Dagobert Soergel
College of Library and Information Services
University of Maryland
Office: (301) 405-2037 Fax (301) 314-9145
Home: (703) 823-2840 Home Fax (703) 823-6427
ds52@umail.umd.edu
www.clis.umd.edu/faculty/soergel/

The rise of ontologies or the reinvention of classification

Journal of the American Society for Information Science. October 1999; 50(12): 1119-1120.

Abstract

Classifications/ontologies, thesauri, and dictionaries serve many functions, which are summarized in this note. As a result of this multiplicity of functions, classifications – often called ontologies – are developed in many communities of research and practice. Unfortunately, there is little communication and mutual learning; thus, efforts are fragmented, resulting in considerable reinvention and less than optimal products.

Classification serves many functions and thus is claimed by many fields, but the communication among these fields is poor, leading to an approach that is marked by fragmented and costly reinvention.

Ontological and lexical structures are the underpinning of scientific and scholarly work, of learning, and of machine intelligence. They serve many critical functions in thinking and in communicating, organizing, and retrieving information by people and machines. The functions of tools providing such structures (dictionaries, thesauri, ontologies/classifications) include the following:

- *Provide a semantic road map to individual fields and the relationships among fields,* thus providing orientation and serving as a reference tool. This includes the following specific functions: *relate concepts to terms and provide definitions*; clarify concepts by putting them in the context of a classification/ontology; relate concepts and terms or icons across disciplines, languages, and cultures.
- *Improve communication and learning:* Assist writers and readers; support learning through providing conceptual frameworks and challenging students to produce such frameworks; support language learning; and support the development of instructional materials.
- *Provide the conceptual basis for the design of good research and implementation*: assist researchers and practitioners in *exploring the conceptual context* of a research project, policy, plan, or implementation project and in *structuring the problem*; support consistent definition of variables/measures for more comparable and *cumulative research* and evaluation results.
- *Provide classification for action*: a classification of diseases for diagnosis, of medical procedures for billing, of staff skills for task assignments, of commodities for customs.

- *Support information retrieval: provide knowledge-based support of end-user searching* (menu trees, guided facet analysis of a search topic, browsing a hierarchy or concept map to identify search concepts, mapping from the user's query terms to descriptors used in one or more databases or to the multiple natural language expressions for free-text searching); support *hierarchically expanded searching*; support *well-structured displays of search results*; provide a *tool for indexing* (vocabulary control, user-centered or problem-oriented indexing).
- *Provide the conceptual basis for knowledge-based systems.*
- *Provide the conceptual basis for data element definition and object hierarchies in software systems*.
- *Do all this across disciplines, languages, and cultures*.
- Serve as *mono-, bi-, or multilingual dictionary for human use* and as *dictionary/knowledge base for natural language processing* — machine translation and natural language understanding for data extraction and automatic abstracting/indexing.

Classification has long been used in library and information systems to provide guidance to the user in clarifying her information need and to structure search results for browsing, functions largely ignored by the text retrieval community but now receiving increasing attention in the context of helping users to cope with the vast amount of information on the Web. Fairly recently, other fields, such as AI, natural language processing, and software engineering, have discovered the need for classification, leading to the rise of what these fields call ontologies.

The Oxford English Dictionary defines *ontology* as "The science or study of being; that department of metaphysics which relates to the being or essence of things, or to being in the abstract." Part of such a study is a classification of things that are into basic types, often starting with *living vs non-living entities*. Thus the term *ontology* assumed the additional meaning of a shallow classification of basic categories. Such classifications or ontologies are needed in linguistics, for example, to formulate rules of the subjects or objects a verb can take, and in data element definition. As such rules became more and more refined, the classification supporting them needed to be more specific, so eventually *ontology* was used to designate any classification, particularly in the communities of linguistics, AI, and software engineering. Indeed, once these communities increased their awareness that there is not only a problem of classification but also of terminology, "ontologies" included lead-in vocabularies as well and became full-fledged thesauri.

But a classification by any other name is still a classification. The use of a different term is symptomatic of the lack of communication between scientific communities. The vast body of knowledge on classification structure and on ways to display classifications developed around library classification and in information science more generally, and the huge intellectual capital embodied in many classification schemes thesauri is largely ignored. Large and useful systems are being built with more effort than necessary. Examples are the CYC ontology (www.cyc.com/cyc-2-1/intro-public.html), whose presentation could be vastly improved, or WordNet (www.cogsci.princeton.edu/~wn or www.notredame.ac.jp/cgi-bin/wn.cgi), a wonderful system whose construction would have profited from applying experience in thesaurus construction and whose synset (concept) hierarchy should be made more easily accessible using standard methods for classification display. Another example is the ANSI Ad Hoc Group on

Ontology Standards (http://www-ksl.stanford.edu/onto-std/index.html), which does not seem to have any information scientist concerned with classification among its members.

There are many types of knowledge bases on concepts and terminology: classification schemes and thesauri, dictionaries and ontologies developed for AI applications, linguistic systems, or data element definition. These different types of knowledge bases — though developed for different purposes — overlap greatly and they follow very similar principles and methods for their construction. Better communication among the various communities involved in these systems could lead to an integrated common access system that would support all the functions discussed above (Soergel 1996).

Soergel, Dagobert. 1996. SemWeb: Proposal for an open, multifunctional multilingual system for integrated access to knowledge base about concepts and terminology. Proceedings of the Fourth International ISKO Conference, 15-18 July 1996, Washington, D.C. Frankfurt/Main: Indeks Verlag: 1996. (Advances in Knowledge Organization, v.5, 165-173). (See http://www.clis.umd.edu/faculty/soergel/semwebab.htm for a fuller exposition of the idea.)