Indexing Medical Information
The Role of SNOMED

D.J. Rothwell
Columbia Hospital

F. Wingert
Munster, F.R.G.

R.A. Cote
University of Sherbrooke, Quebec

R. Beckett
Hartford Hospital

J. Palotay
Oregon Regional
Primate Center

ABSTRACT

Indexing the content of medical reports requires that the data to be organized into appropriate data structures which are capable of recording information at various levels. SNOMED is presented as a data structure suitable for housing medical information and serving as a knowledge base for medical information. Issues related to classification of data and problems related to the computerized processing and automated indexing of natural language are discussed.

Automated natural language systems which accept natural language data input, answer questions about a knowledge base, make inferences, and generate natural language responses are complex, and techniques for designing them are in the early stage of development. Much work is being done on different aspects of the systems to develop more complete grammar, lexicons, and knowledge representation schemes to broaden their domain and improve their efficiency and accessibility.

Natural language processing by computers requires that language data be organized into appropriate data structures which are capable of recording information at various levels. Levels include the lexical, morphological, syntactic, and contextual content of a statement to be processed.

This paper addresses the processing of natural language data in the restricted domain of diagnostic and therapeutic medical statements. In so doing, it avoids many of the problems encountered in free natural language processing and analysis. Furthermore, processing is focused on indexing, which is the most important intermediate objective of such systems.

The Systematized Nomenclature of Medicine (SNOMED) is offered as an example of a multidimensional nomenclature and is a classification system for medical nomenclature. The structure of SNOMED is introduced with emphasis on its capabilities with respect to an artificial medical language.

Introduction to Indexing Medical Information:

Indexing is the processing of medical statements by applying transformations to a language expression that result in the structure of a metalanguage. This is followed by encoding. Therefore, indexing includes much more than just encoding because the actual intellectual achievements are required for the transformation. As an example, the statement "the resulting complication from this medication was a gastric ulceration which bled and perforated" must be transformed to a suitable index such as: stomach, with ulcer resulting in bleeding and perforation caused by medication. But, frequently, indexing and encoding are used as synonyms.

Classification is the systematic ordering of all concepts in a scientific area. Because concepts are communicated by language, a classification is represented by a systematic list of classification rules or a dictionary of denoted terms.

The most important relation used for classification is the hierarchic relation. The following types of classifications are distinguished:

- Monohierarchic classifications are generated by successive refinement and are, typically, one-dimensional. Each concept has exactly one parent, a strict hierarchy.
- Polyhierarchic classifications are generated when the same concept is refined by more than one attribute. Therefore, a concept may have several parents, resulting in a tangled hierarchy.

Hierarchic relations, in turn, can be divided into:

- Partitive - a total-part relation i.e. "is a" relationship (e.g. renal cortex "is a" part of the kidney.
- Generic - a logical relation of parts i.e. a specialization or a superordinated/subordinated concept (e.g. an abnormal pulse is related to the heart and the cardiovascular system).

Monohierarchic classifications are frequently insufficient for complex applications. On the other hand, polyhierarchic classifications tend to be voluminous and redundant and require such an investment for maintenance that, frequently, they become outdated soon after their original design.

In order to realize the advantages of polyhierarchic classification or nomenclature, a significant reduction of the total size without a considerable loss in precision can be reached if there is a **semantic model** of the domain of discourse. Then, a polyhierarchic classification can be replaced by:

- Several disjoint monohierarchic classifications connected by a semantic model.

If such a model is paralleled by a decomposition of statements into parts belonging to the different dimensions of the model then a polyhierarchic dictionary can be replaced by:

- Several disjoint monohierarchic dictionaries the combination of which is described by a metalanguage.

SNOMED is offered as an example of such a structure.

Classification-Free Indexing has been proposed as a solution. There have been frequent discussions on whether indexing requires classifications at all. The supporters of a classification-free indexing in medicine have some very serious arguments on their side:

- Because each classification is goal-oriented, the underlying model as well as its realization is biased towards the objectives of the classification. Two examples of this are:

  1. If a classification shall be used for reimbursement, then by definition - the members of a class must be homogeneous with respect to costs. Medical criteria may totally be neglected.

  2. A statistical classification is, at least in part, constructed such that the class frequencies do not differ too much from an equal distribution in some sort of a reference population. The reason is that the power of the significance tests for frequencies is maximized if the null hypothesis states that all probabilities are equal. There is no doubt that such a requirement has nothing to do with medical criteria.

- Because classification leads to an information loss, the consequence is that data classified according to a specific classification may not be evaluated according to another classification. This situation is very common in an open documentation system where retrieval criteria are not fully known during data acquisition.

- Classifications tend to change over time according to new knowledge or changing objectives. Often since the resources for reclassification of old data are not available there is no possibility of comparing old data with new ones.

- The investment made by using a classification should have a corresponding pay-off. Usually, the minimum requirement is the comparability of data with those collected from other sources. In other words, the value of a classification increases with its acceptance. Unfortunately, acceptance is the inverse of flexibility, especially, if the domain crosses the borders of medical specialities or even political borders. The more classification is standardized and used, the more effort must be put into the process of reaching consensus and the more difficulties arise when adaptation to new knowledge becomes necessary.

These well-known shortcomings of practically every classification have frequently been the background for the argument, "Let's get rid of classifications and let's store and retrieve the original data." This proposal is supported by the enormous growth of storage capacity in modern computer systems that make encoding for technical reasons superfluous.

The answer to these arguments is not simple and requires an analysis of the alternatives. One subgroup argues that it is insufficient to store just codes in a documentation system. Instead, the original data should be stored and retrieved. Codes according to one or more classifications should be hidden to the user. This is a very reasonable argument because a indexing system that stores just codes can neither be checked nor changed as soon as the classification changes. But, this is not classification-free at all.

This subgroup can be characterized by the approach that a list or precoded data is kept in the system, the codes being selected according to one or several classifications. In the latter case, the user has a choice between alternatives during retrieval.

The only real alternative is to store just the original data. Then just the original data can be retrieved. If any type of grouping is required for retrieval then the user is left with the task of enumerating all variations of the external representation of a given information or a superordinated concept he is looking for. If someone is searching all data denoting inflammatory processes of the kidney, he has to identify all possible formulations of this fact. There is no doubt that the result of such a query is almost randomly distributed among a sample of users.

A more intelligent retrieval system should support a user by supplying the knowledge about which elements are relevant for the query. But then the debate starts all over again. Labeling data according to concepts such as kidney or inflammation is nothing else than making explicit the projections of a point in an information space with regard to a predefined set of axes and of discrete points on these axes. The concepts and their values that are selected for such labels together with the relations between the concepts

(kidney is part of the urinary tract) extend the possible queries.

The conclusion is that intelligent indexing systems cannot totally discard classifications. Moreover, the availability of at least one classification is a necessary condition for a good indexing system. Classifications are not a necessary evil but a very effective way of representing knowledge about the domain of discourse.

There are many different reasons for indexing and then representing this information with codes. A selection of these reasons is listed below.

Standardization of terminology:

- Coding requires a careful definition of terms.
- In retrieval the preferred term replaces the internal code.
- Multilingual or multidialectic terminology can be mapped onto the same representation.
- The code can be used as a terminal symbol of an intermediate language by different medical languages in order to facilitate translation.

Facilitation of data storage:

- Codes have usually a much more rigorous structure than language data. Therefore, data elements can be better defined by formal means and the construction of data aggregates is facilitated.
- Storage capacity requirements are reduced.
- Retrieval is facilitated, less expensive and more precise.

Representation of relations:

- Representation of relations is facilitated. This is especially true for synonymy and hyponymy (hierarchic coding). Therefore, retrieval is supported by the use of background knowledge represented in the code structure.

Indexing may be done by human encoders knowing the terminology and the classification. If this approach is used then a common postulation is that indexing is done by those who generate the data, i.e. physicians. There are some good reasons for this requirement but practice has often demonstrated that, in spite of theoretical advantages, the results are worse than when indexing is performed by specially trained medical record officers. Even with special training manual indexing is difficult particularly when the indexing system is complex as is SNOMED.

The alternative is automated encoding which has some promising advantages:

Improvement of consistency:

- Algorithms neither get tired nor change their mind after some time. Therefore, the same input is always mapped onto the same code. This is even true for the errors. Random code assignment is excluded.

Improvement of correctness:

- Some automated systems produce better results than humans when coding large amounts of data.
- In some applications a semiautomated procedure may be the best solution, at least, as long as the system is not sufficiently stable. The automated encoder can be enhanced by implementing learning capabilities.
- Error checking can be formulated explicity.

Increase in encoding adaptability:

- In a manual coding system it is usually impossible to recode a large data set when coding rules or classification change. In an automated system this can be done usually with little difficulty.

For technical reasons, frequently, numerical systems are preferred because they are especially well-suited for representing hyponymy (hierarchic coding, Figure 1). Hierarchic coding considers the following relations:

Synonmymy	(Quasi-)synonyms have the same code.
Hypernymy	The code of a parent is the prefix of the codes of all siblings.
Cohyponymy	Identical code prefixes.

Therefore, hierarchic coding represents more than just a mapping into another character set, i.e., it represents medical knowledge.

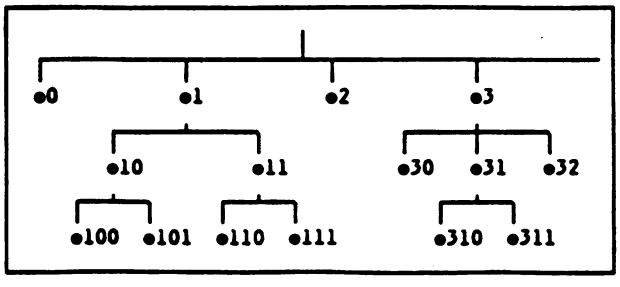

FIGURE 1. MAPPING A HIERARCHIC CLASSIFICATION ONTO A NUMERICAL CODE

SNOMED:

SNOMED is a seven-dimensional categorized nomenclature consisting of about 45,000 terms classified into exactly 1 out of 7 semantic dimensions. The German adaptation is about twice as large because multiple synonyms, general classes, and specific morphemes have been added. If modifiers are considered, the number of lexemes rises to more than 200,000.

Each dimension is a structure tree with up to five levels generated by considering hierarchic relations. The hierarchy is reflected in the hierarchic code structure. Each code consists of a letter indicating the semantic dimension and of up to 5 characters represented by the digits 0 to 9, A thru F. In Topography, the hierarchy is created by considering the partitive hierarchic relation; in the other dimensions, it is created by considering the generic hierarchic relation.

The SNOMED dimensions are:

Topography	Body parts, body systems, organs and regions. This dimension contains the terms for tissues, organs, body fluids and spaces as well as subcellular structures.
Morphology	Alterations of normal anatomy, whether congenital or acquired after birth, i.e., pathologic anatomy of cells, tissues, organs and systems.
Etiology	Terms used to describe causes of morphological and/or functional alterations such as organisms, chemicals, drugs, physical agents.
Function	Terms for (patho)physiology as well as terms used to describe functional states or processes. Its highest order partitioning into chapters is mainly according to the body systems.
Disease	Disease concepts and syndromes. This is a one-dimensional classification of complex nosological units and is mainly partitioned according to the body systems.
Procedure	Administrative, diagnostic and therapeutic activities available to prevent, relieve or cure diseases. Its highest order partitioning is mainly according to the body systems, i.e., the medical specialities.
Occupation	Classification of occupations borrowed from the International Labor Organization. The dimension is of help in coding environmental and occupational relationships to diseases.

If we do not consider the special dimension Occupation the fundamental structural SNOMED model is:

A procedure
for a disease or a morphologic alteration
in a localization
caused by an agent
combined with a functional disturbance.

According to this model, a statement is transformed into a TMEFDP- or SNOMED statement Figure 2. For the analysis it is assumed that each class from one dimension can be combined with each class from another dimension.

It is quite important to note that SNOMED is not a general classification of medical terms. This statement must not be confused with the fact that the system aims at covering the total information space of medical statements, i.e. SNOMED offers the fundamental information structure which is required for indexing medical language data to a significant degree of detail.

Applications for such a powerful multidimensional nomenclature are found wherever language-based communication takes place. Applications range from clinical documentation and health statistics to knowledge organization and automated translation from one medical language to another.

Data mapped into SNOMED may either be retrieved on the basis of low-level information atoms such as: "fracture of the body of the 12th thoracic

Statement	Topography	Morphology	Etiology	Function	Disease
Tuberculosis of the lung	Lung T28000				Tuberculosis D01880
Lipoma of the trunk	Trunk TY1000	Lipoma M88500			
Pharynx paralysis	Pharynx T60000			Paralysis F80840	
Neurosyphilis	Nervous syst. TX0000				Syphilis D01820
Mumps	Parotid gland T55100				Mumps D04330
Streptococcal pharyngitis	Pharynx T60000	Inflammation M4000	Streptococcus E25000		

FIGURE 2. SNOMED CODING EXAMPLES

vertebra due to fall from a ladder"; or on the basis of high level information aggregates such as: "inflammation of the gastrointestinal tract". To do this, just the hierarchic code structure has to be used.

These examples provide some insight into the attribute structure or deep structure of SNOMED. From the point of view of indexing, the corresponding problem to solve is the decomposition of each medical concept strictly into their respective SNOMED entities and their position in a SNOMED space (Figure 3).

SNOMED, as it exists, is not an explicit definition of the deep structure but a structured multidimensional list of denotations of the basic entities, i.e., a list of terms and names from which the deep structure has to be derived. Therefore, an indexing procedure must include a set or a sequence of transformations of a given language string which ultimately result in a formulation reflecting the deep structure by using only SNOMED terms in the denotation of the concept. This may be extended by denoting the relations between the different dimensions.

Multidimensional Nomenclatures:

The principle of multidimensional nomenclatures has been introduced above. They may open a way out of the dilemma that we need classifications for building intelligent documentation systems and at the same time are confronted with the insufficiences of these classifications. In order to play the role of an intermediate language sufficiently well, a nomenclature has to meet certain requirements:

- The utterance model must reflect the medical principles of aggregating information, i.e., it must be multidimensional. The selection of dimensions must be such that the structure is elementary enough to serve as a basis for expressing medical information of all kinds and at the same time be simple enough to be handled with the algorithmic instruments that are available now.

- The nomenclature must incorporate as much medical knowledge as possible. For example, it must have the means for the recognition of synonyms, homonyms, and hierarchical relations.

- The structure of the nomenclature must be free of classificatory elements that lead to a bias which makes mapping into other classifications more difficult or even impossible, e.g. the position of pituitary dwarfism may not prohibit the classification of this entry among dwarfism as well as among diseases of the pituitary gland.

These requirements call for the introduction of associative relations, among which cross references across and within dimensions are of most importance. This is absolutely necessary to fully specify and relate information found in a single dimension (monohierarchy) with information that is contained within a given concept in other dimensions (polyhierarchy).

- The nomenclature must be unambiguous. No information unit may be representable by

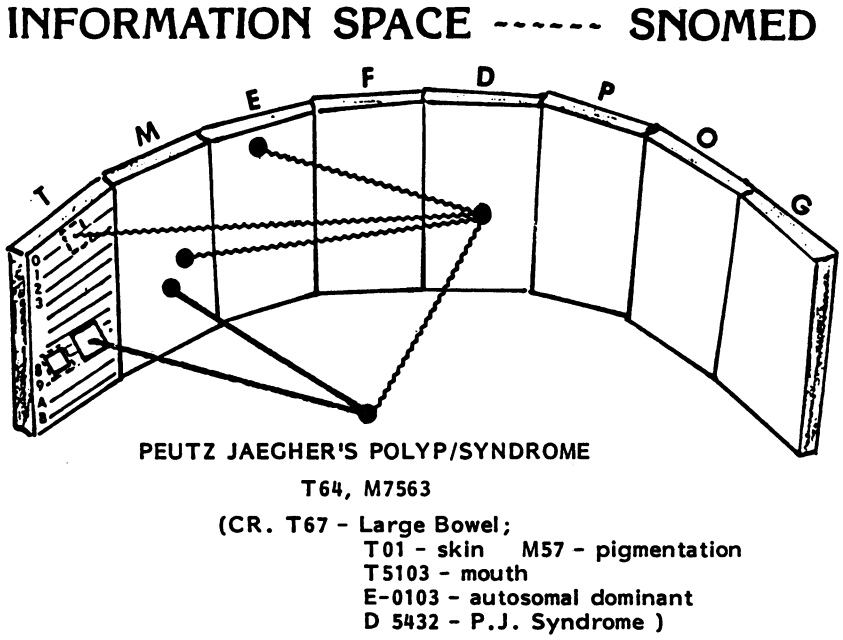

Figure 3.

different projections. This rule is crucial.

- The nomenclature must be complete both with respect to the classes and with respect to the terms. In this context it must be kept in mind that a nomenclature is always explicit, i.e., it is a list of terms and not a set of rules, telling where a term has to be positioned.

- The nomenclature must be free of redundancy in order to reduce maintenance problems.

Feinstein in a recent paper outlines three major requirements for system of taxonomic classification in science. 1) The classification must have a suitable organizing principle. 2) The categories must be labeled with standard titles. 3) The members of each category must be identified with a suitable set of operational criteria. The organizing principle of SNOMED rests on its dimension i.e. where is the lesion (topography); how is the lesion described (morphology); what are its associations or causes (etiology); what functional disturbances does it cause (function); what is the name of the lesion (disease) and what is done about it (procedure). These categories (dimensions) can be construed as the natural parts of medical speech providing standard titles for each dimension. Defining operational criteria for inclusion of each element within such a system has as yet not been addressed in a direct way.

The development of SNOMED has demonstrated a promising direction through which progress can be made. The underlying structure is sound, further elaboration is required prior to widespread clinical implementation.

REFERENCES:

1. Cote, R. SNOMED. Systematized Nomenclature of Medicine. College of American Pathologists, Skokie, Illinois, 1979.

2. Wingert, F. Medical Linguistics: Automated Indexing into SNOMED. Critical Reviews in Informatics 1988; 1:333.

3. Wingert, F., Rothwell, D., Cote, R. Automated Indexing into SNOMED and ICD. IMIA - WG6. Role of Informatics in the Classification and Coding of Health Data. Geneva, Switzerland, September, 1988.

4. Feinstein, A. ICD, POR and DRG. Arch. Int. Med. 1988; 148:2269.

5. Grishman, R. and Kittredge, R. Analyzing Language in Restricted Domains: Sublanguage Description and Processing. Lawrence Erlbaum, Hillsdale, 1986.

6. Harris, M.D. Introduction to Natural Language Processing. Reston Publishing Company, Reston, 1985.

7. Sager, N. Natural Language Information Processing: A Computer Grammar for English and its Application. Addison-Wesley, Reading, 1981.

8. Cote, R. A., Protti, D.J., and Scherrer, J. R. Role of Informatics in Health Data Coding and Classification Systems. North-Holland, Amsterdam, 1985.

9. Pratt, A. W. Medicine, Computers and Linguistics. Adv. Biomed. Eng. 1973; 3:97.

10. Schneider, W. and Sagvall Hein, A.-L. Computational Linguistics in Medicine. North-Holland, Amsterdam, 1977.

11. Wingert, F. Medical Informatics. Springer, Berlin, 1981.

12. Mayr, E. Biological Classification: Toward a Synthesis of Opposing Methodologies. Science 1981; 214:510-16.

13. Cote, R. A. and Robboy, S. Progress in Medical Information Management. JAMA 1980; 243:756-62.

14. International Classification of Diseases. World Health Organization, Geneva, 1978.

15. Hutchins, W.J. Languages of Indexing and Classification. Peter Beregrinus, Ltd, South-gate House, Stevenage, Herts., England, 1975.

16. Sparck Jones, K. and Kay, M. Linguistics and Information Science. Academic Press, New York, 1983.

17. Soergel, D. Indexing Languages and Thesauri: Construction and Maintenance. Melville Publishing Company, Los Angeles, 1974.

18. Sager, N., Friedman, C., Lyman, M. Medical Language Processing. Addison-Wesley Pub. Co., Reading, MA, 1987.

19. Pacak, M.G., Dunham, G.S. Computer & Medical Language. Med. Inform. 1979; 4(1):13-27.

20. Sowa, J. Working Group on Data Bases (WG26). Elsevier, North Holland, 1989.