

Ontologies in Higher Education

*John Milam, Ph.D.
HigherEd.org, Inc.*

Knowledge Management (KM) is based in large part on systems that help users focus their attention on key information that is relevant, timely, and available on-demand. The preparation of this information requires processes for knowledge acquisition, engineering, and representation because “knowledge and expertise are embedded within otherwise diverse and scattered information sources” (Convera, 2004a, p.1).

Necessary to KM strategies is the act of “imposing a structure on the knowledge acquired in order to manage it effectively” (Benjamins et al, 1999, p. 1). This is because most information is unstructured, doesn’t fit easily into database models, and is at best “difficult to manage.” “Leveraging unstructured information is a chronic challenge for companies competing in today’s economy,” explains Venkata (2002, p. S12). Ontologies or taxonomies which categorize information represent “the most promising approach to solving the growing problem of information overload” (Inxight, 2003, p. 2).

In her discussion of taxonomies in the marketplace, Gumport explains that “Higher education often sees itself as an enterprise so unabashedly complex that it can’t be sorted, classified, or pigeonholed” (1997, p. 23). There is, however, a long history of grand classification schemes in higher education, including those of the National Center for Higher Education Management Systems (NCHEMS), the U.S. Department of Education, the National Science Foundation (NSF), and The Carnegie Foundation for the Advancement of Teaching.

This chapter provides an introduction to the use of ontologies and taxonomies in higher education. After a brief introduction to the nature of ontology, examples of ontology in higher education are reviewed. Issues in creating taxonomies, including their incorporation into search engines and concept maps, are then discussed. Software solutions for developing and utilizing taxonomies are presented next, along with problems and issues for implementation. Finally, future trends in the development of KM strategies for ontology are discussed.

I. The Nature of Ontology

An ontology is defined by Noy and McGuinness (2000, p. 1) as “a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic concepts in the domain and relations among them.” The domain is the subject area and ontologies are, basically, systems of categories (Sowa, 2004a). While there is an obvious philosophical underpinning to the nature of knowing, “the subject of ontology is the study of categories of things” and the product of such as study is called an ontology (Sowa, 2004b).

Sowa (2004a) discusses how ontologies contain boxes within boxes of categories, word senses, terms, directories, numbers, and character strings. “All of these lists, hierarchies, and networks are tightly interconnected collections of signs. But the primary connections are not in the bits and bytes that encode the signs, but in the minds of the people who interpret them. The

goal of various metadata proposals is to make these mental connections explicit by tagging the data with more signs” (Sowa, 2004a, p. 1).

Even storytelling, a technique valued in KM for codifying tacit knowledge, is subject to taxonomies. Peter Orton of IBM is quoted by Reamy (2002) stating that “One of the most important yet least appreciated facts about story is that perceivers tend to remember a story in terms of categories of information states as propositions, interpretations and summaries rather than remember the way the story is actually presented or its surface features” (Reamy, 2002, p. 1).

Ontologies and taxonomies help to: (1) share a common understanding of information; (2) reuse knowledge; (3) make assumptions about knowledge more explicit; (4) separate domain and operational knowledge; and (5) analyze domain knowledge (Noy and McGuinness, 2000). “Ontologies can be used as an instrument to make knowledge assets intelligently accessible to people in organizations” (Benjamins et al, 1999, p. 1). Ontologies are, however, expensive to develop and can be difficult to change once they are in place.

Due to their complexity and the need for their evolution within a community of scholars and practitioners, ontologies are “still far from being a commodity” (Angele and York, 2002). Ontologies are present in the category systems of websites such as Yahoo, Amazon, and Google (Benjamins et al, 1998; Leake et al, 2003; Noy and McGuinness, 2000).

Many disciplines have developed ontologies with standardized vocabularies, including medicine and the pharmaceutical industry. There is a taxonomy of non-profit, organizational entities for the Internal Revenue Service (NCCS, 2004). There are several taxonomies of businesses, including the North American Industry Classification System (NAICS), which was “developed jointly by the U.S., Canada, and Mexico to provide new comparability in statistics about business activity across North America” (U.S. Census Bureau, 2004, p. 1). A new North American Product Classification System (NAPCS) is also being developed and will focus first on service industries, with manufacturing products to be added in the future.

A standard ontology for news articles called the Applied Semantics’ News Series is widely used, based on the International press Telecommunications Council subject codes (Lamont, 2003). The Library of Congress and Dewey Decimal System classification schemes for bibliographic records are a taxonomy and coding system, as is the ERIC Thesaurus of Descriptors used to document educational materials.

II. Examples of Ontologies in Higher Education

There are a variety of applications for higher education for ontologies. These include:

- the marketplace of institutions
- academic disciplines
- the documentation of data
- metadata about learning management systems (LMS)
- the nature of the higher education enterprise.
- online resources, such as links and training materials

The marketplace of institutions

Designed to identify categories of colleges and universities, the Carnegie Classification of Institutions of Higher Education is the most widely recognized effort to document the higher education marketplace. First developed in 1971, The Carnegie Foundation for the Advancement of Teaching's classification scheme identifies categories that are "homogenous with respect to the functions of the institutions and characteristics of students and faculty members" (Shulman, 2001, p. vii). Classification reports have been published five times between 1971 and 2001 and a new structure and methodology will be released in 2005. This new edition will "provide a sophisticated, adaptive set of tools that allows users to cluster institutions in different ways" with users being given different "lenses through which to examine and analyze institution mission and other important differences among institutions" (p. viii).

Similar in stature if not in acceptance is the U.S. News and World Report's college rankings effort. Begun in 1983, the U.S. News rankings have evolved from being purely reputational in nature, driven by surveys of college presidents, to include elaborate formula with many complex variables. The tier structure and types of schools represent a category system and taxonomy comparable to the Carnegie classification. While there is great controversy about the methodology and results, U.S. News and other college admissions publishers such as the College Board and Peterson's have "become part of an integral movement – one that aims to provide the public with ever improving information about higher education" (Kleiner, 2004, p. 74).

Gumport (1997) and the National Center for Postsecondary Improvement (NCPI) have recognized that there is a "market for postsecondary education that can be readily described, even quantified" (Gumport, 1997, p. 23). However, "one design would not fit all." The NCPI research project found that "traditional categories for aggregating groups of institutions (size, Carnegie Classification, control) were unable to explain the real differences in student outcomes observed in key national data sets documenting educational attainment and labor market outcomes" (p. 24). As a result of this effort, a new taxonomy for the marketplace was created and "the idea and structure of the taxonomy resonated intuitively" with institutional leaders (National Center for Postsecondary Improvement, 2001, p. xx).

Yet, as Grasel (1999) explains in *The Reality of Brands: Toward an Ontology of Marketing*, "The ontology of marketing, particularly the question of what products and brands are, is still largely unexplored" (p. 1). This is especially true in higher education.

Academic disciplines

Disciplinary taxonomies have been in place for many years. From a national standpoint, these can be traced to efforts by the U.S. Department of Education's National Center for Education Statistics and the National Science Foundation's division of Science Resources Statistics. Student enrollment, degrees conferred, and research expenditure data by institution are collected at the discipline level by one or both agencies. A variety of sample survey data about students, graduates, faculty, and employees (especially science and engineering) are also collected from individual respondents, including information about field of study and field of occupation.

The NCES disciplinary data were originally collected using four-digit HEGIS (Higher Education General Information Survey) codes, part of the HEGIS data collection system used between 1966 and 1985. “These codes were updated into six-digit CIP codes” in 1985, with a “taxonomic coding scheme for secondary and postsecondary instructional programs.” CIP codes are “intended to facilitate the organization, collection, and reporting of program data using classifications that capture the majority of reportable data. The CIP is the accepted federal government statistical standard on instructional program classifications and is used in a variety of education information surveys and databases” (NCES, 2004a, p. 1).

NSF incorporates a three-digit classification scheme in its institutional surveys, one with a finer level of detail than previously offered by CIP codes for science and engineering disciplines, especially medicine. A slightly different three-digit scheme is used for the interagency-funded Survey of Earned Doctorates. NSF also provides a two-digit list of CASPAR discipline codes to which other disciplinary taxonomies may be rolled up for aggregate data. These are available as part of its WebCASPAR online data tool. NSF incorporates several crosswalks between different disciplinary taxonomies, including CIP, HEGIS, and occupation codes, in WebCASPAR.

Many other efforts to map the academic disciplines have been created and are in use today, including those of the National Research Council, the Council of Graduate Schools, Peterson’s, the College and University Personnel Association, and other organizations, agencies, and associations. Whether collecting data on enrollment, degrees, faculty, salaries, research expenditures, or equipment, each of these must include some form of disciplinary taxonomy. For their internal use, most colleges and universities map their departments to the reporting requirements of NCES using CIP codes. States regulate the approval of majors and programs with degree inventories that typically combine CIP code and award level.

Despite these many efforts, it requires extreme care to maintain the currency of these taxonomies. The nature of disciplinary work is becoming more fragmented and compartmentalized into specialized academic niches and fields of research. In order to understand supply and demand issues and the emergence of new knowledge and research, there must be a concerted dialogue about these taxonomies. Much of their evolution is due to mandated reporting requirements which, while sufficient for many purposes, fail to address cutting edge changes because they are often too gross in their level of detail. It is then up to disciplinary bodies to track the nature of their professions through refining data collections over time. Fortunately, web survey software is now more widely available and the result is that it is less expensive for disciplinary associations to conduct this needed work in this disciplinary ontology of higher education.

Documentation of data

In their efforts to promote best practices to improve data collection and reporting, several national organizations have developed standard taxonomies for higher education data. The National Center for Higher Education Management Systems (NCHEMS) has incorporated disciplinary and department data in its vision of resource allocation models since the 1970s, promoting the use of crosswalks between human resource, student, course, and finance data through organ-

izational mapping. It is impossible to conduct complex cost of instruction models without disciplinary taxonomies and departmental crosswalks and NCHEMS has spearheaded this effort for thirty years.

Crystal and Jones (1985) expanded the early NCHEMS work to focus on accreditation, with their monograph A Common Language for Postsecondary Accreditation: Categories and Definitions for Data Collection. One of the early NCHEMS products was its NCHEMS Data Element Dictionary (Thomas, 1971). A new version of Data Definitions for Colleges and Universities was released in a joint effort of NCHEMS and the Consortium for Higher Education Software Services (CHESS). CHESS was designed to foster agreement on terminology and definitions, standardize definitions, and help structure information architecture for an institution. It was released on CD-Rom in Microsoft Access.

CHESS includes several components – (1) data definitions; (2) MetaData Administrator software to maintain institutional files; and (3) the CHESS taxonomy. The taxonomy was first published in 1994 as the CHESS Taxonomy of Administrative Activities for Colleges and Universities and was updated with the second edition released in 2004. It provides “a comprehensive annotated list of academic and administrative activities at a typical college or university. It also provides a detailed guide for categorizing and describing the operations of colleges and universities and the activities that relate to information technology support” (Thomas, 2004, p. 2). There are five levels of hierarchy in the taxonomy, starting with major functional area.

Another national effort to promote effective data practices is that of the National Postsecondary Education Cooperative (NPEC). NPEC was authorized by Congress in 1994 to promote the quality, comparability and utility of postsecondary data and information that support policy development at the federal, state, and institution levels. NPEC receives funding from NCES and as part of its focus on “Quality Data Practices” has undertaken a variety of taxonomy-related projects. These include the work of an NPEC Working Group on “Best Practices for Data Collectors and Data Providers” that asked “What can be done to better coordinate data definitions and surveys on a national basis to achieve greater comparability and relieve institutional data burden?” (NPEC, 1999). Other related NPEC projects include an “Examination of the Data Requirements of the Workforce Investment Act and the Perkins Act of 1998,” a study of “Technology and Its Ramifications for Data Systems,” and an analysis of “Unit Record Versus Aggregate Data: Perspectives on Postsecondary Education Data Collection, Retention, and Release.”

Another long-term effort of NPEC is the ANSWERS (Accessing National Surveys with Electronic Research Sources) website, which includes a variety of online tools to help different types of users or audiences find the data and developer resources they need. At the heart of ANSWERS is a matrix of data dictionary information about almost 25,000 variables from over 110 datasets. Each of these data elements is content analyzed and coded using the unique ANSWERS taxonomy that was developed especially for this purpose, with over 340 subject/topic combinations. ANSWERS is no longer available online as part of the NPEC website, but is maintained by the developer at <http://higher.org/answers>.

Without a tool of this type, it is impossible to keep up with the availability of complex population and survey data about postsecondary education. ANSWERS also includes references

to key citations about developing surveys and using national datasets. It includes a Question Bank of questions used in sample surveys of faculty and students and a Definition Bank of standard glossary terms and definitions. Hard-to-find information about surveys is also included, with information such as average response rates and handling of missing data. With a combination of search and category-driven tools, ANSWERS is an important ontology resource for higher education.

Learning management systems (LMS)

Recognizing that information about learning must be shared between different computer systems, various groups have developed metadata standards about learning management systems (LMS). In the U.S., the National Learning Information Infrastructure and other standards boards are promoting standards for learning courseware. In Europe, ARIADNE and ELENA are examples of applications that support the exchange of knowledge resources.

An early EDUCAUSE article from 1997 spells out the vision and promise for metadata:

The primary purpose of metadata is to provide more helpful information about a work than can be obtained by inspecting the contents of the work, e.g., a Web page may be designed to teach mathematics skills to a third grade audience, but the terms "third grade" and "mathematics" may not appear in any of the text of the Web page. Therefore, traditional Web search engines, which often utilize full-text search indexes, would not return the page if "third grade and mathematics" were used as the search criteria.

Standards for metadata allow information and materials to be easily and consistently located. Unfortunately, where metadata solutions exist today, they are not consistent and are often proprietary. This has created an administrative nightmare for organizations that own or manage large collections of Web-based materials. It is these administrative challenges and the potential benefits to users that are driving the Internet industry to solve this metadata problem. The NLII IMS is building upon the industry's technology efforts by defining the necessary metadata elements to support widespread reuse, discovery and sharing of learning materials via the Internet (Griffin and Wason, 1997, p. 1).

Since 1997, the National Learning Information Infrastructure (NLII) project has made great progress in defining standards for metadata for instructional management systems (IMS). The NLII Annual Review for 2003 documents the current key themes, among them learning materials, software, and service markets; learning objects; and specifications/standards development (NLII, 2004). Metadata standards have now been developed and disseminated and are being implemented widely by developers and institutional IT staff for documenting courseware and learning objects.

New types of Peer-to-peer (P2P) software such as Edutella use the Universal Brokerage Platform to share learning objects across web servers. All of these developments are taking place because an infinitely growing array of learning objects are becoming available in all media types and modes of delivery. Organizational users won't be able to take advantage of them, however, unless there is a common taxonomy for documenting their availability.

Learning Object Metadata (LOM) are used to define a “Base schema that defines a hierarchy of data elements for learning objects metadata” (Ogbuji, 2003, p. 1). These metadata schema must incorporate many types and categories of information, including:

- General information
- Lifecycle – “features related to the history and current state of this learning object and those who have affected this learning object during its evolution”
- Meta-Metadata – “information about the metadata instance itself”
- Technical – technical requirements and characteristics
- Educational – “educational and pedagogic characteristics”
- Rights – “intellectual property rights and conditions of use”
- Relation – “features that define the relationship between the learning object and other related learning objects”
- Annotation – “comments on the educational use of the learning object,” including “when and by whom the comments were created”
- Classification – “describes this learning object in relation to a particular classification system” (Ogbuji, 2003, p. 1).

Wiley (2000) presents a taxonomy to “differentiate possible types of learning objects available for use in instructional design.” The “taxonomy's characteristics' values (such as High, Medium, and Low) are purposefully fuzzy, as the taxonomy is meant to facilitate inter-object comparison, and not to provide independent metrics for classifying learning objects out of context” (Wiley, p. ?).

Learning objects or knowledge chunks represent the most efficient focus of technology in teaching. By focusing on serving and finding learning objects, faculty no longer have to fight for scarce resources. The first steps have been taken with the development of standards for learning object metadata and instructional management systems. Now steps are being taken to build new types of learning object repositories. These require complex taxonomies and new types of search engines which combine the best features of searching and classification.

The higher education enterprise

Halstead's (1979) Higher Education Planning: A Bibliographic Handbook was published in the era when higher education administration was first becoming a professionalized field of study. This document, along with Higher Education: A Bibliographic Handbook Volume II (1981) helped to establish and map the knowledge base of the higher education enterprise, from admissions to space management to student affairs.

Twenty-five years later, the enterprise of higher education is being mapped in new ways, with great interest in virtual colleges and universities (VCUs). Epper and Garn (2004) cite the work of Wolf and Johnstone, whose “taxonomy classifies VCUs along a dimension of collaboration ranging from independence to highly distributed collaboration” (p. 34).

The CHES taxonomy documents the myriad departments and organizational units which exist in typical institutions. From financial information systems which must include a chart of accounts that is mapped into departments and units, to course and student data with alpha codes used to describe majors and academic departments, numerous taxonomies are implemented throughout the higher education enterprise.

The U.S. Department of Education's HEGIS surveys of financial data were the first attempt to categorize types of expenditures and revenues into an agreed-upon taxonomy. The newest IPEDS data categorize finance data in different ways, depending upon the implementation of new required forms from the Financial Accounting Standards Board (FASB) and the Governmental Accounting Standards Board (GASB). Unfortunately, as stated by the NCES documentation for the IPEDS Finance survey, "As data users attempt to compare institutions that cross accounting models, it becomes difficult to put them on the same scale. Some accounting differences cannot be adjusted for, but an understanding of them may help" (NCES, 2004b, p. 1).

The National Association of College Auxiliary Services (NACAS) developed its own taxonomy several years ago. The NACAS Data Bank includes almost 100 "operational categories of auxiliary services" ranging from amusement games to laundry to security.

Online resources

KM initiatives in higher education cover a breadth and depth of online applications, including portals, Intranets, data warehouses, data mining, environmental scanning, document management systems, digital dashboards, content management systems, customer relations management, and e-learning resources (Knowledge Integrity, 2000; Nylund, 2000; Survey Tracks, 2001). All of these systems require developers to make assumptions about how resources will be presented to the user. Some applications are customized and presented based on data from user-compiled profiles. Others are based on the audience, for example portals with categories of links geared to new students, parents, alumni, and the media.

As online resources are created and integrated into existing applications for admissions, registration, online courseware, and faculty advising, they all need to be categorized. Content management systems are used behind-the-scenes to manage the thousands of web pages and database structures necessary for a complex university or college setting. These systems must have a sophisticated and dynamic taxonomy. While much of the portal and administrative information system software delivers a foundation for these categories and ontology, they are only a starting point. For as the emergence of Google and the new breed of search algorithms points out, search techniques only go so far to providing relevant resources. Subject matter experts and others must be used to create taxonomies for the Web that make sense within the context of higher education and within the unique institutional setting. It is critical that developers recognize and document their assumptions about taxonomies when implementing portals, Intranets, and even basic websites for a department or unit. Simple questions such as "How are you going to categorize new information on the site?" can be very difficult to answer.

III. Issues in Creating Taxonomies

Search retrieval and content management

Bernbom writes that KM involves the “discovery and capture of knowledge, the filtering and arrangement of this knowledge, and the value derived from sharing and using this knowledge throughout the organization” (2001, p. xiv). While KM proponents share this goal, organizations are still overwhelmed by the need to “rapidly analyze and classify unstructured information.” This is the result of many forces, including staff retirements and turnover, budget cut-backs, an unqualified labor pool, lack of skills and/or training, and changing mission. However, problems of staffing and the continuity of knowledge within an organization become exacerbated because explicit information is often not readily available and is maintained out of context.

Recognizing that there is “too much information out of context,” it is difficult to “discern high-quality, relevant information from hearsay, inaccurate, unqualified, or outdated information” (Delphi Group, 2003, p. 2). It is also difficult to capture and communicate tacit information. At the heart of documenting knowledge assets is the work of content management. Content and process are “inextricably linked” and many KM proponents believe that “Content Management is all that matters” (Moore, 2003, p. S2). Whether stored in file cabinets, electronic filing systems, document management systems, Intranets, portals, or KM repositories, the critical issue is finding and using content.

Both search engines and ontologies have undergone a significant evolution in the past few years as users became inundated with millions of websites on the World Wide Web and expectations for relevant search results have grown with tools such as Google. With “exponential growth in the amount of data available across the globe,” search engines “return such large numbers of irrelevant results that frustration persistently triumphs” (Inxight, 2003, p. 3).

Major problems with search engines involve: (1) making results relevant; (2) ensuring secure and efficient collection of a breadth of data; (3) allowing various language methods for imputing information; and (4) the ability to scale a search product to very large indexes and volumes of queries (Andrews, 2003). Taxonomies are “often used in tandem with search and retrieval tools... However, unlike search technology alone, taxonomies reveal the overall structure of a knowledgebase in a hierarchy that is visible to the user” (Lamont, 2003, p. 1).

Creating ontologies

There are two activities involved in creating ontologies: (1) coding new documents into a beginning taxonomy; and (2) modifying the taxonomy to handle new types of information. These usually occur “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 the meaning in context” (Lamont, 2003, p. 2). Verity and other search engine tools incorporate auto-classification to generate rules for categories using sample documents. Ontologies have been a central discussion of the artificial intelligence community (Kalfoglou, 2000).

Noy and McGuinness (2000) explain that “The ontology should not contain all the possible information about the domain: you do not need to specialize (or generalize) more than you

need for your application (at most one extra level each way)” (p. 19). The authors break ontologies into information about classes, subclasses, slots, and instances. Most ontology discussion focuses on classes, which “describe concepts in the domain” or subject area.

For example, a class of wine represents all wines. Specific wines are instances of this class. The Bordeaux wine in the glass in front of you while you read this document is an instance of the class of Bordeaux wines. A class can have subclasses that represent concepts that are more specific than the superclass. For example, we can divide the class of all wines into sparkling and non-sparkling wines.

Slots describe properties of classes and instances: Chateau Lafite Rothschild Pauillac wine has a full body; it is produced by the Chateau Lafite Rothschild winery. We have two slots describing the wine in this example: the slot body with the value full and the slot maker with the value Chateau Lafite Rothschild winery. At the class level, we can say that instances of the class Wine will have slots describing their flavor, body, sugar level, the maker of the wine and so on” (Noy and McGuinness, 2000, p. 3).

The steps to developing an ontology therefore include: (1) defining classes; (2) arranging the class into subclasses within a hierarchy; (3) defining slots and the possible value labels for them; and (4) documenting specific instances using the slot value labels (Noy and McGuinness, 2000).

Numerous efforts are underway to develop a standard ontology for application on the Internet. The Suggested Upper Merged Ontology (SUMO) exists to help in building ontologies and includes over a thousand concepts that are “interconnected into [a] semantic network” with over 4,000 axioms (Ahrens and Huang, 2003; Niles and Pease, 2001; Sevcenki, 2003). An online tool called the SUMO Browser is available to help users navigate and use the SUMO.

IV. Software Solutions

Due to the complex nature of knowledge, ontologies can be constructed almost infinitely. Therefore, “Ontology harvesting must identify ontologies that are desirable to share, worth converting, and usable by others” (Kalfoglou, 2000, p. 54).

An interview with Dialog’s architect of content management, Steve Samler, reports that “Our key issues are keeping the taxonomy current and presenting information the way the user wants to see it... Part of our value-added in filtering stories is the judgment of our subject matter experts” (Lamont, 2003, p. 4). Subject matter experts (SME) need to be closely involved in creating taxonomies and concept maps, according to Venkata (2002). Experts “play an active role in the knowledge capture process,” explain Leake et al (2003).

Some software tools such as Convera’s Retrievalware are able to develop a taxonomy dynamically. This “combines searching with classification to produce dynamic classification” (Lamont, 2003). The Convera product literature states that “Users can launch and automatically classify the results based on pre-defined or dynamically generated classifications. The underlying taxonomies can consist of Convera’s pre-packaged industry taxonomies, customer defined

taxonomies or custom taxonomies” (Convera, 2004, p. 4). The benefit of this process is that “Rather than being forced to fit searches within the constraints of inflexible categories, users can create their own information categories based on the context of their search at the moment” (Convera, 2004, p. 5). The query results are presented as a hierarchy of classification folders.

Retrievalware software also offers searching by concept, pattern, and Boolean strings (and/or/not). The concept search “does what we naturally do in conversation with each other – account for the individual differences in the way we express similar ideas.” The pattern search uses a “sophisticated vote and rate scheme that considers a number of different features of the pattern instead of just character pairs.” The Boolean search incorporates “advanced linguistic analysis to ensure high precision and recall” (Convera, 2004, p. 3).

Other software such as The Taxis Categorizer “assigns documents to the categories in a taxonomy and automatically attaches subject codes and other metadata after being trained on sample documents.” There needs to be a “close and interactive relationship between categorization and search” (Lamont, 2003, p. 6)

The importance of pattern and concept mapping is illustrated in an example about effective electronic communications compliance in financial services. Compliance personnel need to “proactively identify patterns of suspicious activity. For example the phrases ‘IPO’ and ‘preferred customer’ appearing within separate but related documents may have little apparent connection to one another” (Delphi Group, 2002, p. 3). It is their proximity and relationship in the same document which help compliance staff find the next Martha Stewart policy breach.

This linguistic approach to mining the results of searches is part of a larger effort to create a “real Semantic Web” (Kasteren, 2003). The Inxight SmartDiscovery software tool is based on work done at Xerox and “automates the creation of structure on otherwise unstructured data sources by leveraging more than 20 years of research in natural language processing and data visualization techniques” (Inxight, 2004, p. 2).

Problems and issues in maintaining taxonomies

The greatest problem encountered with these automated processes is that taxonomies must be dynamic and changing. Venkata (2002) describes how developers must refine and enhance taxonomies by:

- *Adding new topics to capture changing relationships between informational resources being classified, reflecting new subject domains that the taxonomy must accommodate;*
- *Optimizing the taxonomy structure to more accurately reflect both the informational content as well as organizational requirements;*
- *Deleting and/or aggregating topics that are no longer of value;*
- *Increasing topic coherency and optimizing statistical training sets to maintain or enhance classification accuracy as content changes over time (Venkata, 2002, p. S13).*

Angele and Sure (2001) evaluate the limitations of software tools for ontology and document the many problems they must overcome. These include:

- language conformity, with standardized syntax
- consistency with respect to semantics
- interoperability for exchanging ontologies between tools
- turn around ability, so that users see it consistently over time
- performance through benchmark tests
- requirements for memory allocation in hardware to perform according to benchmarks
- scalability to more complex and larger taxonomies
- ease of integration into frameworks of other tools
- connectivity to other tools

Benjamins et al (1998, 1999) document similar problems. These include technological risks, including tool support, maintenance, and scaling up; and social and organizational risks, including the need for a minimum number of participants in creating a taxonomy, the climate of competitive mentality, and incentive systems. Collaborative thinking about ontology needs to be free of a competitive environment, according to the authors. An incentive system is necessary because “Given the high workload of today’s employees, it may be easily felt that contributing to a knowledge management effort is a waste of time, or at least does not have priority” (Benjamins et al, 1998, p. 17).

There are many more issues to consider in constructing ontologies and the reader is referred to the website of Sowa (2004), who discusses many issues, including:

- the relationships of process types
- distinctions in roles and relations
- causality
- agents
- thematic roles
- alignment of concepts, relations, and commonalities between ontologies
- properties, features, and attributes to differentiate categories
- hierarchy
- identity conditions

Concept Maps

Papadopoulos (2003) discusses how “The relations between terms help to describe the conceptual interactions between words or expressions and thus will directly impact precision and recall” (p. 9). There are five types of term relationships: (1) part-whole, such as bumper and automobile; (2) collocation, occurring frequently in a sentence; (3) paradigmatic relations such as sun and solar; (4) synonymic; and (5) antonymic. “The principal challenge lies in assessing the effect of these relationships on information retrieval results” Papadopoulos (2003, p. 9).

Milam et al (2000) and Carnot et al (2003) describe the use of concept maps in educational software applications. Milam et al explain that:

Even with clear assumptions and good qualitative research methodology, there are a myriad of ways to create a single type of map of the same content. It is important to either involve a group of scholars in developing a map and/or to recognize that the resulting map is simply a pattern for documenting the links between complex ideas. The groupware features for the collaborative creation of concept maps have great potential for developing these consensual maps and need to be explored further within an education context (Milam et al, 2001, p. 63).

Benjamins et al (1998) explain how ontologies can be represented visually through software such as Ontobroker, which displays a hyperbolic query interface. Clicking on the main node takes the user to related classes. “Ontology browsing” involves a visual representation of a taxonomy, based on the principles of hyperbolic geometry. “This visualization technique allows a quick navigation to classes far away from the center as well as a closer examination of classes and their vicinity. Classes can be dragged around while the size of the visualization nodes changes corresponding to their location, that is, the more centric the bigger they appear” (Benjamins et al, 1999, p. 7).

Leake et al (2003) explain that concept maps are similar to “vector-space models” in the way that knowledge is represented. Some systems are weighted, with higher weights given to top keywords. These systems can “consider the number of outgoing and incoming links to a concept node, strengthening the weights of keywords in nodes for concepts with many connections to other concepts in the map” (p. 5). In generating and suggesting new concepts using data mining techniques, the CMapTools software lets users “control how far the retrieval algorithm descends in the hierarchy tree to search for related concept maps...” (p. 5). A “keyword correlation metric” is used that measures the distance between concepts on a map. These and other automatic categorization techniques help present relevant topics to the user. There are limits to the results, however, and other scholars have created software such as EXTENDER (Extensive Topic Extender from New Data Exploring Relationships) that “identifies and suggests novel topics” (Leake et al, 2003).

Specific issues in taxonomies for higher education

In practice, the development of taxonomies for higher education is a much more imprecise process than is suggested by the previous discussion of dynamic classification software and semantics for the web. The following section addresses very specific issues in taxonomies related to the author’s experience with portals, campus-wide information systems, cataloging Internet resources, disciplinary taxonomies, and using national datasets.

Portal developers, whether open source or vendor-driven, have not evolved elaborate and complex categories for providing links to web resources. The department, college, and university staff who build websites often do not understand the principles of content management. Many of these sites are not database-driven, which provides a mechanism for standardization. This work is done piece-meal, in a haphazard fashion using student workers, and without much

foresight about managing site content and appearance over time. The website taxonomy, if there is one, is changed only when there is a site makeover, and often this process is compressed into a timeframe that leaves little time for reflection and analysis of site navigation problems, much less user input.

Almost all colleges and universities have a web presence. This is different from portals geared to different sets of user needs and applications and from Intranets with administrative information systems for operations. Campus-Wide Information Systems (CWIS) are the most visible symbol of an institution on the Internet and must represent many perspectives. There are inherent categories of content for a CWIS, from special audience pages to lists of academic departments to topical information such as directions, admissions, and student services. In the author's experience, the choice of categories is often a political process, geared not so much to usability and the needs of a majority of users as to the images and perception which the institution wants to promote as its public "face." In Milam's study of the "politics of websites," the conclusion emerged that CWIS are a form of sense-making, focused more on aspirational ideas than reality.

Another taxonomy developed by the author is the website Internet Resources for Institutional Research, which has been maintained since 1995 and includes thousands of links in numerous categories. When begun, there were few efforts to catalog the web by subject area and there were few links related to higher education, so the job was relatively easy. With a growing number of links, it became necessary to implement a rudimentary cataloging system. Hundreds of links were grouped by subject on single page. With interest in listservs, institutional fact books, and institutional research office websites, special pages were added and kept up separately. By 1997, an Access database was developed and ColdFusion software was used in a menu structure to document and serve over 100 pages by topic. Any link could be coded with multiple topics. Attempts were made to share the upkeep process with other volunteers, but the decision rules for coding and editing links were not made explicit. At some point, certain categories were hot, such as web database information; while others were little used. The site and work benefited when the ERIC Clearinghouse for Higher Education began linking to specific pages within the site.

The Internet Resources website should have deleted unnecessary categories and continued to evolve new ones, but the process was time-consuming. Briefly, the site was turned over to part-time staff of the Association for Institutional Research to maintain, but this was never fully implemented and the expectations of the author were not communicated clearly. The site is still maintained and new links are added and bad links removed. The evolution of the taxonomy is stalled due to the amount of effort needed to maintain it correctly. This requires the author's or other's subject matter expertise and a degree of discernment about relevance and interest to users. Any link changes or additions provided by users are made quickly, but the vision for Internet Resources has not been modified since the late 1990s. While it is highly used by institutional researchers and some faculty who teach higher education administration, it badly needs a makeover.

The work of NCHEMS, NCES, and NSF provides badly needed standards for data collection that implement de facto taxonomies. However, these are typically five or more years out

of date, owing to the development and approval processes involved. They are more often a map of where higher education has been, not where it is going. The recent release of the CIP Code 2000 disciplinary taxonomy is a case in point. By the time the 2000 CIP Codes were developed, reviewed, made available for comment, finalized, approved, and implemented into the software architecture of collections of student and degree data, it was already obsolete. There were thousands of changes, however, between CIP 1990 and CIP 2000 and the upgrade involved a massive effort. It is important to recognize the development of complex taxonomies as an evolutionary process.

In developing the taxonomy of subject and topic codes used for the NPEC ANSWERS project, the author found himself immersed in ontology issues. Relying on the principles of naturalistic inquiry, polychotomous coding categories were developed. These were not mutually exclusive, but allowed for multiple, redundant coding of variables into different combinations of subject and topic. It was recognized that hierarchical models make inherent value judgments about the best way to describe a piece of data. Therefore, every effort was made to categorize variables in as many ways as possible. The most difficult part of this work was the re-work needed after a new coding category was added. This required that all previous variables be analyzed to see whether they could also be coded under the new category. This is a routine part of content analysis, the constant comparative method, and ethnography. Special software was developed to import, export, pre-select, and anticipate coding categories based on previous variables.

Recently, approximately 5,000 variables from NCES sample surveys were added to the ANSWERS matrix. The cost and time involved in hand-coding each variable were tremendous. However, another taxonomy was already in place from other Data Analysis System (DAS) software which made these same variables available. Multiple attempts were made to crosswalk the ANSWERS and DAS taxonomies. While this worked for some variables, many others had to be recoded. All attempts to automate the process were inadequate, though they represented a valid starting point. The difficulty of modifying taxonomies and building relationships between taxonomies was made clear. This is another reason why great care must be taken to construct categories. Numerous subject matter experts were used during the creation of the ANSWERS taxonomy, due to the help of NPEC Working Group members, and this is essential for ontology.

In using national datasets such as IPEDS and the Survey of Earned Doctorates, researchers are reluctant to compromise on the level of detail they desire. For example, in developing a study of faculty supply and demand, the author brought together datasets about students, degrees, faculty, and employees to estimate how many doctoral recipients were interested in entering academe by discipline and how many entry level faculty positions were available to them. Data on retirements, rank mobility, and other factors were also included. Each dataset that contributed to the model had its own level of disciplinary taxonomy. But in building crosswalks between all of the different datasets, it was only possible to rely on either two-digit CASPAR or two-digit CIP Codes. Much meaningful data are lost this way. Only a relatively small number of two-digit U.S. Census occupation codes are provided to document postsecondary faculty employment. This meant that after many hundreds of hours spent learning about and implementing complex disciplinary taxonomies, data could be analyzed at only a superficial and gross level. It was impossible to create the model with comparable disciplinary taxonomies in each dataset.

Similarly, the author recently worked with a consulting firm and national association to help them use IPEDS data. The researchers wanted to provide historical trends of IPEDS financial data. The initial data suggested many unpredicted anomalies and outliers. What they researchers did not know was that the data categories themselves had underlying taxonomy issues. These included the changing nature of the data collection, which involved a move from paper to the web; implementation of new forms based on different accounting standards; cutbacks in the NCES budget for IPEDS for a specific year which resulted in decreased collection and editing of some data; and decisions made not to release a certain year of data in final format because it could not pass adjudication requirements. The categories and types of data over time appeared to be comparable. Only a sophisticated user of IPEDS trend data would be aware of these concerns.

V. Future Trends

This is an unprecedented era of what Gandel et al (2004) call “Information Abundance” for higher education. This requires new ways of thinking about “boundless information” and how institutional repositories can be rebuilt “from the bottom up.” Meta-tools for capturing metadata, especially ontologies, are badly needed.

In building taxonomies, it is critical that developers not try and reinvent the wheel. They need to understand existing efforts, where they succeed and where they are less useful than expected. Since ontologies are not yet a commodity and since they cost so much to develop, these efforts at KM must be valued. An organizational culture must be prized in which KM is rewarded, especially those hidden and less glitzy projects such as building a taxonomy.

In order for taxonomies to be fluid and changing to meet many different sets of needs, developers must incorporate the principles of dynamic classification. Exciting new software such as Convera and Inxight is now available to combine the best of taxonomies and search engines.

Developers need to incorporate concept maps, pattern recognition, Boolean logic, subject matter experts. They need to understand the natural problems which occur in creating and interfacing between ontologies. While the CHESS/NCHEMS taxonomy of data structures is very impressive and the metadata efforts promoted by NLII for IMS are essential and remain at the forefront of current thinking about learning and technology, these are very expensive to develop and maintain. These initiatives need to continue to evolve and require a substantive commitment of resources and vision.

While sometimes deemed too costly, ontology is shown in this chapter to be central to KM strategies in higher education for capturing and utilizing knowledge assets. The case still needs to be made for the return on investment (ROI) of KM for higher education. This is very critical for content management through taxonomies. Work on ontologies must be given the resources and attention it deserves if content management is going to succeed in helping institutions handle the onslaught of information which is overload existing systems and personnel. The loss of critical knowledge assets with employee turnover and retirement must be stemmed

through capturing and leveraging knowledge. This is only possible through the use of dynamic classification.

The development of ontologies for higher education is still a nascent field. There is much exciting, groundbreaking research to be done. It is important that institutional leaders and policymakers recognize the value that ontology holds for future work in KM and make the necessary investments now. This starts with a shared vision of what ontology offers to higher education and how taxonomies are interwoven throughout administrative information systems and all web-based learning applications.

References

- Ahrens, K, Chung S.F., and Huang C. 2003. Conceptual Metaphors: Ontology-based Representation and Corpora Driven Mapping Principles. In *Proceedings of the ACL Workshop on the Lexicon and Figurative Language*. Accessed 10/26/04. Available online at: <http://acl.ldc.upenn.edu/acl2003/lexfig/pdf/Ahrens.pdf>
- Angele, Jurgen and York, Sure. (2002). "Whitepaper: Evaluation of Ontology-based Tools." Workshop presentation at the 13th International Conference on Knowledge Engineering and Knowledge Management. Accessed 10/26/04. Available online at: http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/eon2002_whitepaper.pdf
- Benjamins, V. Richard et al. (1998). "Knowledge Management through Ontologies." Proceedings of the 2nd International Conference on Practical Aspects of Knowledge Management. Accessed 10/26/04. Available online at: <http://citeseer.ist.psu.edu/benjamins98knowledge.html>
- Benjamins, V. Richard et al. (1999). "(KA)²: Building Ontologies for the Internet: a Mid Term Report." Accessed 10/26/04. Available online at: <http://citeseer.ist.psu.edu/276747.html>
- Bernbom, Gerald, editor. (2001). Information Alchemy: The Art and Science of Knowledge Management. EDUCAUSE Leadership Series #3. San Francisco: Jossey-Bass.
- Carnot, M. J. et al. (2003). "A Summary of Literature Pertaining to the Use of Concept Mapping Techniques and Technologies for Education and Performance Support." Technical Report submitted to the Chief of Naval Education and Training, Pensacola, FL. Accessed 10/26/04. Available online at: <http://www.ihmc.us/users/acanas/Publications/ConceptMapLitReview/IHMC%20Literature%20Review%20on%20Concept%20Mapping.pdf>
- Crystal, Melodie E. and Dennis P. Jones (1985). A Common Language for Postsecondary Accreditation: Categories and Definitions for Data Collection. Boulder: National Center for Higher Education Management Systems.
- Convera. (2004a). "Mission-Critical Search & Categorization for the Enterprise." Accessed 10/26/04. Available online at: http://www.ihssolutions.com/canada/documentation_library/index.cfm
- Convera. (2004b). "RetrievalWare's advanced Categorization and Dynamic Classification." Accessed 10/26/04. Available online at: http://www.convera.com/Products/rw_categorization.asp
- Inxight. (2003). "Inxight SmartDiscovery: Discover the True Value of Information." Sunnyvale, CA: Inxight Software, Inc.
- Delphi Group. (2002). "Enabling Electronic Communications Compliance." *Snapshot*. Boston: Delphi Group.
- Delphi Group. (2003). "Maximizing Organizational 'Know How' in Government Entities." *Snapshot*. Boston: Delphi Group.

- Epper, Rhonda M. and Myk Garn. (2004). "Virtual Universities Real Possibilities." *EDUCAUSE Review*, vol. 39, no. 2 (March/April 2004).
- Gandel, Paul B., Richard N. Katz, and Susan E. Metros. (2004). "The Weariness of the Flesh: Reflections on the Life of the Mind in an Era of Abundance." *EDUCAUSE Review*, vol. 39, no. 2 (March/April 2004): 40–51.
- Grasel, Wolfgang. (1999). "The Reality of Brands: Toward an Ontology of Marketing." *American Journal of Economics and Sociology* 58. Accessed 10/26/04. Available online at: <http://ontology.buffalo.edu/brands.html>
- Griffin and Wason. (1997). "The Year of Metadata." *Educom Review* 32(6). Accessed 10/26/04. Available online at: <http://www.educause.edu/LibraryDetailPage/666&ID=ERM9763>
- Gumport, Patricia J. (1997). "In Search of Strategic Perspective: A Tool for Mapping the Market in Postsecondary Education." *Change* November/December, 1997.
- Halstead, D. Kent. (1979) Higher Education Planning: A Bibliographic Handbook. Washington, D.C.: U.S. Department of Education, National Institute of Education.
- Halstead, D. Kent. (1981) Higher Education: A Bibliographic Handbook Volume II. Washington, D.C.: U.S. Department of Education, National Institute of Education.
- Inxight. (2003). "Inxight SmartDiscovery: The Complete Solution for Enterprise Information Discovery." Sunnyvale, CA. Accessed 10/26/04. Available online at: <http://www.inxight.com/products/smardiscovery/>
- Kalfoglou, Yannis. (2002). "Maintaining ontologies with organisational memories." Accessed 10/26/04. Available online at: <http://www.ecs.soton.ac.uk/~yk1/kalfoglou-kluwerKMOMbook.pdf>
- Kasteren, Joost van. (2003). "Semantic Web Should Be Based On Well-Founded Ontologies: An Interview with Nicola Guarino." *DigiCULT: Towards A Semantic Web for Heritage Resources*. Thematic Issue 3, May 2003.
- Kleiner, Carolyn. (2004). "Decades of Rankings." America's Best Colleges 2004 Edition. Washington, D.C.: U.S. News and World Report.
- Knowledge Integrity, Inc. (2000). "Collecting Quality Customer Data." *Knowledge Management*, (3):2. pp. 78-80. Accessed 10/26/04. Available online at: http://www.destinationcrm.com/km/dcrm_km_article.asp?id=226.
- Lamont, Judith. (2003). "Dynamic taxonomies: keeping up with changing content." *KM World* 12(5). Accessed 10/26/04. Available online at: http://www.kmworld.com/publications/magazine/index.cfm?action=readarticle&Article_ID=1508&Publication_ID=90
- Leake, David B., et al. (2003). "Aiding knowledge capture by searching for extensions of knowledge models." Proceedings of the International Conference on Knowledge Capture. Accessed 10/26/04. Available online at: <http://portal.acm.org/citation.cfm?id=945655&dl=ACM&coll=GUIDE>
- Milam, John H., et al. (2000). "Concept Maps for Web-Based Applications: ERIC Technical Report." Accessed 10/26/04. Available online at: <http://highered.org/docs/milam-conceptmaps.PDF>
- Moore, Andy. (2003). "The Next Big Thing... Again." *KM World* supplement. April 2003. Accessed 10/26/04. Available online at: <http://www.kmworld.com/publications/whitepapers/ECM03/moore.pdf>
- National Center for Postsecondary Improvement. (2001). "Resurveying the Terrain: Refining the Taxonomy for the Postsecondary Market." *Change* April/May 2001.

- NCPI. (1998). "The User-Friendly Terrain: Defining the Market Taxonomy for Two-Year Institutions." *Change* January/February, 1998.
- NCCS. (2004). "National Taxonomy of Exempt Entities – Core Codes." Accessed 10/26/04. Available online at: http://nccs2.urban.org/ntee-cc/irs_code.htm
- NCES (2004a). IPEDS Glossary. Accessed 10/26/04. Available online at: <http://nces.ed.gov/ipeds/glossary/>
- NCES. (2004b). "IPEDS Finance Data FASB and GASB - What's the Difference?" Accessed 10/26/04. Available online at: <http://nces.ed.gov/ipeds/web2000/gasbfasb.asp>
- NLII (2004). 2003 NLII Annual Review. Washington, D.C.: National Learning Infrastructure Initiative. Accessed 10/26/04. Available online at <http://www.educause.edu/nlii/>
- NPEC. (1999). "Best Practices for Data Collectors and Data Providers." Washington, D.C.: National Postsecondary Education Cooperative. Accessed 10/26/04. Available online at: <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=1999191>
- Niles, Ian and Adam Pease. (2001). "Towards a Standard Upper Ontology." Paper presented at FOIS conference, October 2001.
- Noy, Natalya F. and Deborah L. McGuinness. (2000). "Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001. Accessed 10/26/04. Available online at: <http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html>
- Nylund, Andrea L. (2000). "Finding Patterns in a Deluge of Data." *Knowledge Management*, (3):2. pp. 69-71. February, 2000. Accessed 10/26/04. Available online at: http://www.destinationcrm.com/km/dcrm_km_article.asp?id=189.
- Ogbuji. (2003). "XML knowledge management flourishes in learning technology initiatives." Accessed 10/26/04. Available online at: <http://www-106.ibm.com/developerworks/xml/library/x-think21.html>
- Papadopoulos, Alkis. (2003). "Meaningful Search: Why PET Scanners are not about Cats & Dogs." Carlsbad, CA: Convera.
- Reamy, Tom. (2002a). "Imparting knowledge through storytelling, Part 1." *Mold* 11(6). Accessed 10/26/04. Available online at: http://www.kmworld.com/publications/magazine/index.cfm?action=readarticle&Article_ID=1306&Publication_ID=73
- Reamy, Tom. (2002b). "Imparting knowledge through storytelling, Part 2." *KMWorld* 11(7). Accessed 10/26/04. Available online at: http://www.kmworld.com/publications/magazine/index.cfm?action=readarticle&Article_ID=1328&Publication_ID=74
- Shulman, Lee S. (2001). The Carnegie Classification of Institutions of Higher Education. Menlo Park, CA: The Carnegie Foundation for the Advancement of Teaching.
- Sevcenko, M. (2003). "Online Presentation of an Upper Ontology." *Proceedings of Znalosti 2003*, Ostrava, Czech Republic, February 19-21, 2003. Accessed 10/26/04. Available online at: <http://ontology.teknowledge.com/Sevcenko.pdf>
- Sowa, John F. (2004a). "Ontology." Accessed 10/26/04. Available online at: <http://www.jfsowa.com/ontology/>

Sowa, John F. 2004b). "Ontology, Metadata, and Semiotics." Accessed 10/26/04. Available at: <http://users.bestweb.net/~sowa/peirce/ontometa.htm>

"Survey Tracks Trends in E-Learning." (2001). *Knowledge Management* (4):1. p. 11. January, 2001. Accessed 10/26/04. Available online at: http://www.destinationcrm.com/km/dcrm_km_article.asp?id=641

Thomas, Charles R. (1971). NCHEMS Data Element Dictionary. Boulder: National Center for Higher Education Management Systems.

Thomas, Charles R. (2004). CHESS Data Definitions - Second Edition. Boulder: National Center for Higher Education Management Systems.

U.S. Census Bureau. (2004). "North American Industry Classification System (NAICS)". Accessed 10/26/04. Available online at: <http://www.census.gov/epcd/www/naics.html>

Venkata, Ramana. (2002). "Taxonomies, Categorization, and Organizational Agility." *KM World* supplement. October, 2002. Accessed 10/26/04. Available online at: <http://www.kmworld.com/publications/whitepapers/KM2/venkata.pdf>

Wiley, D. A. (2000). "Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy." Accessed 10/26/04. Available online at: <http://www.reusability.org/read/chapters/wiley.doc>