Work Coding: Beyond SIC and SOC, BOC and DOT

Philip Harber, MD, MPH; Giedra Miller, BS; and Justine Smitherman, MS

Traditional coding schemes for occupational health related factors (eg, SIC, SOC, CAS) are structured similarly as hierarchical, exclusive, and exhaustive systems. Each is limited to information of one particular type (eg, industry) and, therefore, the relationships implicit in the coding scheme are limited. This study empirically determined that "natural" coding schemes used by occupational health professionals do not share these characteristics but are more akin to semantic network data bases. There is therefore a need to re-evaluate how occupational data are to be coded.

Coding information about an employee's work is necessary for three major purposes. First, it permits the *succinct summarization* of exposures of the worker for clinical assessment purposes, epidemiologic studies, worker surveillance activities, and for establishing descriptive data about work. Second, it permits *aggregation of data* into general classes (eg, to compare health outcomes of groups). Third, it often facilitates *inference* about similarity of jobs by implying that items with similar codes are similar. Thus, in addition to simply substituting numeric codes, schemes actually imply relationships. Attention to the method of coding may improve its function. Although free text prose may be useful in describing the work of an employee for clinical purposes, when used for epidemiologic and aggregate analysis purposes, such an approach is untenable. Therefore, several commonly used schemes have been developed for coding information about work.

From the Occupational Medicine Branch, Department of Medicine, University of California, Los Angeles (Dr Harber, Associate Professor of Medicine; Giedra Miller, Research Staff Member; Justine Smitherman, Staff Research Associate).

Address correspondence to Philip Harber, MD, MPH, Occupational Medicine Branch, Department of Medicine, UCLA, Los Angeles, CA 90024-1690.

0096-1736/91/3312-1274$03.00/0

The major coding schemes with direct relevance to occupational health are Standard Occupational Classification (SOC),[1] Standard Industrial Classification (SIC),[2] Chemical Abstract Service (CAS),[3] International Classification of Diseases (ICD),[4] Bureau of Census (BOC),[5] and Dictionary of Occupational Titles (DOT).[6] Several other special-purpose coding schemes are available.[7–11]

The traditional coding systems currently in use share several general characteristics (Table 1):

1. They have been developed for purposes other than occupational health. In general, those describing jobs or industries (eg, SOC, SIC) have been systematized for economic reasons, linking industries that are closely related in the economy although they are not related from the health standpoint. For exposure agents, coding often is based on the CAS numbers, which are aggregated based on chemical structure, not health effect.

2. The information encoded in the codebooks is fully ordered: the listing of numbers and corresponding entities is based on a predetermined sequence. Therefore, addition of a new entity (for example, new agent) might require redesignation of a large group of terms; a new term cannot simply be appended.

3. The systems are *hierarchical*. Thus, the entire universe of possibilities is broken into a series of ordered classes. Within each class, there are subclasses, and within each subclass are further subdivisions. The system implies that information about a more general class should apply to each of its members. For example, in the SIC system, the general class of manufacturing includes several major groups, including Food and Kindred Products, Furniture and Fixtures, and Primary Metal, and each is further broken into subunits.

4. Numeric proximity implies similarity. Codes within a specified range are lumped together as homogeneous, and conversely, codes that are distant are assumed to represent dissimilar entities. It is common in epidemiologic studies to aggregate on the first or first two digits

TABLE 1
Characteristics of Traditional Coding Schemes

1. Schemes not developed for occupational health purposes
2. Fully ordered
3. Hierarchical
4. Numeric proximity implies similarity
5. Exclusive location for each entity
6. Exhaustive
7. Strongly typed
8. Single code for a worker at any one time
9. Virtual knowledge limited in extent

of a five-digit SIC code although the entities represented may not in reality have the same health implications.

5. The systems imply an *exclusive location* for each entity. Thus, a particular job title should be assigned a single numeric code. Because of the hierarchical system, this requires that it occupy only one place in the scheme, thereby limiting the information available by inference from proximity.

6. The system is *exhaustive*. Every entity in the universe must have a location within the coding scheme. This constraint, in conjunction with the fully ordered characteristic (2) makes additions very difficult. Hence, the schemes are static.

7. The coding systems are *strongly typed* (to borrow a computer term): the coding schemes are specific for the type (class) of entity. There are independent coding schemes for job titles (eg, SOC), industry (eg, SIC), chemical agent (eg, CAS). Thus, the relational information implied by the sequence and hierarchy can only imply relationships *within* a specific type, rather than allowing any information about cross-type relations (eg, chemical agents likely to be present in a particular industry).

8. An additional, implicit implication of these systems is that at any time, a worker is assigned to one and only one code for a particular type (that is, one job code, one industry code).

Coding schemes are useful because they allow formal statistical and aggregate analyses. However, they inherently lead to loss of some information content; (For example, there is more information in the statement, "he worked as a TIG or oxyacetylene welder outdoors. . .," than in the corresponding coded statement, "SOC Code 7714.") Optimally, the information loss should be minimized. One source of information loss is the mapping from a large number of possibilities onto a smaller number of categories. For example, if all job titles were collapsed into 10 classes by first digit of SOC, much information would be lost (for example, jewelers (6822) and pipe organ builders (6839) have the same first two digits in the SOC, but obviously are quite distinct).

In addition to loss of information content due to the many-to-one mapping, characteristic 5 of Table 1 implies loss. For example, the job title "agricultural biologist" might fit into an "agriculture" class or into a "science" class, however, because it is allowed only one location, its existence in the class from which it is excluded is lost.

Coding schemes include both direct knowledge and "virtual" knowledge. Virtual knowledge is that which is not explicitly coded but may be deduced from information inherent within the scheme. For example, the hierarchical nature of SOC implies that a beet farmer (code 5513, which includes crop, vegetable, fruit, nut and tree farmers) is a member of the more general class of code 55 (farm operators and managers), which includes codes 5512 (general farmers) to 5525 (horticultural; specialty farm managers).

In addition to virtual knowledge inferred from the hierarchical nature of many schemes, additional virtual knowledge derives from proximity (that is, entities with adjacent codes are similar). For the coding schemes currently in use, virtual knowledge is limited to these two types.

This study focuses on the structure and function of occupational coding schemes. It compares them with the manner in which health professionals implicitly organize their knowledge (their "natural coding schemes"). We conclude that the currently used schemes perform relatively poorly in comparison and that attention to their basic design is warranted.

Methods

Two methods were used to determine empirically the manner in which work related information is encoded "naturally," thereby providing comparison with the "traditional" coding schemes. The two methods used were collection of empiric phrase associations and analysis of the written information contained in the NIOSHTIC abstract database.[12] The major benefit of coding is providing a formal means for expression of relationships among entities. Both methods of this study permitted determination of the types of relationships that exist in occupational health.

Empiric phrase association (EPA) was evaluated by giving a brief questionnaire to a group of 30 pulmonary and occupational health professionals. The questionnaire was distributed without explanation of the underlying purpose. The questionnaire listed three object terms ("wood, welding, and wheeze") and asked subjects to list up to 10 words they associated with each of these terms when "thinking about lung problems." Such words are referred to as "empiric phrase associations." Subjects also characterized themselves on the questionnaire along two visual analog scales: clinical versus nonclinical and occupational versus pulmonary. Some questionnaires, coded to reveal the discipline of the responding subject, were sent directly to individuals selected via directories; others were distributed to two rows of participants at a regional occupational medicine meeting. All subjects were occupational or pulmonary physicians, or industrial hygienists. The group must be considered a "convenience sample" because it was not randomly selected.

The NIOSHTIC data were collected by using a generally available CD-ROM based occupational health abstract collection[12] produced by the National Institute

for Occupational Safety and Health (NIOSH). The CCINFO version allows searching the CD-ROM disc for occurrences of a particular term. "Wood," "wheeze," and "welding" were used as search words, and the first 35 abstracts found for each were then examined. Each occurrence of the search word in the abstract (excluding its use as part of an introductory identifier) was located, and the nearest phase was recorded as an associated phrase. This permitted the development of a list of associated phrases for each search word. (Articles and prepositions were excluded.)

Several evaluative methods were used. All object terms and empiric phrases were assigned a type (for example, "wheeze" is a symptom, "asthma" a disease) to permit the analysis of typing relationships in empiric associations. The presence of virtual knowledge also was evaluated. Using outside occupational health knowledge, the investigative team developed a series of possible causal chains to explain why certain terms might be included. The relationships between the object terms and their empiric associations, and the key words and the NIOSHTIC-associated phrases were examined to determine how information is naturally related.

The number and types of phrases generated by each empiric study participant were tallied. The self-stated orientations (clinical versus nonclinical, pulmonary versus occupational) and primary disciplines (occupational and pulmonary physicians, industrial hygienists) were evaluated for their influence on the types of relationships seen.

Results

Analysis of the data shows several differences between traditional coding methods and actual knowledge.

1. Empiric phrase associations and NIOSHTIC data are *not* strongly typed. Table 2 illustrates the distribution of subjects' phrase associations by type, with examples of words from each type. (Type is the general category of a term, for example, occupation, agent, disease). Agents and diseases were the most commonly mentioned types of terms. Symptoms, clinical findings, and descriptive terms also were often mentioned; no other type of term was mentioned more than 20 times. Table 2 shows that information organization is not strongly typed. Although traditional coding schemes nearly exclusively interrelate terms of the same general type (eg, occupation, subclass of occupation), natural structures include extensive between-type interrelationships. The empiric phrase associations demonstrated that 46% of terms empirically associated with "wood," a term whose type = agent, were of other types. Similarly, 50% of the phrases associated with "welding," a job/industry, were agent, and 25% were diseases. Fewer than 1% of words mentioned in relation to welding were types of welding or industries with which welding is associated. Figure 1 illustrates the breadth of types associated with welding. They have been placed in regions corresponding to different entity types. Quite comparable results were seen with the associations found in the NIOSHTIC data base. Forty percent of words associated with welding were agents, with only 5% related to types of welding and welding industries. Thus, it is clear that information is naturally organized with considerable cross-references rather than being strongly typed.

2. Empiric phrase associations and NIOSHTIC data include more than just parent-child relations. Another disparity between natural knowledge coding and systems commonly in use is the multiplicity of relations, even within a type. Traditional coding schemes tend to be organized hierarchically, only using a series of parent-child (set-subset) relationships. In both the empiric phrase association and NIOSHTIC data, however, items can have many different types of relations, even within the same general type. This is illustrated by Figure 1, which shows a possible grouping of terms associated with welding. Even when viewing only the phrases of type = agents, relations to associated agents (eg, flux),

TABLE 2
Term Distribution by Type for Empiric Phrase Association and NIOSHTIC Data Base Phrase Associations

	Empiric Phrase Associations			NIOSHTIC Phrase Associations			Example of Type for Term "Wheeze"*
	Welding	Wheeze	Wood	Welding	Wheeze	Wood	
Agent	100	41	99	8	7	10	Dust
Clinical finding	1	28	1	0	0	0	Bronchoconstriction
Descriptive	16	9	7	8	0	0	Musical
Disease	49	87	39	1	1	0	Asthma
Job/industry	3	2	21	1	6	4	Spray painting
Lifestyle	2	1	0	0	0	0	Smoker
Product	0	0	1	0	0	4	Floor (wood)
Protection	15	0	3	0	0	0	Dust mask (wood)
Result	1	0	0	0	0	0	Thermal decomposition (welding)
Route	5	2	3	0	0	0	Inhalation
Symptom	6	26	7	0	3	0	Shortness of breath
Target	0	7	1	0	0	0	Lung
Test	0	7	0	0	0	0	Spirometry
Time course	1	0	1	2	0	0	Years' exposure (welding)
Treatment	0	19	0	0	0	0	Bronchodilator

* Examples are given for each type for the term "wheeze" except when no examples for wheeze existed; in this case another example was given and the key term is given in parentheses.

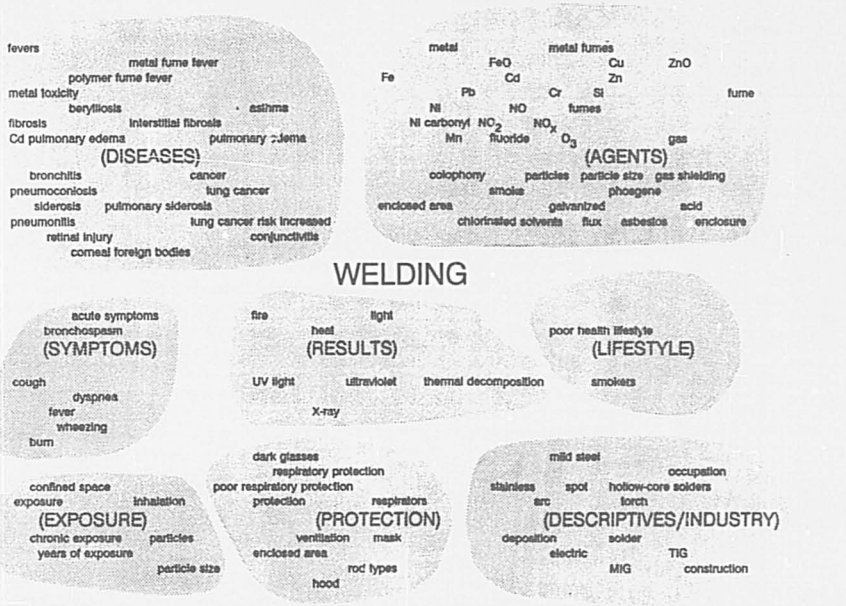

Fig. 1. Terms empirically associated with "welding." These terms have been grouped according to type for purposes of demonstrating that words associated with welding are frequently of different types than the term "welding" itself.

agents released (eg, heat, light, UV light), and agents operated upon (eg, mild steel) are seen. Thus, parent-child relations are actually only a small part of those that exist. There is much more information incorporated in the natural relations descriptors than in a simple coding scheme. Similarly, "wood" (an agent or product, by general object type) was associated with some items that might be considered subclasses (eg, redwood dust), but also was empirically associated with a large number of agent-product type items that would not have been related to it in traditional monotonic single-relationship systems. For instance, "wood preservatives," "plicatic acid," and others would not have been related by traditional coding schemes (that is, these are neither subsets nor a superclass of the word). However, plicatic acid is an agent found in wood dusts that causes respiratory sensitivity, so it is reasonable to include plicatic acid as words related to wood.

Similar findings showing a multiplicity of types of relations, even within object type, are again illustrated by the clinical symptom, "wheeze." First, there are several associated clinical terms that are neither subsets nor supersets of the term but are very appropriately related to it; for example, airway reactivity is appropriately incorporated as an associated clinical finding, but would not be linked in a simple parent-child system. Furthermore, there may be multiple directions associated within object-type relations. Terms such as "airway reactivity" and "asthma" show one type of relation, illustrating disease that may be associated with "wheeze," although "dyspnea" and "difficulty breathing" are equally appropriately related but have an inherently different relationship, exemplifying symptoms often found with wheeze but frequently associated with other disorders.

3. Empiric phrase associations and NIOSHTIC do not require that each entity be placed in exactly one location. Traditionally, any specific object has only a single place in a scheme, as a child and perhaps as a parent but never having any other relatives. This is not consistent with the empirically derived structures illustrated by Figure 1. A single term may reasonably appear in several different positions. For example, it can be argued that "particle size" is a member of the "exposure" category because the distribution of particle size may determine extent of exposure; however it may also reasonably be placed into the "protection" category in that smaller particles may warrant special filters, etc. Similarly for "wood," subjects listed many subclasses (oak, redwood), but they also listed plicatic acid. The latter term also would be appropriately included elsewhere. Viewed as an industry, the associated terms, "sawmill" and "cabinet maker" might be located in several places.

4. The extensiveness and types of encoded relationships depend on the perspective of the user. When the types of empiric phrases mentioned by our subjects were analyzed against the backgrounds of the subjects, a

TABLE 3
Absolute Frequencies of Associated Terms of Several Types by Discipline

Discipline*	Class			
	Agent	Clinical	Disease	Symptom
Weld				
Occupational	28	0	20	1
Pulmonary	16	0	7	2
Hygiene	18	0	4	0
		$P = .15$		
Wheeze				
Occupational	3	6	25	9
Pulmonary	1	11	21	5
Hygiene	11	2	13	3
		$P = .0004$		
Wood				
Occupational	30	0	12	2
Pulmonary	15	1	10	3
Hygiene	14	0	7	0
		$P = .43$		

* "Occupational" and "pulmonary" refer to specialties of physicians, and "hygiene" refers to industrial hygienists. The number of phrases of selected types is shown according to primary discipline of subject. χ^2 calculations were performed to determine P values.

number of interesting patterns emerged. Not surprisingly, the specific disciplines of the subjects in the empiric phrase study determined the types of words they were most likely to mention. Results are summarized in Table 3. For example, when "welding" was given as the object term, industrial hygienists listed a higher percentage of protective devices than did occupational or pulmonary specialists. Similarly, when "wheeze" was the object term, pulmonary specialists named a higher proportion of disease and clinical finding words than did the other two subject groups, whereas industrial hygienists named considerably more agents that cause wheeze than did the other two groups.

5. Extensive virtual knowledge is embedded in empiric phrase association and NIOSHTIC coding. A direct relationship is one that is explicitly encoded, whereas a virtual relationship is inherent in the coding system but is not explicitly expressed. Traditional coding schemes include only direct connections. A large number of virtual relationships were identified in our empiric studies. Many of the associated terms (in both empiric phrase associations and NIOSHTIC data) do not superficially appear related to the index term. However, examining the relationships that are indirectly inferred provides insight into the manner in which information is organized. For example, the subject who suggested that "nasal cancer" came to mind in association with "welding" may have actually implicitly created a causal chain composed of links that are individually more obvious (for example, welding causes fumes, metal fumes include chromates, and chromates cause nasal cancer; therefore, welding and nasal cancer are related terms). That is, many relationships existed in their minds without having been explicitly encoded in advance.

Although the subjects were not asked to explicitly state the basis for the relationships they described, the investigators used common occupational health knowledge to develop postulated networks to explain the terms given by the subjects. Figure 2 shows how terms, all of which were suggested by subjects, could be explained. The empiric results suggest the underlying mechanism employed by the subjects to produce a chain of sequential links among more closely related terms, thereby creating a chain leading to the associations reported. Thus, although the subjects did not directly encode the associations seen but generated them when faced with the question, the information was truly present. These are thus "virtual" relationships rather than directly coded relationships.

Encoding of relationships may be viewed as a mapping function whereby connections among objects represent mapping of objects in a domain class onto objects in a range class. For example, the parent-child relationship may be expressed as a mapping of an object onto its parent superclass identifier, and similarly, its children (subclasses) are all mapped onto it. The inherently hierarchical structure of the traditional coding schemes implies a one-to-one or many-to-one mapping structure. This means that any specific object is mapped onto only a single other object (the superclass identifier) but may have many objects mapped onto it (its children or subclasses). Figure 3 graphically illustrates a many-to-one mapping. The multiplicity of relationships seen in Figure 2 suggest that a many-to-many mapping structure is more appropriate. This is particularly evident when one considers the cross-type as well as the within-type relationships empirically found in this study.

Figure 3 also demonstrates many of the difficulties in creating a system of this type. As classifications become more specific, more and more qualifications are necessary to place items in distinct categories. For example, "beef" is included under both "meat packing" and "sausage and other prepared meat products"; in the former case its qualification is that it is made in the same establishment as the basic materials, whereas the qualification in the latter case is that the beef is made from purchased materials. Furthermore, the classification scheme may force placement of items into wholly inaccurate categories, such as "fruit pops, frozen" within the "dairy products" category. Also, notice that a major group was required to described miscellaneous food products, even including the smaller category, "food preparations, not elsewhere classified." This type of categorization, inclusive of many different food industries, makes it almost impossible to validly state that the food industries within the code are "similar" to food industries with nearby codes. These examples illustrate some of the difficulties that arise in attempting to create systems that logically and exhaustively place items in exclusive locations.

Discussion

Convergent changes in the fields of linguistics, computer science, natural language understanding,[13] and artificial intelligence over the past decade have led to interest in knowledge engineering, and specifically, to the structure of knowledge. Knowledge structure can have significant impact on the use of information. In the

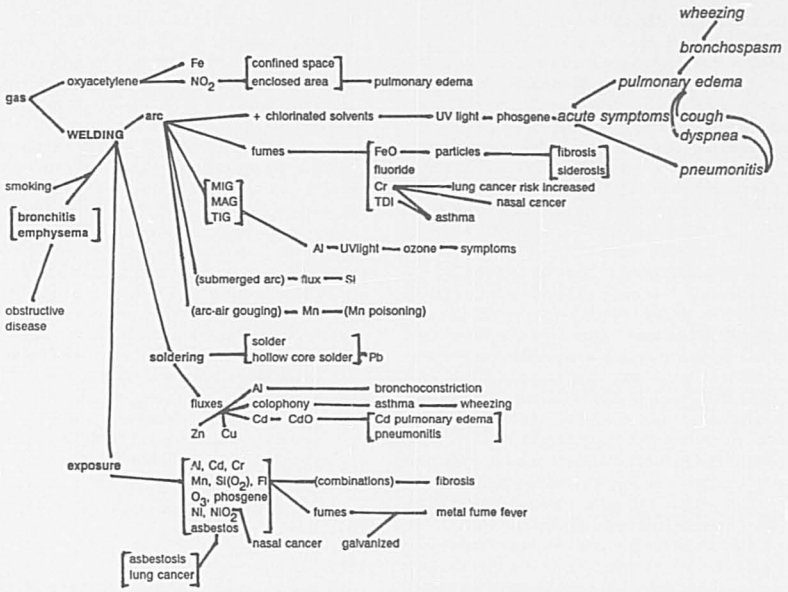

Fig. 2. Virtual knowledge: inferred relationships for "welding." Each subject gave five to ten of these words. Some words were mentioned by more than one subject whereas others were mentioned only once. Note that there are other possible arrangements; for example, the italicized group of words to the right of the figure are all words that fit elsewhere on the figure but also can be reasonably placed here as well.

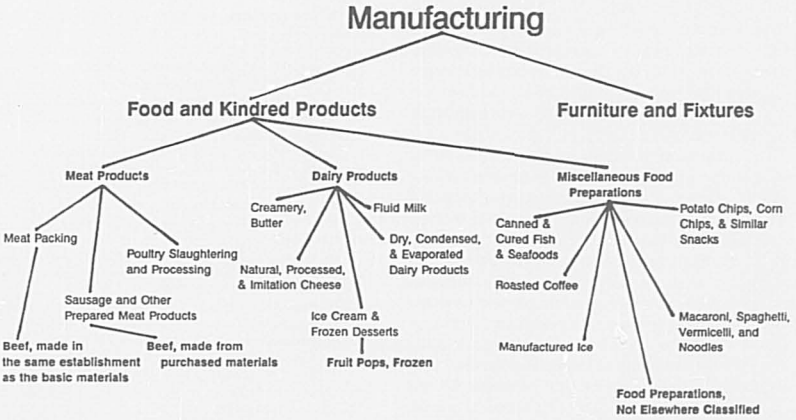

Fig. 3. Many-to-one mapping taken from SIC. The hierarchy shown gives a breakdown of the major group "food and kindred products" under the industry division, "manufacturing." Bold entries toward the bottom are discussed in text.

realm of occupational health, a number of schemes commonly are used for organizing information about occupational health. They have not been examined for efficacy and efficiency in any systematic way. This study was a pilot endeavor to evaluate the extent to which the current coding schemes reflect the natural means for organizing information. Specifically, the coding schemes implicitly used by occupational health professionals were compared with the formal coding schemes that have conventionally been applied.

The study demonstrated there is significant disparity between the manner in which information is structured in the typical hierarchical coding schemes and the actual structure of data used by professionals. Meaningful differences emerged: Unlike traditional information encoding structures, natural organization is not strongly typed. Relationships are not hierarchical as in conventional systems, and a specific entity may simultaneously belong in several locations. The depth and breadth of the structures is dependent on the characteristics of the user. Finally, relationships may be derived by virtual rather than by explicitly coded knowledge.

What is the significance of these findings? The empiric studies provide insight into the natural process by which information is organized. Our data underscore the highly artificial nature of the traditional schemes. In view of the extensive evolutionary effort represented by native thought processes, it is a priori likely that they are more complete and more efficient than a highly contrived scheme.

Furthermore, the manner in which information about work is coded has significant influence on its usefulness in research, clinical practice, and public policy. In the research sphere, classification is important because relations between exposures and outcomes are often evaluated for subjects who are placed in exposure categories based on coding information. Our data suggest there may be considerable loss of information content when "natural information" is translated into the highly artificial hierarchical coding schemes. Such information loss can lead to misclassification and decreased power of epidemiologic and other studies. For example, virtual relationships are inaccessible to the established schemes. Improved coding schemes, such as those shown in Figure 2, could significantly increase both the power and significance of studies. Moreover, epidemiologic research studies also frequently involve adjustment for confounding variables, again, generally based on data that are aggregated.

Clinically, information based on work coding often is used in establishing diagnoses or in making work placement decisions. For example, determination of possible exposures and possible consequent health effects is often based on such information. The knowledge with which an individual clinician is familiar may be limited, as illustrated by our analysis of the relationship between subject background and stated empiric phrase associations; consequently, outside sources designed for use by such clinicians should be organized to enhance the ability to make correct decisions in unfamiliar areas. Our data imply there is not a direct link between formal coding schemes and clinical thought. In terms of public policy, information that is structured and/or coded inaccurately can lead to laws and employment regulations that do not adequately account for all the many issues related to occupational health. For policies to be maximally beneficial to employers and workers alike, it is essential that policy be based on information about work that is structured to reflect reality.

Thus, we believe the data strongly suggest the need for research into alternative coding schemes that are more efficient and more akin to those naturally used. Particularly in view of recent advances in information science (both conceptually and in computer implementation) the opportunity exists for critically examining coding schemes. Data base management systems exist in several general categories[14]: relational data bases (eg, dBase), hierarchical (eg, SIC, SOC), and semantic networks. The latter method is based on a format in which terms are related in as many ways as they may be related naturally. Although hierarchical models are traditional, there is no a priori basis for this choice. Indeed, it appears that the semantic network data base management format appears similar to that empirically used and should be investigated further. If it is possible to construct coding schemes for occupational health that mirror the natural structure of knowledge of health professionals, there is significant potential for gaining more benefit from available information.

References

1. Office of Federal Statistical Policy and Standards. *Standard Occupational Classification Manual.* Washington, DC: US Department of Commerce; 1980.

2. *Standard Industrial Classification Manual.* Washington, DC: Office of Management and Budget, Executive Office of the President; 1987.

3. *Chemical Abstract Services Index Guide.* Washington, DC: American Chemical Society; 1989.

4. *International Classification of Diseases: Manual of the international classification of diseases, injuries, and causes of death,* 9th rev. Geneva: World Health Organization; 1977.

5. *1980 Census of Population.* Alphabetical index of industries and occupations, PHC80-R3 final ed. Washington, DC: US Department of Commerce, Bureau of the Census; 1982.

6. *Dictionary of Occupational Titles,* 4th ed. Washington , DC: US Employment Service, US Department of Labor; 1977.

7. Keefe AR, Grace JR, Band PR, et al. A hierarchical coding system for occupational exposure. *J Occup Med.* 1991;2:127–133.

8. Hoar SK. Job exposure matrix methodology. *J Clin Toxicol.* 1984;21:9–26.

9. Hoar SK, Morrison AS, Cole P, et al. An occupation and exposure linkage system for the study of occupational carcinogenesis. *J Occup Med.* 1980;22:722–726.

10. Hoar SK. Epidemiology and occupational classification system. In: Peto R, Schneiderman M, eds. *Banbury Report 9. Quantification of Occupational Cancer.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory; 1981:455–470.

11. Gerin M, Siemiatycki J, Kemper H, et al. Obtaining occupational exposure histories in epidemiologic case-control studies. *J Occup Med.* 1985;27:420–426.

12. *CCINFO* disc, NIOSHTIC database. Ontario, Canada: Canadian Centre for Occupational Health and Safety, 1991.

13. Allen J. *Natural Language Understanding.* Menlo Park, Calif: The Benjamin/Cummings Publishing Company, Inc.; 1987.

14. Ozkarahan E. *Database Management: Concepts, Design, and Practice.* Englewood Cliffs, NJ: Prentice Hall; 1990.