Building a Knowledge Management Taxonomy for the Department of Navy

Dr. Geoffrey P Malafsky

TechI^2 LLC

Achieving Knowledge Superiority, both for the Warfighter and support forces, requires us to capture, organize, and disseminate critical knowledge in a timely and succinct manner. We cannot merely expand access to knowledge, information, and data (KID) by building large repositories, since without a clear and easy method to find exactly what people need at any given moment our forces will continue to succumb to information overload and not achieve the objectives of Knowledge Superiority. The proliferation in the quantity of electronically available information is overwhelming people and network systems and is making it very difficult for users to find necessary information in the time they have available, especially in Knowledge Management (KM) Systems that strive to deliver answers and targeted links. The key to this success is to organize information according to how users think about it, which often varies from command to command, person to person, and day to day to facilitate the rapid and precise navigation of huge volumes of potentially relevant material to the few definitely pertinent items.

As part of the Enterprise KM and Integration efforts, DoNCIO is working with Task Force Web (TFW), PEO-IT, Navy-Marine Corps Intranet (NMCI), OPNAV, OSD, and other stakeholders to design architectural and content management standards and policies to allow all DoN personnel to effectively use the wealth of KID in the DoN, both explicitly available in electronic form and the tacit knowledge of our people. This will leverage the vast breadth and depth of our knowledge to achieve greater mission success, efficiency, and innovation.

A key part of this strategy is the methods and tools used to organize and classify the vast volume of KID throughout the DoN enterprise. DoNCIO is coordinating the development of the Enterprise Knowledge Management Taxonomy (EKMT) to serve as the common framework for effective user access and interactions with the NMCI Enterprise Portal and the applications web-enabled by TFW. This taxonomy embodies the Best Practices and Lessons Learned in organizing and classifying enterprise-scale information repositories within the DOD, Federal Government, and corporations. The EKMT bridges KM and Information Management by using both sets of design and architectural precepts to build a classification scheme that is logical and hierarchical, as well as centered on user's intuitive knowledge mapping. In addition, knowledge sharing requires the context in which the information was created and will be used, and the relationships among component items.

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

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 (Davis and Wiegard, Encyclopedia of Library History, Garland Publishing, NY, 1994). 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. When we ask a colleague what the Commanding Officer wants us to do, we don't want to be given the latest PowerPoint presentations or status reports, but rather a direct answer such as "The Admiral wants us to immediately get the readiness status of the Battle Group for a potential operation tomorrow. We need to contact both the J and N codes to get the newest logistics data and METOC analysis. If METOC can't accurately predict tomorrow's weather in the mission area, send out the new Micro-UAV with the miniature covert METOC system and have it feed data directly into the Course of Action and Sensor Performance Prediction systems right up to mission time." This is an answer that a human gives that does much more than point to the individual reports or web sites, and allows the questioner to immediately start acting and deciding their next activity.

A different way to solve this problem is to use automated search engines to find the best information that fulfills a user's query. This has been a very popular approach in the last few years with the growth of commercial search engine and portal tools like IBM's Textminer, Microsoft's Sharepoint, Verity, Convera, Altavista, Google, Ask Jeeves, and Autonomy. Yet, despite their marketing claims, performance metrics collected annually by the Federal

Government's Research and Development agencies DARPA and NIST show that these tools still cannot satisfy user needs on realistically large volumes of dense topic areas. The TREC results show precision levels of only approximately 40% for automatic searches and 60% for manual searches (K. S. Jones, Summary Performance Comparisons TREC-2 Through TREC-8, 1999, http://trec.nist.gov/pubs/trec8/t8_proceedings.html). It is easy to show why these systems have failed to solve the information retrieval need: a 10,000 item repository (small for enterprises like the DON) with 10 items directly pertaining to a query requires a 99.9% filtering accuracy to deliver these items to the user. Lower values result in either the user not getting the information at all or having the search engine deliver a larger number of lower relevancy ranked items (recall percentage) to ensure that the desired items are in the retrieved set. However, this latter approach, which is the one most often used, forces the user to wade through a large number of irrelevant responses, and has led to high levels of user frustration and disenchantment with these systems.

Now that we know we still need to classify information to help sort through the large number of items, the question becomes what framework to use. There are many existing standards from the Federal Government, DOD, consortia, and professional societies. For example, the Defense Technical Information Center (DTIC) has a technology taxonomy that is a standard for the DOD, 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 within the DOD will be easier with DTIC than LOCC since it was designed just for this purpose.

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 hair-splitting 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 2 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. If someone asks "How do we detect and track diesel submarines?", we can answer them by telling them what we know about state-of-the-art sonar transceivers and underwater acoustic wave signal processing, a listing of approved Navy ASW systems, a report on operational procedures, a statement of Navy organizations under CINCPACFLT involved in ASW, or even which acquisition programs develop and provide systems to the Fleet. In each case, the person asking the question will be implicitly expecting their perspective to be the central theme since it is most important to them. If the actual classification framework, say an acquisition-centric one, 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 in DOD, National Intelligence services, corporate intranets, and the Internet has shown that enterprise taxonomies must define which user perspective, or perspectives, will form the framework for the classification scheme (G. M. Sacco, Dynamic Taxonomies: A Model for Large Information Bases, IEEE Transactions Knowledge and Data Engineering, 12 (2000) 468; R. L. Glass, and I. Vessey, Contemporary application-domain taxonomies, IEEE Software , 12 (1995) 63). 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 knowledge taxonomy, we must define what we mean by knowledge and how knowledge differs from information and data. Does a KM system provide automated access to all electronically available information across the enterprise from a portal? Does it require full-time content creators and editors to produce summaries and analyses? Is a corporate personnel directory knowledge? The answer to all of these questions is: it depends! It depends on what the user needs to know at that moment and if that piece of information is all they need or only a small component of what they need. The following figure shows how information progressively moves from individual pieces of data that are devoid of context and relationships, up the cognitive staircase to information where pieces are grouped together, to knowledge where disparate information sources are brought together and fused in a validated way, and finally into a human's cognitive processes as understanding. At each step, there are greater connections made among the variety of related items with authenticity and strength of relationships explicitly made. One type of knowledge taxonomy is the famous Bloom Taxonomy of educational objectives that outlines the major cognitive areas of thinking and analyzing (B.S. Bloom, et al, Taxonomy of Educational Objectives: Handbook 1: Cognitive Domain, David McKay Co, NY, 1956). Bloom actually starts with knowledge and moves sequentially upward in cognitive skills (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) with the following major areas:

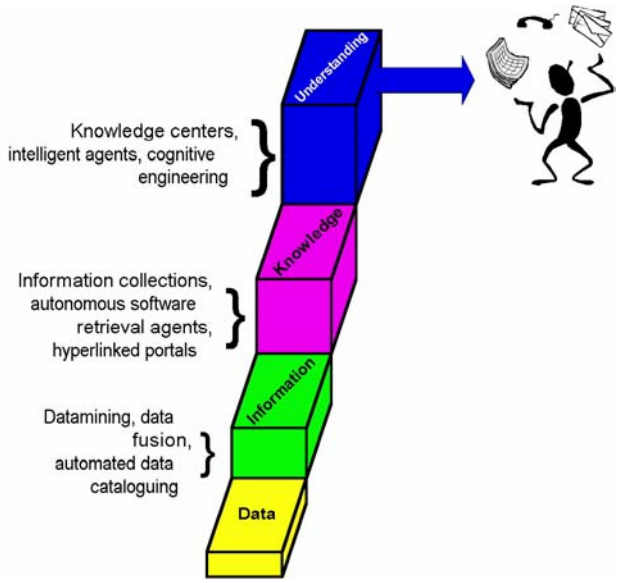

Figure 3 The cognitive staircase shows how data is transformed into information, knowledge, and ultimately human understanding as it progressively is fused, validated, connected, and placed in context.

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).

Which brings us to ontologies. Ontologies are the conceptual framework that people are really trying to express in a classification scheme. When we talk about Animal Husbandry or ASW systems, we are actually considering all the context and relationships to other topics that we have as a general understanding of these topics in our society. When engineers talk about sonar systems, they do not have to keep asking about how this topic relates to sound waves in water since that is common knowledge in their field. Yet, this contextual link is critical to understand why acoustic transceivers are important and how they relate to submarine detection and tracking and other topics. In contrast, a non-engineer will likely not have this knowledge and therefore not understand why the others are discussing seemingly disparate topics like signal processing and Sensor Performance Prediction algorithms. It is the group's general understanding of the concept of ASW systems that is the basis for classifying topics and determining which topics are more general and detailed to establish a hierarchy. These concepts inherently have connections to many other concepts with different strengths of relationships, as shown in the following figure. Ontologies can be created for many applications and have many coordinating themes, such as business topics, technology functions, and tactical military capabilities.

The ontology is translated into a hierarchy of descriptive categories that forms the taxonomic schema used to control the classification process. Even with a detailed taxonomy, the classification scheme cannot convey the relative importance of the taxonomy nodes within the document nor the relationship among the nodes, which is exactly the contextual information needed to transform information into knowledge. A great deal of knowledge and context is lost as the concept, which often takes a group of people hours to discuss to refine its meaning, is distilled into one or a few words that act as its representation in the taxonomy. For example, the SSIC has a node titled Data/Information Archiving under Operations under Operations and Readiness. As a user, this can also describe an Information Technology (IT) system function and therefore belongs under IT or some other heading that starts with an information theme. Similarly, this topic can be about new data storage techniques, both hardware and software, and therefore belongs under a Research and Development heading. Each case is correct and useful but difficult to determine which is best without more knowledge on the context of how the topic is being used. One common method to alleviate some of this discrepancy is to use a thesaurus of terms to augment the terms used for the taxonomy nodes. This allows a wider set of words to form the basis of determining what is relevant to a particular node in the same way as we might

use synonyms and antonyms to help someone understand a new word.

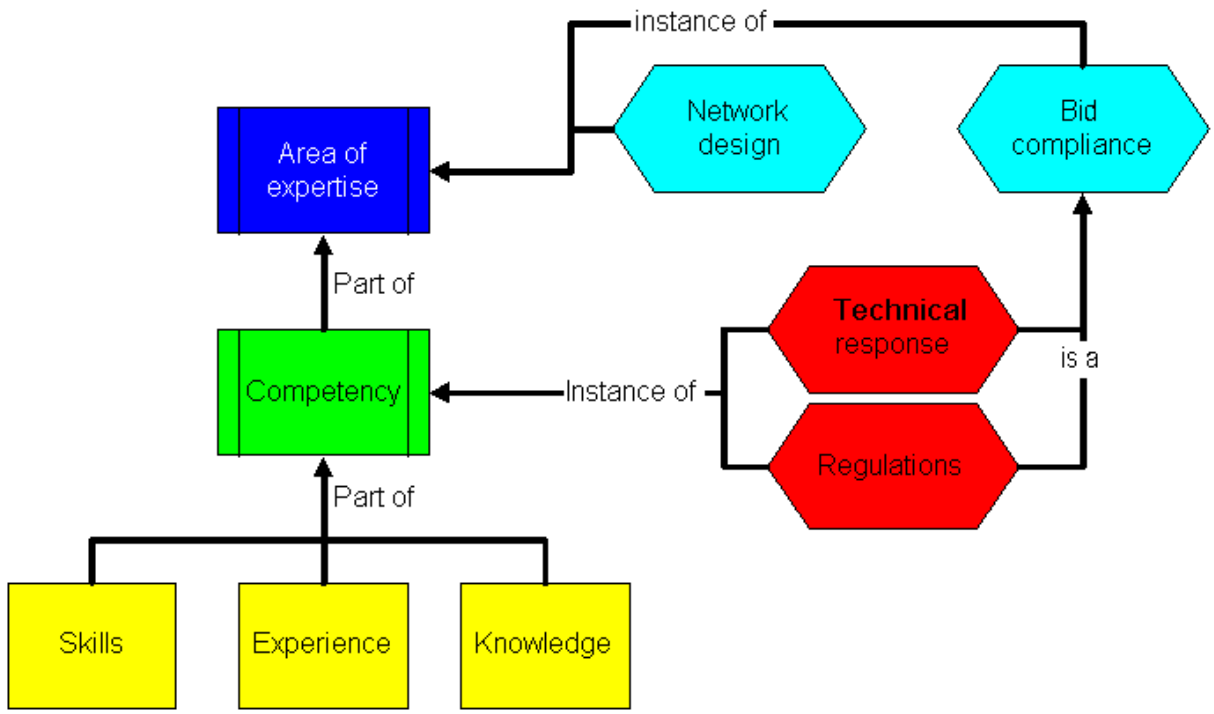

Figure 4 An ontology is the conceptual framework for a taxonomy and how people truly comprehend relationships among topics. The connections between items are explicitly noted as one item being a component of another item (part of), an example of a broader category (instance of), or a specific type of item (is a). Adapted from J. Vasconcelos, et al, A Group Memory System for Corporate Knowledge Management:An Ontological Approach, Proc 1st European Conf Knowledge Management, Sep 2000

Thus, users need a classification framework for the KID that is consistent across the enterprise but also allows individuals to intuitively navigate large volumes of resources. These seemingly conflicting objectives can be reconciled by constructing a knowledge taxonomy that blends the need for context and individuality with a consistent and structured framework.

What is a knowledge taxonomy?

A taxonomy is a structured set of names and descriptions used to organize sources in a consistent way. A typical taxonomy uses a logical arrangement but doesn't account for users' particular decision-making and action-taking needs. A knowledge taxonomy focuses on enabling efficient and interoperable retrieval and sharing of knowledge, information, and data across the enterprise by building in natural workflow and knowledge needs in an intuitive structure.

In order to ensure that the Lessons Learned from many enterprise scale projects are incorporated and current Best Practices are used, the Enterprise KM Taxonomy uses the following primary design principles:

1. User effectiveness in retrieving, sharing, and storing data, information, and knowledge is the primary metric of success
2. Multiple perspectives of organizing schema are needed to create intuitive navigational and classifying structures for the variety of user types
3. Local commands should be able to develop and use their own organizing schema in addition to the schema within the Enterprise KM Taxonomy
4. All of the domains, including locally developed sub-domains, must be completely cross-referenced to allow people to transparently access information across the enterprise without having to struggle with different and non-interoperable schema

These principles lead to the following taxonomy architectural characteristics:

1. Multiple domains and sub-domains
2. Significant overlap among domains is allowed to facilitate intuitive user navigation
3. Standard taxonomies are incorporated, such as Standard Subject Identification Code, Library of Congress Classification, and North American Industrial Classification System
4. New domains are created when user effectiveness could significantly decrease by coalescing partially similar schema
5. Semantic flexibility is incorporated by including taxonomic thesauri and planning for an ontological framework
6. Policies will be issued to define standard taxonomic and XML methods for interoperability

One key component of the approach is using a modular architecture of highly cross-referenced enterprise scale and local workgroup level domains, as shown below. However, this flexibility and user-centered architecture cannot be permitted to degenerate into a large number of disparate and non-interoperable classification schemes. All schemes must adhere to a set of adaptive but consistent standards and content management policies. The schemes can have substantial overlap in their domain entities if this can provide a significantly easier and more effective system. A mixture of customized and standard domains can be used to concurrently classify the data, information, and knowledge repositories thereby allowing users to choose one or more of the domains depending on their particular perspectives and needs at that moment.

The EKMT uses this mechanism to provide all users with an intuitive mapping of KID resources. The EKMT has nine primary domains that include custom developed topics for the DON's functional areas, as well as standard taxonomies such as SSCI, DTIC, LOCC, and the new North American Industry Classification System (NAICS) which was jointly developed by the USA, Canada, and Mexico to facilitate North American commerce. As shown in the figure below, these domains are all mapped to the full KID resources across the enterprise. To avoid users having to learn other taxonomy frameworks, they are completely cross-referenced in a central metadata registry that acts like a exhaustive index of all categories and how they map

across taxonomies. These domains are chosen to provide a variety of perspectives to the same information. This multi-faceted classification is known to represent KID content better than the typical single theme taxonomies like the Dewey Decimal System, LOCC, and SSIC (O. E. Taulbee, Classification in Information Storage and Retrieval, Proc ACM National Conf, 1965). Indeed, a formal approach to multi-faceted classification dates back to the 1920's when the Colon classification system was developed. This method breaks down the content into a set of terms with primary characteristics that can then be arranged in any hierarchical pattern that suits individual users (S. R. Ranganathan, Prolegomena to Library Classification, 2nd ed., Library Association, London; 1957).

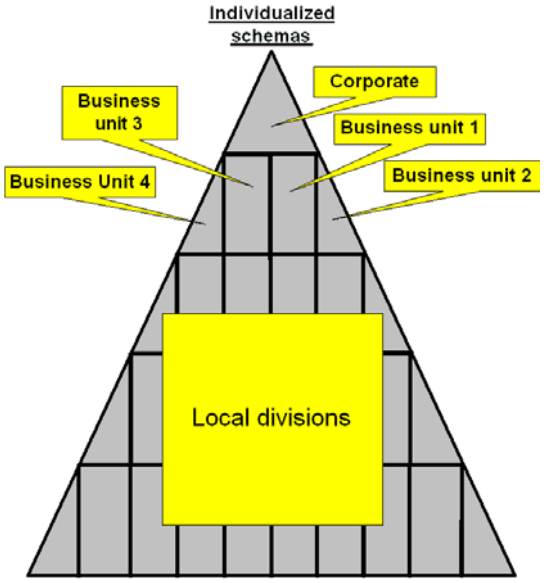

Figure 5 An enterprise scale taxonomy should incorporate corporate and local workgroup level classification schema to allow users to choose the most intuitive organizing framework for their particular needs. In this figure, different schemas are defined for various portions of the enterprise. In order to avoid a chaotic system of disparate schema, all schema developers must adhere to a flexible but consistent set of standards and policies.

Initially, the EKMT used the following set of primary domains.

1. DON organization
2. Geography (standard country codes and DON locations)
3. DON functional areas (22 sub-domains): Acquisition; Administration; Allies; Civilian Personnel; C3; Financial; Information Warfare; Intelligence & Cryptology; Logistics; Manpower; Medical; METOC; Modeling & Simulation; Naval Nuclear; Reserves; Readiness; Religion; Requirements, resources, assessments; Science & Technology; Test & evaluation; Training; Weapons

4. Library of Congress (government and general purpose standard)
5. Defense Technical Information Center: (DOD standard for technology systems)
6. Universal Naval Task List
7. North American Industrial Classification System

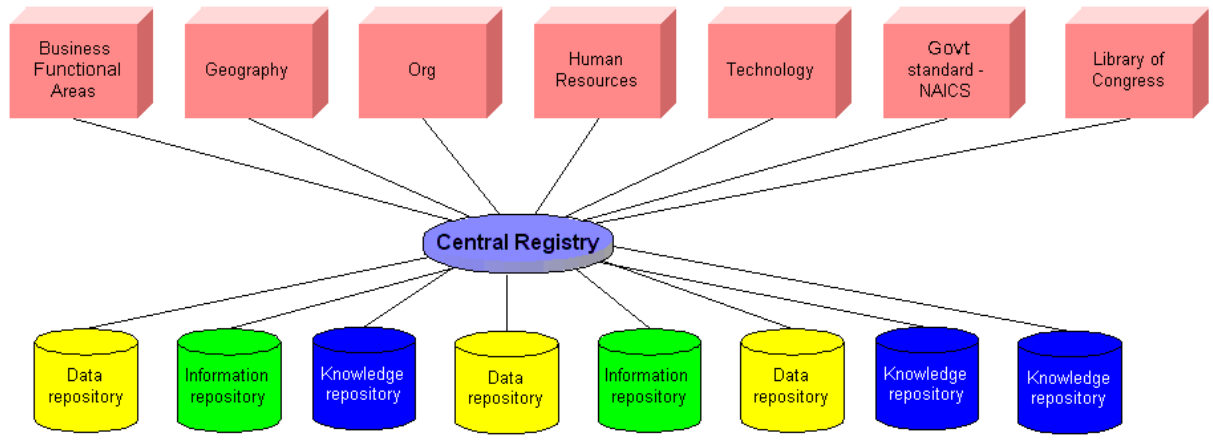

Figure 6 A KM taxonomy should use multiple primary domains (top row) to map all the data, information, and knowledge repositories to allow users to choose the most intuitive domain or domains for their particular needs. These domains can include customized schemes focused on the corporation's specific work environment and standardized schemes such as the Library of Congress.

The taxonomy domains and the original twenty-two functional areas were reviewed by a working group comprised of major stakeholders according to the major design precepts listed earlier. Through this process, the working group learned that the original DON functional areas did not accurately reflect the primary task areas across the DON enterprise. They determined that the entire functional area domain should be changed, and that the number of sub-domains limited to about ten to promote greater usability. However, the existing twenty-two functional areas are already being used in the DON and are possibly an OPNAV standard. Consequently, in keeping with the KM principle of focusing on user effectiveness, this domain was kept but renamed to allow users who need this thematic framework to have access to it. The new DON functional areas were defined through a usability sampling of stakeholders and became:

1. Logistics
2. Operations
3. Installations & Facilities
4. Administration
5. People
6. Acquisition
7. Education & Training
8. Science & Technology, Research & Development, Test & Evaluation (STRDTE)
9. Medical
10. Intelligence

11. Finance

The EKMT will be implemented on the federated architecture of application services and metadata repositories and registries being developed by TFW and NMCI, as shown in the following figure. This architecture uses physical databases and information repositories linked to a virtual network. The EKMT will be the classification framework unifying the metadata within the federated architecture.

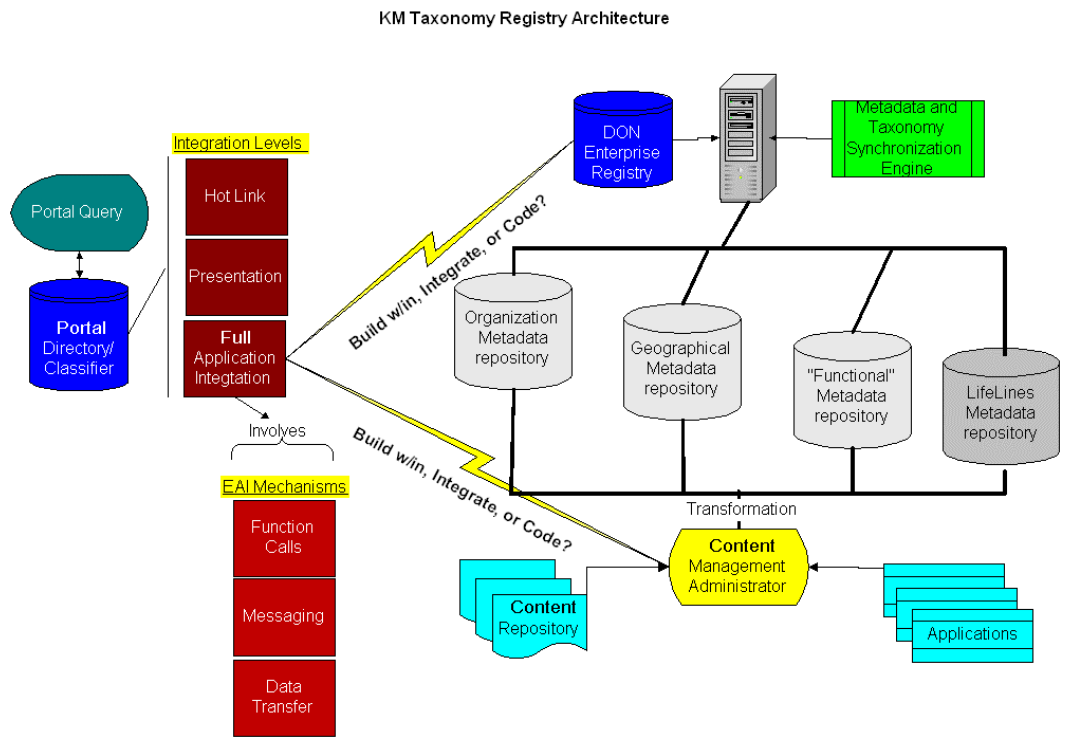

Figure 7 Schematic of taxonomy integration with the metadata repository architecture being developed for the DON by TFW and NMCI.

Ultimately, the EKMT will be implemented in XML to be part of the Enterprise Portal and application architecture of TFW and NMCI. This work is coordinated with the XML Working Group of the Data Management and Interoperability IPT as part of Application Integration planning. The central issue is the ability of the portal system to incorporate the functionality of metadata and XML repositories and registries for information retrieval as well as ecommerce and datawarehousing. The EKMT will exist within a XML schema that establishes the data structures for applications to use the predefined elements and attributes. Once a schema is populated with actual data, it becomes an XML document and can be used for operations.

The final interim version of the taxonomy is being distributed by DONCIO along with a policy statement for its use with information retrieval and KM systems throughout the DON. The

next phase of this project is working with TFW and NMCI to build the EKMT into XML metadata schema, namespace, repository, and registry to integrate with the Enterprise Portal and its embedded search and classification engines. Performance measurements are now being collected on the combined taxonomy-portal system and used to analyze and modify both the taxonomy and the portal architecture and setup. In addition, the working group is starting to define the next set of policies and standards to incorporate greater contextual meaning through the use of an ontological framework in XML. This is the forefront of information and knowledge management systems and uses prototype ontology frameworks such as OIL and Ontolingua.

Creating an intuitive yet consistent classification framework for all DON knowledge, information, and data will all us to finally corral our information systems and exploit their great potential to enable greater DON efficiency, effectiveness, and innovation. We cannot blindly pursue this path or we will fall to prey to the same narrow focus that hampers so much of our IT system, and which NMCI and TFW were specifically created to streamline. Only both continuously and vigilantly measuring and adapting our tools to user processes and needs can be ensure that we are truly achieving the goals of KM to quickly and precisely share and reuse knowledge throughout the DON enterprise whenever and wherever it is needed.