courts or other agencies request the information, it is copied and exchanged on paper. 90% of this paper was originally generated electronically, yet a half-dozen agencies each undertake the labor to scan and index or file this paper in their own systems—and then exchange the data on paper.
- A bank forecloses on someone's unpaid mortgage, while that same bank issues them a new credit card.

3.5 Creation and Maintenance of a Corporate Taxonomy

The task of creating taxonomy can be daunting. Whereas Web sites are often organized using at most a few hundred topics, enterprise taxonomies can often contain upward of 5,000 to 15,000 nodes. Developing such taxonomies manually is an extremely labor-intensive effort. We have encountered situations where companies developing taxonomies have averaged three person-days per topic for several thousand topics. Finally, in the absence of a neutral analysis of a document corpus, manual taxonomy development can easily run afoul of internal political agendas.

There are some significant efforts in building ontologies and ontology systems like CYC [Guha et al, 1990], [Lenat and Guha, 1990] [Lenat, 1995] and IEEE SUMO [Niles and Pease, 2001a], [Niles and Pease, 2001b]. The Upper CYC project captures 3000 most significant concepts of common knowledge and creates ontology. The SUMO (Suggested Upper Merged Ontology) project at IEEE Standard Upper Ontology Working Group addresses the problem of capturing high level, meta, philosophic and general kind of concepts and presenting them in more details. Some domain specific ontologies are also built using these formalisms. From these previous efforts, we know that building large scale ontologies pose a new set of problems, especially in an environment where ontologies are viewed as "live repositories" rather than frozen resources [Farquhar et al., 1996]. For example, any large scale ontology design teams will have to consider the following major issues: adding a new domain in the existing ontology network, changing the knowledge format, performance issues and guaranteed quality of service, scalability and ease of making any modifications. There are some engineering approaches [Varma, 2002] to build ontologies and to provide access to them by making use of existing resources such as WordNet [Fellbaum, 1999] and Open Directory Project called Directory Mozilla [DMOZ].

Maintaining a corporate taxonomy consists of two primary activities: incorporating new documents into the existing structure and changing the structure to accommodate new information that cannot fit into the existing one. Those processes are usually carried out through a combination of automation and human intervention. Classification techniques include keywords, statistical analyses that look for patterns of words, and use of a semantic network or ontology that analyzes words for their meaning in context. Analytical capabilities can help determine when a new category is needed, and how the documents would be redistributed. According to Laura Ramos, director at the Giga Information Group, maintenance is the most expensive part of a taxonomy project, yet is often overlooked in the planning process.

Leading industry analysts and solution providers focus on taxonomies that can bring a consistent and predictable sense of structure. For example, a geographic taxonomy is a hierarchical, general-to-specific representation, such as: world > continent > region > nation > state > city. Employing such taxonomy against all of an organization's information repositories allows the search system to automatically identify documents with references to the taxonomy nodes, thus allowing the information to be organized and analyzed (in this case) from a geographic perspective. The application of additional taxonomies, such as MeSH2, GO3 or DTIC®4 and others, can organize

information that is relevant to an organization's primary areas of interest and operation.

Much standard or industry-accepted taxonomy are readily available. Applying these enables a new realm of information categorization that can be either industry- or domain-specific, or general and horizontal. Taxonomies provide well-defined, stable organizational frameworks that cut across disparate data sets and functional areas, adding structure and the ability to find information that would otherwise be difficult to recognize. Once categorized, information can be populated into browsable folders or classifications allowing users to intuitively navigate to relevant concentrations of information. These classifications can mirror the taxonomy hierarchy used to categorize the information, or be constructed and populated to meet the specific organizational structures and perspectives of an enterprise.

Although difficult, creating a corporate taxonomy is just the first, albeit crucial, step to leveraging informational resources in support of organizational agility. The ongoing challenge that an enterprise must overcome is how to keep the taxonomy accurate and up-to-date. Because taxonomies exist within dynamic organizational and market environments, they must constantly change to accurately reflect the state of informational resources as well as organizational imperatives.

4. USE OF ONTOLOGY IN KNOWLEDGE MANAGEMENT

Typical Knowledge organization software can be divided into two types: content staging that includes content characterization, indexing, metadata creation, concept extraction, categorization, and summarization; and content delivery that includes data visualization, retrieval, broadcasting, and packaging. An example of content staging is categorization that automatically funnels documents into pre-defined hierarchical structures. We discussed this issue in detail in section three and saw how taxonomies can help in categorization. An example of content deployment is visualization tools such as animated or hyperbolic trees. These tools graphically represent large amounts of data and translate enterprise knowledge into an animated tree or web structure. In a visually stunning interface, wire frame graphics link a category with all its sub-categories.

Simple taxonomies that have a fixed relationship (e.g. "is-a") or no clearly defined relationship (e.g. "is related to") have many limitations effectively staging and delivering knowledge. However, ontologies are

richer than taxonomies as they allow different kinds of relations between concepts. In addition, these relations are governed by definable set of axioms such as Disjoint-ness, covering, equivalence, subsumption etc.

Ontology help in organizational knowledge management in several ways both in content and information staging as well as in content deployment. Though the way in which ontologies are used may be completely different. For example, ontological parameters for automatic document classification and for visualization tools such as animated or hyperbolic trees can be very different, though both share the goal of easing information access; they employ different techniques of information organization. Another difference between classification and visualization tools is that animated trees do not materially reorder content. Classification directories physically catalog documents in an enterprise portal. In contrast, visualization tools graphically link similar content together regardless of where material is located. A simple taxonomy will not be bale to provide this flexibility.

4.1 How Ontologies help KM

Similar to the role played by taxonomies in knowledge management applications, ontologies also act as repositories to organize knowledge and information based on a common vocabulary. They provide access to and optimize knowledge retrieval and support the mechanisms for communications and, therefore, the exchange of knowledge. They also help in reusing existing knowledge and facilitating reasoning and inferences on existing knowledge.

Ontology in knowledge management contributes directly to the application functionality. Ontologies help in all three fundamental knowledge management processes, namely, communication, integration, and reasoning. Once ontology has been created, it serves as a base for communication, facilitating knowledge transfer. To do this, it provides precise notation for queries on the domain of interest. Likewise, it facilitates the interpretation of messages, establishing a proper interpretation context. Then it serves to integrate varied knowledge sources. Finally, the most complex applications can use the ontologies to find new rules or patterns that had not appeared before.

The main purpose of any ontology is to enable communication between computer systems in a way that is independent of the individual system technologies, information architectures and application domain. The key ingredients that make up ontology are a vocabulary of basic terms and a precise specification of what those terms mean. The rich set of relations between these terms guide knowledge workers and knowledge systems

navigate through the corporate semantic space. Categories or directories provide a meaningful context for retrieved information because they delineate conceptual relationships. Within a category, searchers can hop from one associated concept to another, learn about related terms, or begin their search at a broader term in the hierarchy and move down to more specific instances of a concept.

Searching is an iterative venture, and people often cannot fully articulate what they are looking for. This type of information structure takes the pressure off users. Directories contextualize the search and knowledge management process. Because classification schemes explicitly delineate a structure of conceptual relationships, the meaning of a discrete term is not presented as "a thing in itself" but is understood in the context of surrounding terms.

Ontology also helps in improving communication. Directories and ontologies function to hook people and context together. They provide a common language, and workers better relate concepts across departments, divisions and companies. For example, what one company calls "CRM" another may term "infrastructure management," "one-to-one marketing" or "front-office applications". Ontologies promote collaboration by matching up users with similar interests. Research is more cumulative when, for example, an analyst studying European Union banking policies is linked to an employee researching French fiscal policies.

Ontology classifies information into logical categories that allow users to readily browse through content. They are often used in tandem with search and retrieval tools (keyword- or concept-based) to help locate target information. However, unlike search technology alone, ontologies reveal the overall structure of a knowledgebase, in a hierarchy that is visible to the user. The user navigates through sub-categories to narrow the search, a process that helps avoid false hits that are outside the area of interest. When used with search and retrieval tools, ontologies aid in efficiency by limiting the volume of material that must be searched.

4.2 Central versus Distributed Ontologies

There are two technologically different approaches in designing KM solutions. The first one looks at the organizational knowledge converging from various sources into a central repository. This Centralized KM approach focuses on providing central control, standardization. The second approach takes a distributed approach to managing organizational knowledge and attempts to coordinate the exchange of knowledge across different autonomous entities.

It was shown [Ehring et al, 2003] that the KM solutions need to move away from traditional approach of centralized knowledge repository to the realization of one or a few repositories of documents, organized around a single ontology or other meta-structures. In the past, it has been the experience of many KM implementers that the centralized KM systems are often deserted by end-users. The reasons for this may have been the social, distributed and subjective nature of working groups in any large organization, each with its own languages, process or tools. There is a major directional change for all the KM solutions from a stand alone KM application to Global-Ontology-Local-Application to Local-ontology-local-application to completely distributed applications.

The Semantic Web and Peer-to-Peer (SWAP) project (swap. semanticweb.org) demonstrates that multiple and distributed ontologies will allow support for decentralized KM applications. Participating knowledge work groups can maintain individual knowledge structures, while sharing knowledge in such ways that minimize the administration efforts and knowledge sharing and finding is easy. Considering this major paradigm shift in building knowledge management systems, we need to find a way in which we can work with multiple, distributed ontologies and corporate taxonomies. There should be a provision to shift from one ontology to another by the core knowledge staging and deployment engines with ease. The next subsection presents one such framework.

4.3 UTON - Universal Taxonomy and Ontology Network

Content and knowledge management systems should be able to operate with multiple taxonomies and ontologies at the same time. It should be possible to switch between taxonomies or ontologies depending on the context and the input document. Hence it is important to come up with a framework where multiple taxonomies or ontologies can co-exist and accessed using unified protocols.

In this section we briefly describe a framework that can be used to co-locate different taxonomies and ontologies, called Universal Taxonomy and Ontology Network (UTON). UTON stores concepts, relations among these concepts, cross linkages, language dependencies in its repository and provides interfaces to storage and retrieval functionality and the administrative functionality (including user and version management). The knowledge and semantic information is stored within the network in the form of a DAG (Directed Acyclic Graph). The storage and retrieval interfaces provided by ontology network are used by various media indexing and categorization components. Ontology developers, editors and administrators have different interfaces depending on their needs.

All these interfaces interact with higher level UTON objects such as Ontology, Concept, term and relation. If ontology consists of concepts belonging to more than one domain or sub domains, then another higher level object called context will come into play to help disambiguate concepts belonging to more than one domain. In the following, we describe each of the higher-level objects.

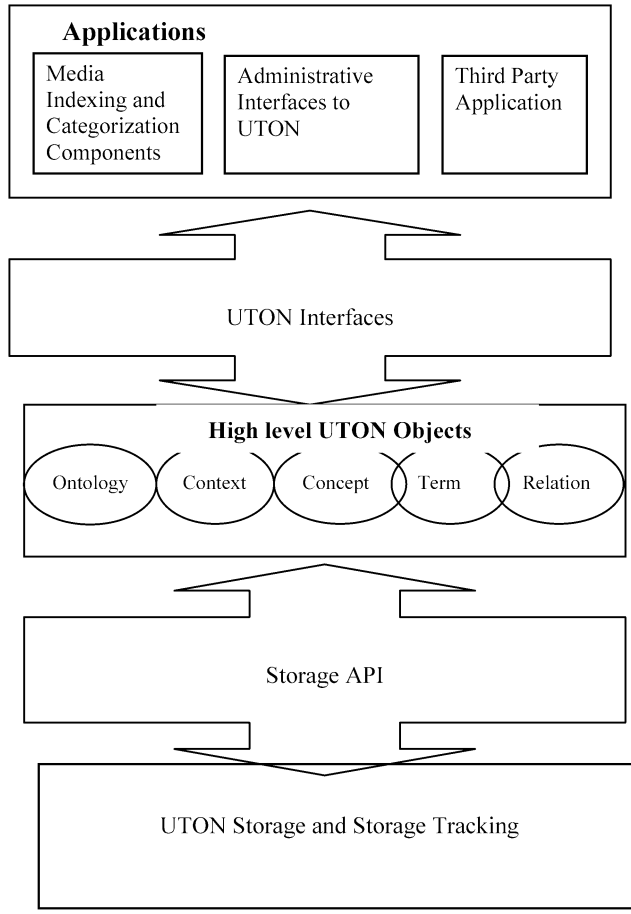

Figure 2-2. Architecture of UTON

As shown in the above figure, the general architecture components are:

- UTON Storage: The storage system is the place where the UTON data is stored – typically a Relational Database Management System (RDBMS).
- Storage API: Provides a unified access to the basic structures of UTON. The API should be accessible from any high level programming language.

- Higher level UTON objects: UTON objects are expressed in a data description language format, or as objects in any high level programming language. They are retrieved and stored using the storage API.
- Applications: Applications can use the UTON by integrating the ontology objects returned from the storage API in their program code.

This architecture and design of UTON [Varma, 2002] enables multiple ontologies and taxonomies to co-exist and makes it possible to access them in a unified manner. Our major focus is to build a network of large scale ontologies and taxonomies that are highly scalable and with high performance and guaranteed quality of service. All the components can be distributed and can be running on a set of server forms to obtain the required scalability and performance.

The UTON objects Ontology, Context, Concept, Term and Relation are independent entities and are related to each other in a loosely hierarchical fashion. Any two objects in neighboring layers typically have many-to-many relationships between themselves. The details of each of these objects are given below.

Ontology: The ontology is the topmost entity, necessary because the intention of UTON is to contain a network of taxonomies and ontologies, likely to be contributed by different sources. Depending on the number of domains the ontology contains a set of contexts will form the ontology itself. As attributes, the ontology has a name (mandatory and unique), a contributor, an owner, a status ("under development", "finished" ...) and documentation (an arbitrary string in which the contributor or the owner can specify relevant information).

Context: A context is actually a grouping entity; it is used to group terms and relations in the ontology. Within a given ontology, every context should have a unique name. The context object comes into picture when there is a possible existence of ambiguous concepts (see below for the description of concept), terms and relations among them when a given ontology covers more than one domain or sub domain, which is typically the case.

Concept: A concept is an entity representing some "thing": the actual entity in the real world and can be thought as a node within the ontology structure. Every concept has a unique id. A concept also has a triple "source-key-value", which is the description for that concept. The “source” identifies the source from which the description originates, the “key” is a string which gives a hint to the user on how he should interpret the value, and finally the “value” is the description of the concept. One concept can have more than one source-key-value triple, and thus have its meaning described in different ways. As an example, let's consider WordNet [Fellbaum, 1999]. In WordNet synsets denote a set of terms (with their "senses") which are equivalent. Every term also has a glossary, which is an informal description of the

meaning for that (particular sense of the) term. In this respect, from WordNet, we can extract two different descriptions for a concept, two different source-key-value triples, namely the glossary (Source: WordNet – Key: Glossary – Value: "<informal description denoted as a glossary in WordNet>") and the synset (Source: WordNet – Key: Glossary – Value: <enumeration of synonyms forming the synset>). As a different example, when a concept exists in various media (text, video, audio and image), a concept represented using source-key-value triple will give the appropriate media value, when retrieved using appropriate key.

Term: A term is an entity representing a lexical (textual) representation of a concept. Within one context, a term is unambiguous and, consequently, it can only be associated with one concept and of course, several different terms within one context can refer to the same concept, implicitly defining these terms as synonyms for this context. Terms in different context can also refer to the same concept, and in this way implicitly establish a connection between these two contexts.

Relation: A relation is a grouping element; it can be interpreted as a set of triples consisting of a starting term (also called the "headword" of the relation), a role (relation name) and a second term (also called the "tail" of the relation).

UTON was developed in the context of building information extraction, indexing and categorization engines for a content management system that is heavily rich media oriented. There was a compelling need to switch between ontologies depending on the domain and the context in which the application was running.

5. FUTURE TRENDS

As organizations realize the need to break down existing silos of information and question the established limits on information flow in the organization, there is an emerging trend to take a more holistic view of information architecture. This trend is due to the failure of many companies to respond quickly enough to changing conditions and being able to adopting themselves to information overflow.

The rising interest in ontologies and information classification, hopefully, will result in a better management of information assets across an organization. With the emergence of usable standards for information representation, information exchange and application integration (such as XML, RDF, WSDL, SOAP), we can finally start to overcome the recurrent barriers to developing a unified approach to managing information and knowledge across an organization. The increasing use of XML as a standard

for information description holds out the hope of developing semantically rich infrastructures in which new forms of information publishing, information discovery and information sharing will be possible.

All these developments in terms of standards and awareness in organizations regarding the role of ontologies bring up various possibilities in building and utilization of ontologies within the corporate world. Some such important possibilities are:

Multifaceted ontologies: Multi-faceted ontologies enable the user to navigate through a number of facets of ontology. An example of this is a search feature in a music library by artist, genre, instrument or composer. They also allow the different facets to be cross-referenced to narrow or widen a search as the user browses the categories. For example, in a cooking recipe portal, one can browse and select recipes by a combination of ingredients, cuisine and occasions. Developments in multifaceted taxonomies are also closely linked to new analytical and visualization capabilities that offer to transform our experience of search and navigation through large volumes of information.

Workflow and collaboration: Developing and managing ontologies is a collaborative project involving multiple stakeholders. It also needs clear procedures for change management. Integrated workflow tools and collaborative editing tools make it easier to manage ontologies in large organizations and places where ontologies have to monitored and adapted on a regular basis, such as shopping sites.

Search analytics and ontology management: Search analytics refers to the collection, analysis and exploitation of information about the way search technologies are used. The initial impetus for this development came from the need for e-commerce sites to know how users are searching their sites. The next step is for those techniques to be used within the enterprise. Better information on what users are searching for, and the ability to tailor results and navigation paths, offers a relatively easy way to improve information retrieval within an organization. There is a great opportunity for using search analytics in the design and maintenance of better ontology structures.

Visualization tools: Improved visualization capabilities can enhance the value of ontologies at two levels:

- **Usability–**providing visualization capabilities to the user enhances their ability to take advantage of the investment in an underlying ontology. Ontologies provide a basis for implementing existing visualization tools in a useful way and open the way for new tools that can help users visualize the multidimensional space in which they are searching.
- **Design and management–**improved means of visualizing ontology structure make it easier to ensure an efficient balance among categories and better fit with user expectations. Such developments are closely

linked to improved support for the rapid design, test and refinement of ontologies.

In order to develop the promise of the "knowledge-based economy," organizations have to develop a much better understanding of the nature of information and knowledge capital. Ontologies are linked into that development at a number of levels: The development of ontologies and corporate taxonomies is part of a general move toward developing methods and techniques for the management of intellectual capital. Those developments are also linked with the evolution of a mature information management architecture based on technologies such as content management, portals, search and data warehousing. Thirdly, a new generation of enterprise IT architectures based on Web services and other open standards is making possible new levels of information integration and interoperability.

As organizations evolve their information management processes, methodologies and technologies in coming years, ontology development will be given a much more prominent role within organizations. Ontology methods and technologies will themselves have to evolve if they are to meet the requirements of this general transformation in information management.

6. CONCLUSIONS

Observing current day KM implementations makes it clear that there is a sluggish adoption of ontology-based tools and technologies within the mainstream KM community. Poor understanding of the relationship between KM and ontology research is the major culprit behind this.

Modern organizations cannot survive without the help of technology in managing its information and knowledge. Architecting information and Knowledge management systems is increasingly becoming complex with a constant growth in the amount of content that gets generated. Corporate taxonomies and ontologies play a key role in building effective content management and knowledge management systems for organizations. In this chapter, I have discussed technology enablers in building content and knowledge management systems. These include search, categorization, creating knowledge network and infrastructure, crawlers and agents, summarization, personalization, multiple language support. Each of these features can benefit from the corporate taxonomies and ontologies.

There is a movement away from centralized Knowledge Management solutions towards distributed and peer-to-peer knowledge management systems. There is an urgent need to technically support this paradigm shift. UTON kind of frameworks can help in providing uniform interface to co-

located ontologies and help in implementing distributed knowledge management solutions.

The emerging trends in ontologies such as multi-faceted ontologies, workflow and collaboration, search analytics, ontology management tools and visualization tools will hopefully make the creation, usage and maintenance of ontologies in organizations easier and help in creating the much needed awareness on benefits of ontology use.

ACKNOWLEDGEMENTS

I would like to thank Bipin Indurkhya for reviewing this chapter at various stages and providing very useful comments. I thank the co-editors and other reviewers whose comments helped in making this chapter more readable.

REFERENCES

[Berkely] Berkey study on information growth http://www.sims.berkeley.edu/research/projects/how-much-info-2003.

[Davies et al., 2002] Davies, J., Fensel, D., and van Harmelen, F. (eds.), 2002. Towards the semantic web: ontology-driven knowledge management. John Wiley & Sons.

[DMOZ] The Open Directory Project, http://dmoz.org.

[Ehring et al., 2003] Ehrig, M., Tempich, C., Broekstra, J., van Harmelen, F., Sabou, M., Siebes, R.,Staab, S., Stuckenschmidt, H.: "SWAP – ontology-based knowledge management with peer-to-peer technology". In Sure, Y., Schnurr, H.P., eds.: Proceedings of the1st National "Workshop Ontologie-basiertes Wissensmanagement (WOW2003)".

[Farquhar et al., 1996] Farquhar, A., R. Fikes, J. Rice. "The Ontolingua Server: a Tool for Collaborative Ontology Construction". KAW96. November 1996. Also available as KSL-TR-96-26.

[Fellbaum, 1999] Fellbaum, Christiane (Ed). "WordNet: An electronic lexical database", MIT Press, 1999.

[Feldman, 2001] Feldman, Susan "Content Management" in eInform Volume 3, Issue 7, IDC News letter.

[Guha et al., 1990] Guha, R. V., D. B. Lenat, K. Pittman, D. Pratt, and M. Shepherd. "Cyc: A Midterm Report." *Communications of the ACM* 33, no. 8 (August 1990).

[Lamont, 2003] "Dynamic taxonomies: keeping up with changing content" Judith Lamont, KMWorld May 2003, Volume 12, Issue 5.

[Lenat and Guha, 1990] Lenat, D. B. and R. V. Guha. "Building Large Knowledge Based Systems". Reading, Massachusetts: Addison Wesley, 1990.

[Lenat, 1995] Lenat, D. B. "Steps to Sharing Knowledge." In *Toward Very Large Knowledge Bases*, edited by N.J.I. Mars. IOS Press, 1995.

[Niles and Pease, 2001a] Niles, I., and Pease, A., "Origins of the Standard Upper Merged Ontology: A Proposal for the IEEE Standard Upper Ontology" *in Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology*, 2001.

[Niles and Pease, 2001b] Niles, I., & Pease, A., "Toward a Standard Upper Ontology", *in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001).* 2001.

[PWC, 2003] Price Waterhouse Coopers, "Technology forecast 2003-2005".

[Sanda, 1999] Harabagiu Sanda M, Moldovan Dan I, "Knowledge processing on an extended WordNet" appeared in [Fellbaum, 1999].

[SemWeb] Semantic web: http://www.semanticweb.org.

[SWAP] Semantic web and Peer-to-Peer: http://swap.semanticweb.org.

[Staab et al., 2001] "Knowledge Processes and Ontologies" IEEE Intelligent Systems, 2001.

[Varma, 2002] Vasudeva Varma "Building Large-scale ontology networks", Language Engineering Conference, University of Hyderabad, December 13-15, 2002. IEEE Computer Society Publications, Pages: 121-127, ISBN 0-7695-1885-0.

[Venkata, 2002] "Taxonomies, Categorization and Organizational Agility", Ramana Venkata, Best Practices in Enterprise Knowledge Management, Volume II, October 2002, A Supplement to KMWorld, Vol 11, Issue 9.

[Woods, 2004], "KM past and future—Changing the rules of the game" Eric Woods, KMWorld-Volume 13, Issue 1, January 2004 Content, Document and Knowledge Management.

[Winkle, 2002] "Maximizing the Value of Enterprise Information by Leveraging Best of Breed Search and Categorization Software", Jon Van Winkle, Best Practices in Enterprise Content Management, Vol. IV, A Supplement to KMWorld, March 2004, Vol. 13, Issue 3.

http://www.springer.com/978-0-387-37019-4

Ontologies
A Handbook of Principles, Concepts and Applications in Information Systems
Kishore, R.; Ramesh, R. (Eds.)
2007, XIX, 930 p., Hardcover
ISBN: 978-0-387-37019-4