Developing A Standard Data Structure For Medical Language - The SNOMED Proposal

Rothwell DJ[a], Cote RA[b], Cordeau JP[c], Boisvert MA[c]

Columbia Hospital, Milwaukee, Wisconsin 53211
[b]University of Sherbrooke, Sherbrooke, Canada J1H-5N4
[c]University of Montreal, Sacre Coeur Hospital, Montreal, Canada H4J-IC5

ABSTRACT

The Systematized Nomenclature of Medicine, Third Edition, SNOMED International, is a comprehensive structured nomenclature of human and veterinary medicine, the terms of which are detailed, fine grained and semantically typed. Terms are assigned to eleven independent modules (fields), each of which is systematized. Terms may be linked to on another to represent complex entities or manifestations or alternately complex terms dissected into their elemental parts. Terms are illustrated utilizing a frame representation. Efforts are in progress to build both a conceptual graph and a frame-based semantic network encompassing each SNOMED term, effectively building a knowledge base. In this way, the knowledge contained in each alphanumeric representation is made explicit. SNOMED is a linked data structure capable of faithfully representing the activities, observations and diagnoses found in the medical record in a computer processable form.

INTRODUCTION

Natural language is the most expressive and powerful knowledge representation method available to us today.[1,2] The ability to express subtle shades of meaning, to avoid contradiction and the ability to reason are its principle strengths.[3] Its principle faults, however, are its inherent ambiguity and absence of a computable data structure.[4] The subset of natural language we recognized as medical terminology, like any technical vocabulary, seeks precision in order to promote clarity and to assist in correct decision making.[5] Since much of medicine deals with incompletely understood natural phenomena medical terminology is often descriptive of phenomena rather that definitive. The challenge is to devise a systematic hierarchical unambiguous terminology which can define as well as describe medical phenomena and to express the relationships among its terms and concepts. An optimal computer processable data structure incorporating each of these features must then be selected.[6,7]

KNOWLEDGE REPRESENTATION

What features must a good knowledge representation possess? In general, representation of knowledge requires that relevant objects in the knowledge domain be named, described, and organized and that relationships between objects be expressed including constraining relationships that govern storage and retrieval of object properties.[8] The goal of a knowledge representation method is to carry out these functions in an efficient manner. In a parallel manner, a well controlled medical terminology suitable for computer processing requires many of these same properties.

Useful knowledge in a particular domain may occur in a variety of forms; declarative (nonadaptive) and procedural (adaptive). Declarative knowledge is that in which the content is predetermined e.g. narrative text, diagrams and images. Its presentation or view may differ depending upon the user and context but the actual content does not change. It may be unstructured (e.g. text or image) or structured (e.g. data base, semantic nets, frames). Procedural knowledge is that in which content is computed and modified depending on user action and context (e.g. inference, data analysis).[9] A significant portion of medical knowledge is still largely embodied in narrative documents. A structured terminology that can process both declarative and procedural knowledge is necessary. This becomes a critical issue when attempting to build, access or manipulate a medical knowledge base.

SNOMED CONTENT

The third edition of the Systematized Nomenclature of Medicine, SNOMED International,[10] is presented as a comprehensive medical terminology having virtually all of the desirable features listed above and in addition a data structure that is modular, opened ended and possess a flexibility suitable for expressing simple as well as complex concepts and their relations in a highly structured linked data model. It

0195-4210/92/$5.00

SNOMED INTERNATIONAL

MODULE	Number of characters for termcode	Number of characters for termcode	Records
T = Topography	5	120	12,385
M = Morphology	5	120	4,991
F = Function	5	120	16,352
L = Living Organisms	5	120	24,273
C = Chemical Drugs & Biological Products	5	220	14,138
A = Physical Agents, Forces and Activities	5	120	1,355
D = Disease/Diagnosis	6	200	28,622
P = Procedures	6	220	27,033
S = Social Context	5	120	433
J = Occupations	5	120	1,886
G = General Modifiers	4	80	1,173
		TOTAL RECORDS	132,641

Figure 1

has the features that we believe are necessary for a well grounded medical terminology that is capable of 'packaging' concepts i.e. information units, both simple and complex into computer processable entities. A terminology constructed in this manner can serve as a structured nomenclature as well as a classification system.[11,12]

SNOMED currently contains eleven modules, each an independent taxonomy, representing the semantic categories necessary for describing and indexing virtually all of the events found in the medical record. The eleven fields with their respective sizes are shown in Figure 1.

All of the important concepts in human and veterinary medicine are placed within this structure. Brief Definitions of each of the modules (semantic categories) are: Topography - detailed anatomical terms in human and veterinary medicine; Morphology - terms used to describe structural changes in the body and in addition the tumor nomenclature found in the morphology section of the International classification of Disease for Oncology (ICD-O). Living Organisms - an unabridged classification of the animal kingdom including bacteria and viruses encompassing all of the pathogens and animal vectors of disease; Chemical/Drugs - a compilation of drug terms with cross referenced generic and trade names each assigned to its class and for compound drugs the constituents of each; Function - clinical terminology related to signs and symptoms along with the terms used to describe biochemical and physiologic processes; Disease/Diagnosis - a clinical terminology file of the named diseases encountered in medicine incorporating each entity found int the ICD-9-CM; Occupation - the International Labor Office (ILO) list of occupations; Procedures - a comprehensive list of administrative, therapeutic and diagnostic procedures used by all health care personnel in all specialties and each medical discipline; General - syntactic linkages and qualifier terms; Physical Agents, Forces and Activities - a listing of those devices, forces and activities commonly associated with disease; Social - an embryonic list of social conditions related to ethnic or religious heritage, family status or economic condition.

SNOMED STRUCTURE

SNOMED International is a structured vocabulary that is: 1)Systematic, 2)Modular, 3)Linked.[13,14] By Systematic we mean it contains all (i.e. it is designed to accept all) of the accepted terms and concepts in an ordered polyhierarchy that are used in medicine. Synonyms and related terms are assigned their position in the hierarchy. Both simple and complex concepts are represented. Each term is assigned an alphanumeric code, often with extensive crossreferences to other terms within its own module or to

D-11120	
D -	- DIAGNOSTIC TERM
D-10000	- METABOLIC/NUTRITIONAL DISORDER
D-11000	- DISORDER OF MINERAL METABOLISM
D-11100	- DISORDER OF IRON METABOLISM
D-11120	- IRON STORAGE DISEASE

Figure 2A

F-10363	
F -	- FUNCTION (PHYS. UNITS:S/S)
F-10000	- UNIT OF METABOLISM
F-103	- ELEMENT, ION, SIMPLE COMPD.
F-10363	-IRON
F-10363	- IRON, INCREASED

Figure 2B

different modules (see below for example). The assignment of a term to a particular module and its placement within that module effectively profiles a semantic type for that term. Using this information it is possible to build a conceptual graph or a frame-based semantic network for each term in SNOMED International.[15] By Modular we mean that division and assignment of terms into the categories each representing, for a given patient's condition, a natural part of medical speech i.e. where is the lesion (T), how is it described or what is its name (M,D), with what signs, symptoms or physiologic alterations did the patient present (F), were living organisms involved (L), were chemicals or drugs involved (C), what forces, agents or activities are implicated (A), what is the occupation and social context (J,S), what was done (P), and finally linking or qualifying each of these elements describing a patient's condition (G). Each module is systematized. This modular approach provides for relative ease of maintenance for each term set. By Links we mean the provision of a set of basic relational and qualifying terms used to connect basic concepts with on another so as to formulate more complex concepts and/or manifestations of disease processes. This is a particularly powerful tool when SNOMED International is used to construct conceptual graphs or frame-based semantic networks.

LINKED DATA STRUCTURE

The following two examples illustrate the structure of SNOMED International. 1) The diagnosis, Iron Storage Disease is assigned two alpha numeric codes, D-11120; F-10363. The information contained in these codes is shown in Figure 2A and 2B.From this illustration it is clear that the information packaged within the code is far more than the mere substitution of a number for a word. The coded representation tells us that Iron Storage Disease is a metabolic nutritional disorder of mineral metabolism composed of an ion, element or simple compound, specifically an increase in storage iron. Each of these concepts is independently retrievable and form an effective knowledge base. The site(s) of iron storage can be specified and linked to the two frames illustrated above giving further definition and specificity to the condition. 2) The second example is Subacute Bacterial Endocarditis involving the mitral valve caused by Strep Viridans (SBE). This is indexed as D-34567; T-35322; L-25127; M-41000 as shown in Figure 3.

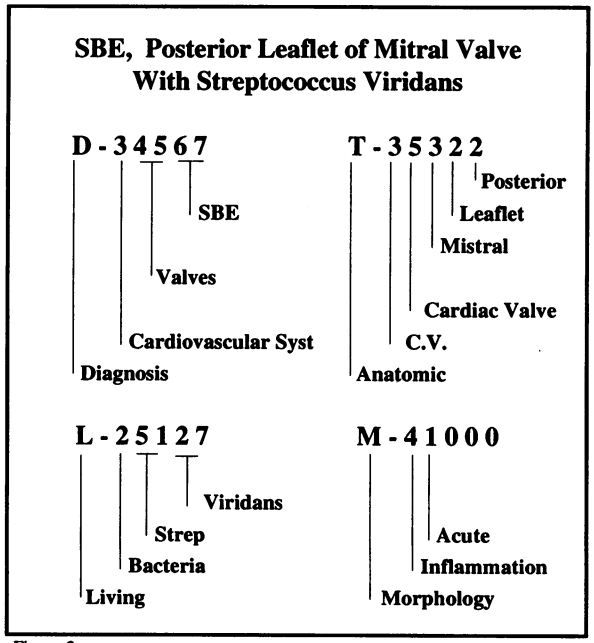

Figure 3

The information contained in this representation over and above that found in the expressed term are cardiovascular system, valve, mitral valve, bacteria and acute inflammation. As in the previous example, each are separately recoverable and make explicit and retrievable the information implicit in the diagnostic phrase.

Figure 4 illustrates how this representation for the same condition (SBE) (D) at Mitral Valve (T) associated with Strep Viridans (L) resulting in Acute Inflammation (M) can be extended with links to include complications, e.g. complicated by (CB) cerebral embolism, hemorrhage or any other condition.

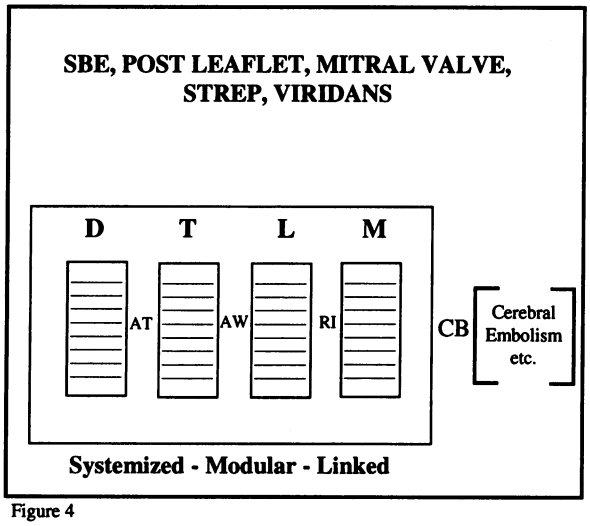

Figure 4

Using this data structure, frequently encountered complex entities can be constructed and conversely complex entities can be decomposed into their elemental parts. This we feel is a principle strength of the SNOMED approach.

APPLICATIONS

A significant proportion of the medical knowledge found in a medical record is textual, e.g. history and physical examination, progress notes and interpretation of x-ray images or of slides from a pathology department. The techniques used for representation of this knowledge in a computer system derived from text or images is critical to understanding its meaning and for the ease by which it can be retrieved. Whether representation of this static (stated) knowledge in natural language is sufficient or a coded representation of the information content is better raises one set of issues. Use of this same knowledge as a basis for making decisions and for use in expert systems makes the manner in which this knowledge is represented even more critical; it raises however a different set of issues and the two should not be confused. The diagram in figure 5 illustrates in a figurative way this difference, i.e. issues related to declarative knowledge (vocabulary, text, images) are shown in the lower half of the diagram, those related to procedural issues in the upper half. The structure and notation of SNOMED International described in this paper is, we believe, an efficient and effective method for representing textual medical information from any source such that it can be used in the wide variety of applications illustrated in the upper half of the diagram.

SNOMED International is presented as a candidate for a structured vocabulary for declarative knowledge.

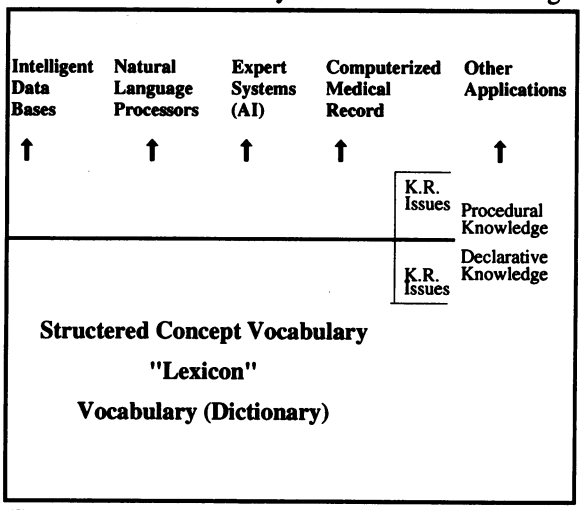

Figure 5

CONCLUSION

SNOMED International is a detailed, fine grained, semantically typed comprehensive and computer processable set of terms used in human and veterinary medicine. They are structured in such a way that complex diagnostic entities and their manifestations can be constructed and these same complex entities can be dissected into their individual parts. It will permit the composition of new terms from existing ones giving precise characterization of the new terms semantics. The computer-based patient (CPR) requires a standardized vocabulary/data dictionary. SNOMED International is a candidate for this role.

REFERENCES

[1] Wingert, Rothwell DJ, Cote R. Automated Indexing into SNOMED and ICD. IMIA-WG6 Role of Informatics in the Classification and Coding of Health Data, Geneva, 1988.

[2] Boguraev BK. Theoretical Issues in Natural Language Processing: The Definitional Power of Words, Wilks Y, Ed. Lawrence Erlbaum Associates Publisher. Hillsdale:NJ, 1989:7-10.

[3] Bench-Capon TJM. Knowledge Representation, An Approach to Artificial Intelligence. New York: Academic Press, 1990.

[4] Newmeyer FJ. Linguistic Theory in America, 2nd ed. New York: Academic Press, 1986.

[5] Fox J, Glowinksi A, Gordon C, Hajnal S, Oniel M. Logic Engineering for Knowledge Engineering: Design and Implementation of the Oxford System of Medicine. Artificial Intell. in Med. 1990, 2:232-339.

[6] Evans DA, Hersch WR, Monarch IA, Lefferts RG, Handerson SK. Automated Indexing of Abstracts via Natural-Language Processing Using a Simple Thesaurus: In AMIA Proceedings. San Francisco, June 1991.

[7] Rector AL, Nowlan SK, Kay S. Unifying Medical Information Using An Architecture Based On Description. In Proc SCAMC 90. Miller RD, Ed. IEEE Computer Society Press 1990:190-194.

[8] Parsaye K, Chignell M. Expert Systems for Experts. New York: John Wiley and Sons, 1988.

[9] Greenes RA, Deibel SRA. The DeSyGNER Knowledge Management Architecture: A building block approach based on an extensible kernel. Arficial Intelligence in Med. 1991, 3:95-111.

[10] Cote RA, Rothwell DJ, Beckette R, Palotay J (eds.) SNOMED International. College of American Patholgists, Chicago, 1993.

[11] Wingert F. Medical Linguistics: Automated Indexing into SNOMED. Critical Reviews in Informatics 1988, 1:333-403

[12] Rothwell DJ, Hause LL. SNOMED and Microcomputers in Anatomic Pathology. Medical Informatics 1983, 8:23-31.

[13] Rothwell DJ, Cote R. Optimizing the Structure of a Standard Vocabulary - The SNOMED Model. In Proc SCAMC 90. Miller RD, Ed. IEEE Computer Society Press 1990:181-184.

[14] Rothwell, DJ, Wingert F, Cote R, Beckett R and Palotay J. Indexing Medical Information - The Role of SNOMED. In Proc SCAMC 89. IEEE Computer Society Press 1989:534-539.

[15] Evans DA, Rothwell DJ, Monarch IA, Lefferts RG, Cote R. Toward Representations for Medical Concepts: In AMIA proceedings. San Francisco, June 1991.