Organizing Knowledge with Ontologies and Taxonomies

Dr. Geoffrey P Malafsky
Mr. Brian D Newman
TECHi2
Fairfax, VA
techi2@techi2.com

Abstract: As with many management fields, KM must overcome one of the greatest challenges to effective management, namely, organizing a large amount of related but disjointed information into something that is useful, accurate, and trustworthy. Knowledge differs from simple information or data since it conveys the context, timeliness, confidence, and relationships among the individual pieces of information. Yet, this need for context and confidence is what makes using KM in practice so difficult since people "know" things at a personal level. Managing knowledge begins by defining a structure to organize information into categories of main concepts (e.g. by type, age, cost) and then by terms to group like items (e.g. fruit, meat, dairy). The concepts are defined in an ontology that maps the main ideas and their relationships. Once this conceptual map is made a set of terms must be created that defines how to label items according to the concepts described in the conceptual map. This structured set of terms is a taxonomy. This paper describes the basics of ontologies and taxonomies for KM and how to develop and implement them.

1 Introduction

Knowledge Management (KM) is a field that began as a means to collect, manage, and share the key knowledge of many people in a systematic manner. By bringing common ad hoc activities into a more systematic process, knowledge can managed and used as a valuable organizational asset to greatly improve people's productivity and effectiveness.

As with many management fields, KM must overcome one of the greatest challenges to effective management, namely, organizing a large amount of related but disjointed information into a useful, accurate, and trustworthy set of knowledge. The first challenge is to define what knowledge is and how it differs from information and data. Knowledge conveys the context, timeliness, confidence, and relationships among the individual pieces of information [1]. In particular, the contextual and confidence nature of knowledge is the fuzzy boundary between words as information for one person to knowledge for another. People "know" something when they believe the information to be true and useful in a specific context to make decisions or take actions.

Yet, this need for context and confidence is what makes using KM in practice so difficult since people "know" things at a personal level. Trying to understand and capture the personal views and needs of everyone in a group is extremely difficult, but doing so and packaging and tagging the bits and pieces for others to find and use at a later date is very difficult. KM as a management method therefore does not try to get all knowledge but only the key knowledge that is relevant and useful to the organization, most people, and in the most important activities. Even with this distillation and prioritization, it is a daunting challenge to package and tag knowledge into understandable, findable, and reusable units.

Managing knowledge includes defining a structure to organize information into categories of main concepts (e.g. by type, age, cost) and then by terms to group like items (e.g. fruit, meat, dairy). The concepts are defined in an ontology that maps the main ideas and their relationships. It also include the creation of a set of terms that defines how to label items according to the concepts described in the conceptual map. This structured set of terms is a taxonomy.

This paper describes the basics of ontologies and taxonomies for KM and how to develop and implement them.

2 What are Ontologies and Taxonomies

Ontologies and taxonomies provide a structure to the concepts and language used to organize knowledge. Without them, the knowledge will inevitably be difficult to find and reuse as people have very different perspectives on how the knowledge is related in the context of their situations.

2.1 Ontologies

Ontologies specify the primary concepts and the relationships among the concepts in a particular domain. The term means several things depending on the field in which it is used. In philosophy, ontology is concerned with the metaphysical nature and relationships of being. In contrast, computer science uses ontologies to describe specific conceptual terms and relationships in a standardized machine readable format [2-3].

Any KM effort must grapple with the challenge that there are several viable and valid perspectives on any given topic or business domain. To make the knowledge useful and an effective enabler of organizational success, the KM manager must create a single shared understanding among people of what the knowledge means to the organization within the context of its business domain and how it is intended to be used. An ontology provides this unifying map of concepts and relationships.

The ontology can be represented either graphically or in a structured text format. The former is usually used when the primary goal is to forge a shared understanding of the domain and provide guidance to the members of the group. The latter approach is most often used for computer applications that perform language analysis and concept matching, such as the goal of greater automated semantic capabilities on the Internet (i.e. the Semantic Web).

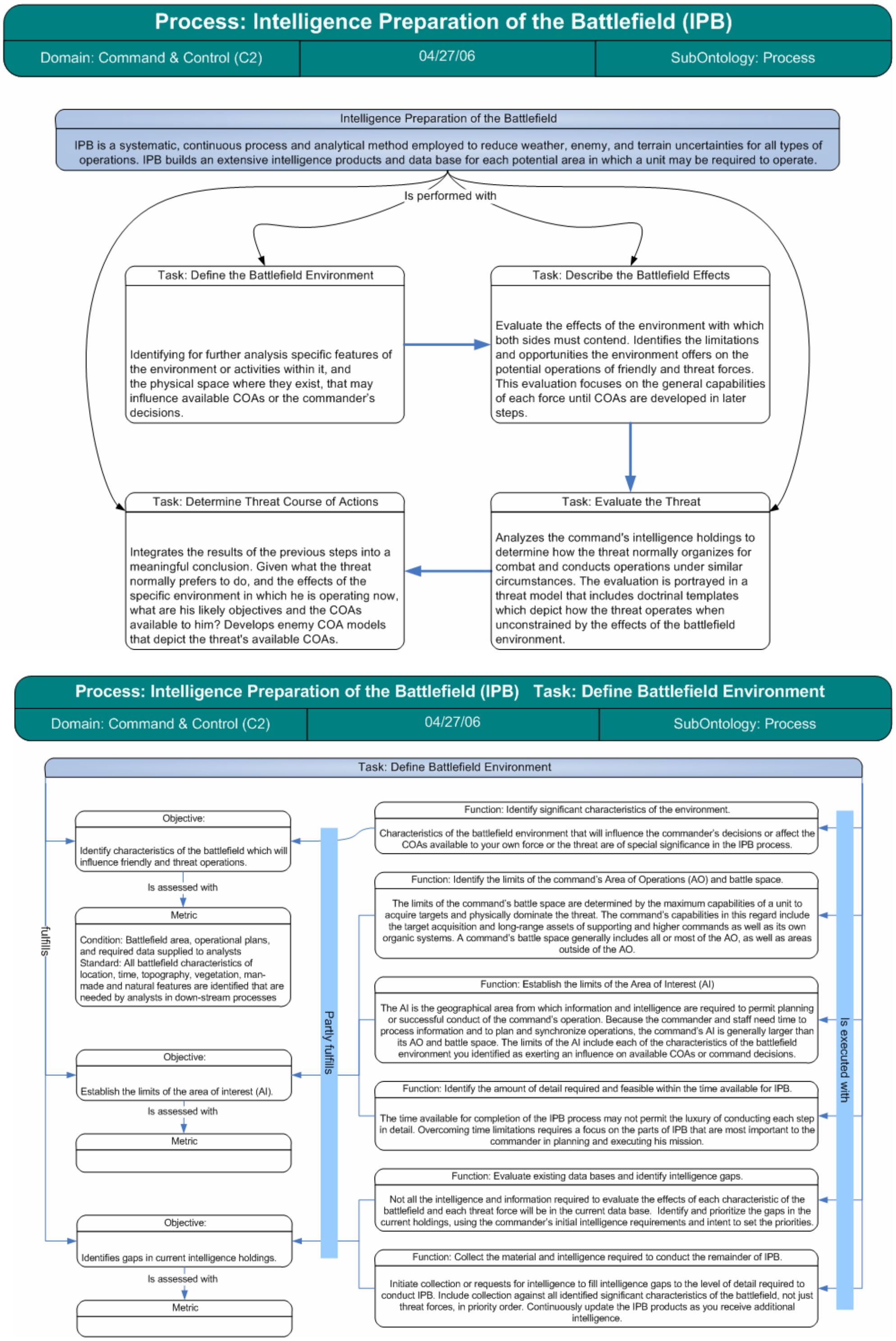
Machine readable ontologies require a computer language to define the concepts and associated relationships. One standard language is the OWL Web Ontology Language developed by the World Wide Web Consortium (W3C) [4]. An example of a small part of an OWL ontology is:

```
<owl:Class rdf:ID="WineGrape">
  <rdfs:subClassOf rdf:resource="&food;Grape" />
</owl:Class>
<WineGrape rdf:ID="CabernetSauvignonGrape" />
```

This example defines a class of items (WineGrape) and specifies that it is a type of another class of items (food:Grape) and then defines a single type of the new class (CabernetSauvignonGrape is a type of WineGrape).

While this format is a powerful enabler of computer semantic analysis, it is not very useful for people to use as a common baseline to forge a shared understanding and roadmap of how to use the concepts. For this purpose, we use a less rigorous textual but more graphical ontology definition to focus on defining the main concepts and relationships for a single perspective of a business domain. This perspective then serves as the foundation to guide the members of the organization in their activities.

One example of the human-centric ontology is shown in figure 1. This example is part of the author’s ontology framework for large-scale knowledge-based data environments. The ontology uses a standard military process and maps it to the ontology framework. Then, the specific work tasks comprising the major business domain are shown in the context defined by the process including individual tasks and the objective and performance metrics of each task.



Process: Intelligence Preparation of the Battlefield (IPB)

Domain: Command & Control (C2)

04/27/06

SubOntology: Process

Intelligence Preparation of the Battlefield

IPB is a systematic, continuous process and analytical method employed to reduce weather, enemy, and terrain uncertainties for all types of operations. IPB builds an extensive intelligence products and data base for each potential area in which a unit may be required to operate.

Task: Define the Battlefield Environment

Identifying for further analysis specific features of the environment or activities within it, and the physical space where they exist, that may influence available COAs or the commander's decisions.

Task: Describe the Battlefield Effects

Evaluate the effects of the environment with which both sides must contend. Identifies the limitations and opportunities the environment offers on the potential operations of friendly and threat forces. This evaluation focuses on the general capabilities of each force until COAs are developed in later steps.

Task: Evaluate the Threat

Analyzes the command's intelligence holdings to determine how the threat normally organizes for combat and conducts operations under similar circumstances. The evaluation is portrayed in a threat model that includes doctrinal templates which depict how the threat operates when unconstrained by the effects of the battlefield environment.

Task: Determine Threat Course of Actions

Integrates the results of the previous steps into a meaningful conclusion. Given what the threat normally prefers to do, and the effects of the specific environment in which he is operating now, what are his likely objectives and the COAs available to him? Develops enemy COA models that depict the threat's available COAs.

Process: Intelligence Preparation of the Battlefield (IPB) Task: Define Battlefield Environment

Domain: Command & Control (C2)

04/27/06

SubOntology: Process

Task: Define Battlefield Environment

Objective:

Identify characteristics of the battlefield which will influence friendly and threat operations.

Metric

Condition: Battlefield area, operational plans, and required data supplied to analysts
Standard: All battlefield characteristics of location, time, topography, vegetation, man-made and natural features are identified that are needed by analysts in down-stream processes

Objective:

Establish the limits of the area of interest (AI).

Metric

Objective:

Identifies gaps in current intelligence holdings.

Metric

Function: Identify significant characteristics of the environment.

Characteristics of the battlefield environment that will influence the commander's decisions or affect the COAs available to your own force or the threat are of special significance in the IPB process.

Function: Identify the limits of the command's Area of Operations (AO) and battle space.

The limits of the command's battle space are determined by the maximum capabilities of a unit to acquire targets and physically dominate the threat. The command's capabilities in this regard include the target acquisition and long-range assets of supporting and higher commands as well as its own organic systems. A command's battle space generally includes all or most of the AO, as well as areas outside of the AO.

Function: Establish the limits of the Area of Interest (AI)

The AI is the geographical area from which information and intelligence are required to permit planning or successful conduct of the command's operation. Because the commander and staff need time to process information and to plan and synchronize operations, the command's AI is generally larger than its AO and battle space. The limits of the AI include each of the characteristics of the battlefield environment you identified as exerting an influence on available COAs or command decisions.

Function: Identify the amount of detail required and feasible within the time available for IPB.

The time available for completion of the IPB process may not permit the luxury of conducting each step in detail. Overcoming time limitations requires a focus on the parts of IPB that are most important to the commander in planning and executing his mission.

Function: Evaluate existing data bases and identify intelligence gaps.

Not all the intelligence and information required to evaluate the effects of each characteristic of the battlefield and each threat force will be in the current data base. Identify and prioritize the gaps in the current holdings, using the commander's initial intelligence requirements and intent to set the priorities.

Function: Collect the material and intelligence required to conduct the remainder of IPB.

Initiate collection or requests for intelligence to fill intelligence gaps to the level of detail required to conduct IPB. Include collection against all identified significant characteristics of the battlefield, not just threat forces, in priority order. Continuously update the IPB products as you receive additional intelligence.

Figure 1 Ontology example showing TECHi2's ontology framework with a standard military process.

2.2 *Taxonomies*

Taxonomies are the classification scheme used to categorize a set of information items. They represent an agreed vocabulary of topics arranged around a particular theme. Although they can have either a hierarchical or non-hierarchical structure, we typically encounter hierarchical taxonomies such as in libraries, biology, or military organizations. This type has a tree-like structure with nodes branching into sub-nodes where each node represents a topic with a few descriptive words. For example, the following figure shows a portion of the familiar Dewey Decimal System that was introduced in 1876 as a general catalog of knowledge and is the most common system used in libraries.

600	Technology (Applied sciences)			
	630	Agriculture and related technologies		
		636	Animal husbandry	
			636.7	Dogs
			636.8	Cats

Figure 2 The hierarchical relationships of the Dewey Decimal System are expressed through structure and notation where numbers with more significant digits are a subclass of a number with fewer digits. The underlined digits demonstrate this notational hierarchy. [5]

The need to classify information is not new. One of the first large organized cataloguing and classification projects was in the center of ancient knowledge at the library in Alexandria, Egypt. Its first bibliographer Callimachus compiled the Pinakes, a 120 volume subject catalog of all the library's books. He is considered the founding father of librarians since he did not just list the books, but included the author, data on the text, and comments on authenticity to guide users [6]. However, many others throughout history solved the classification problem by strictly limiting the number of books by religious, political, or economic reasons, and then organizing the set by acquisition date, size, or other simple criteria.

Thus, classifying information becomes more important as the number of items increases and people have more trouble remembering what they have and where to find it. This is now crucial as we buckle under the immense volume of information available to everyone by the electronic networking of the world. We have become the fabled man dying of thirst while at sea as we search for the one or two items that answer our needs from within this sea of information. Indeed, KM is specifically focused on not only giving people the right information, but going to the trouble of distilling it into validated contextually connected knowledge that fuses information and data from a variety of distinct topical areas.

In order to classify information a framework must be defined. There are many existing standards from the Federal Government, consortia, and professional societies. For example, the Defense Technical Information Center (DTIC) has a technology taxonomy while the Standard Subject Identification Code (SSIC) is the standard for all DOD information including memorandums and records management. Similarly, the Library of Congress Classification (LOCC) is a commonly used general purpose system. However, taxonomies inevitably have a central theme that guides how the tree structure is arranged. For example, the LOCC and Dewey Decimal System are built from a perspective of classifying knowledge itself in a general purpose manner. Thus, the major LOCC headings include topics such as: Philosophy, Psychology, Religion; Auxiliary Sciences of History; History (General); and Fine Arts. In contrast, DTIC's major headings are more focused on technical issues and include: Aviation; Agriculture; Chemistry; and Electrotechnology and Fluidics. Clearly, trying to find a technology issue will be easier with DTIC than LOCC.

As we build a classification scheme, we define topics and order them based on relative importance to our organization and their level of detail. Thus, Dogs and Cats are included in the Dewey Decimal System under Animal Husbandry because they are specific instances of the general field. But, how far do we go in listing animals? Should we scour the world for every possibility and create a node for all animals? Do we include pets or do we create a separate heading for them, and if so, at what level of the taxonomy? These issues quickly arise while defining a taxonomy and lead to difficult decisions about what nodes should be included and which are subordinate to others. As a consequence, taxonomies grow in size and complexity to the point that people cannot remember the classification scheme and cannot use it to mentally map their interests and needs. For example, the LOCC has greater than 6000 nodes while SSIC has 2500 nodes. Even specialized taxonomies that are small parts of general purpose taxonomies like the LOCC become large as they attempt to cover all the important topics in a field, such as with the physics taxonomy from the American Institute of Physics, a portion of which is shown in the following figure. Note how the nodes gets extremely detailed to the point that a non-physicist probably cannot understand what they mean, but for a physicist the nodes are still broad definitions since there are many sub-specialties under a topic as specific as III-V semiconductors (node 81.05Ea).

80. INTERDISCIPLINARY PHYSICS AND RELATED AREAS OF SCIENCE AND TECHNOLOGY

81.	Materials science
81.05.✐t	Specific materials: fabrication, treatment, testing and analysis
∀∀∀∀∀	*Superconducting materials, see 74.70 and 74.72*
∀∀∀∀∀	*Magnetic materials, see 75.50*
∀∀∀∀∀	*Optical materials, see 42.70*
∀∀∀∀∀	*Dielectric, piezoelectric, and ferroelectric materials, see 77.80*
∀∀∀∀∀	*Colloids, gels, and emulsions, see 82.70.D, G, K respectively*
∀∀∀∀∀	*Biological materials, see 87.14*
81.05.Bx	Metals, semimetals, and alloys
81.05.Cy	Elemental semiconductors
81.05.Dz	II–VI semiconductors
81.05.Ea	III–V semiconductors
81.05.Gc	Amorphous semiconductors
81.05.Hd	Other semiconductors
81.05.Je	Ceramics and refractories (including borides, carbides, hydrides, nitrides, oxides, and silicides)
81.05.Kf	Glasses (including metallic glasses)
81.05.Lg	Polymers and plastics; rubber; synthetic and natural fibers; organometallic and organic materials
81.05.Mh	Cermets, ceramic and refractory composites
81.05.Ni	Dispersion-, fiber-, and platelet-reinforced metal-based composites
81.05.Pj	Glass-based composites, vitroceramics
81.05.Qk	Reinforced polymers and polymer-based composites
81.05.Rm	Porous materials; granular materials

Figure 3 Portion of the physics taxonomy from the American Institute of Physics.

This highlights the enormous complexity of creating an orderly method of classifying human knowledge and writings. We use the same words to convey different concepts depending upon the context of the discussion, what we expect other people to already know or not know, and how it relates to other activities and thoughts. People implicitly expect their perspective to be the central theme since it is most important to them. If the actual classification framework doesn't match the user's perspective, they will have to hunt to find something they feel should be easy to find. Extensive experience with enterprise taxonomies have shown that enterprise taxonomies must define which user perspective, or perspectives, will form the framework for the classification scheme [7-8]. For example, an enterprise taxonomy can be based on the core business areas, the organization hierarchy, primary product lines, or even an external schema. Previous projects have shown that it is very difficult for a single classification scheme to capture the many concepts embodied in a document and the multiple perspectives needed to create an intuitive navigation scheme for all of a system's users.

In order to construct a taxonomy for a KM endeavor we must define how knowledge differs from information and data. This is described by the Bloom Taxonomy of educational objectives that outlines the major cognitive areas of thinking and analyzing [9]. Bloom starts with knowledge and moves sequentially upward in cognitive skills with the following major areas [10]:

1. Knowledge: remembering previously learned material, recall facts or theories; bring to mind.
2. Comprehension: grasping the meaning of material; interpreting; predicting outcome and effects (estimating future trends).
3. Application: ability to use learned material in a new situation; apply rules, laws, methods, and theories.
4. Analysis: breaking down into parts; understanding, organization, clarifying, concluding.
5. Synthesis: ability to put parts together to form a new whole; unique communication; set of abstract relations.
6. Evaluation: ability to judge values far purpose; base on criteria; support judgment with reason (no guessing).

3 Organizing Your Knowledge

When it comes time to implement ontologies and taxonomies there are three options:

- develop the ontology then develop the supporting taxonomy.
- develop taxonomy and then develop the over-arching ontology
- develop the two in parallel.

While numerous rationale have been developed for the first two options, in practice it is the latter that normally prevails, and is reflected in the approach highlighted in the following paragraphs.

3.1 Define your scope

The first step in developing the combined ontology and taxonomy is to clearly scope the effort. A clearly defined scope is critical to the success of the effort. The question that can best help shape the scope of the effort is simply, *what purpose will the combined ontology and taxonomy serve?* The answer to this question serves several purposes:

- It sets bounds on the effort. These bounds are necessary to answering the basic managerial questions of how long will it take, and how much will it cost.
- It helps identify the primary domains and perspectives to be included.
- It should identify the specific business activities that will make direct use of the ontologies and taxonomies and how the resulting knowledge will be used to accomplish their mission.

Normally the answer to this question will fall into one of three categories:

- To serve as a common framework for knowledge sharing.
- To enable reuse of existing domain knowledge.
- To a better understand what the organization knows by separating domain knowledge from the operational knowledge and making assumptions explicit.

The first two of these, knowledge sharing and reuse are at the heart of most KM initiatives. Answers falling into the third category are indicative of advanced KM or academic initiatives. The following steps are applicable to all three scenarios and are specifically targeted toward those efforts focusing on knowledge sharing and reuse.

3.2 Check for Existing Ontologies and Taxonomies

Business operations today are also often dependent on, or required to adhere to one or more industry standards and may interface with applications that make use of existing ontologies or controlled vocabularies. For that reason, it is often best to use preexisting taxonomies and ontologies before launching into an extensive and possibly expensive development effort.

3.3 Identify Important Terms

If it is determined that existing ontologies and taxonomies are insufficient to meet the scope of the effort, then it is time to start collecting the raw materials for the new structures. This starts with identifying the key terms that are used to express the knowledge needed to enable specific business activities. At this point, it is important to list of all terms used to make statements or to explain to someone else what is needed to accomplish the business activities included within the scope of the effort. These can normally be found in corporate policy or operational instructions and from people with expertise in the activities.

While gathering these terms it is important to stay within the scope of the effort. Incorporating the full lexicon for a large multi-national conglomerate may look like a major accomplishment, but it will not help build an effective knowledge environment if the scope of the effort is smaller. Rather, the broader scope will make it more difficult to marry the concepts of the field to those used by the people you are trying to support.

3.4 *Define the Class Hierarchy*

Armed with the terms and concepts that are critical to expressing the knowledge needed to enable those business activities to be supported, the next step is to define a class hierarchy. Again, there are three ways to do this:

- Work from the top down. Start by identifying the general concepts (super-class identification) and then determine which of the others fall within those categories (subclass identification).
- Work from the bottom up . Start by developing cluster of related concepts (subclass grouping) and then look for the higher-order concepts that under which a given cluster of concepts might fall (super-class identification)
- Work from the top down and bottom up. This starts with the identification of the more important concepts first and then generalize and specialize them appropriately. The authors have found that this approach is the easiest to follow and has a lower risk of getting mired in the "semantic swamp."

When organizing terms and concepts the basic formula is:

> *If a class A is a super class of class B, then every instance of B is also an instance of A.*

In other words, the class B represents a concept that is a “kind of” A. As the class structure for your domain begins to take shape, check how well it expresses the way the people that are actually involved in the associated activates actually talk about what they are doing. Make sure that it reflects the actual context in which it will be used. The language of practical ontology or taxonomy should not to require the user to stop and translate between the way they normally think about things and the "official" term.

This is also the time to make sure that the vocabulary used to express the hierarchical class structure agrees with the real-world (and agreed upon) vocabulary and accurately serves to classify the information that will need to be exchanged between the people and applications involved in meeting the targeted activities. When these conditions occur, the class hierarchical effectively becomes the upper-layers of the taxonomy. This is a critical factor in insuring on-going semantic alignment between the resulting ontology and taxonomy.

3.5 Define Class Properties

The properties of a class are the types of information that distinguish one instance of a class from another (e.g., when used as a property, a person's DNA code or profile could be used to distinguish one person from all other members of the class "human") Properties can include:

- Intrinsic properties. The intrinsic property of a thing is a property that is essential to the thing, which loses its identity when the property changes. In our earlier example this would include things like the favor of the wine.
- Extrinsic properties such as the area in which the wine is from, etc.
- Part, which can be physical or abstract.
- Relationship. This includes relationships between individual members of the class and other items. (e.g., the maker of a wine, representing a relationship between a wine and a winery, and the grape the wine is made from.)

For specifics relating to the format and syntax by which properties are associated with classes, it is best to refer to the documentation provided for the specific technology.

4 Conclusion

This paper has described the basics of ontologies and taxonomies, how they differ and how they relate one to another in support of KM. While the development of these two important artifacts is often seen as separate activities, based on practical experience, they are really parallel components of the same effort. The process should clearly define the scope of the effort, use existing ontologies and taxonomies as much as possible, and maintain close alignment with the real-word context in which they will be used.

5 References

(1) Davenport, T. H. and L. Prusak, *Working Knowledge: How Organizations Manage What They Know*, Cambridge, Mass: Harvard Business School Press, 1998

(2) Berners-Lee, T., J.Hendler, and O.Lassila, “The Semantic Web” , *Scientific American*, May, 2001

(3) Holsapple, C, and K. D. Joshi, “A Knowledge Management Ontology”, in *Handbook of Knowledge Management* , 89, Springer-Verlag, NY, 2003, ed by C. Holsapple

(4) “OWL Web Ontology Language Guide”, W3C, at http://www.w3.org/TR/2004/REC-owl-guide-20040210/

(5) Introduction to the Dewey Decimal System, OCLC First Press, http://www.oclc.org/oclc/fp/about/about_the_ddc.htm

(6) Davis and Wiegard, Encyclopedia of Library History, Garland Publishing, NY, 1994

(7) G. M. Sacco, Dynamic Taxonomies: A Model for Large Information Bases, IEEE Transactions Knowledge and Data Engineering, 12 (2000) 468

(8) Glass, R. L. and I. Vessey, “Contemporary application-domain taxonomies”, *IEEE Software* , 12, 4, July 1995, 63

(9) B.S. Bloom, et al, Taxonomy of Educational Objectives: Handbook 1: Cognitive Domain, David McKay Co, NY, 1956

(10) R. A. Rademacher, Applying Bloom’s Taxonomy of Cognition to Knowledge Management Systems, Proc 1999 ACM SIGCPR Conf on Computer Personnel Research, 1999 , New Orleans, Louisiana