Medical Infobahn for Europe
A. Hasman et al. (Eds.)
IOS Press, 2000

Graphical tool for navigation within the semantic network of the UMLS metathesaurus on a locally installed database

T. Frankewitsch, H.U. Prokosch
Department of Medical Informatics and Biomathematics, University of Muenster, Germany

Abstract. Knowledge in the environment of information technologies is bound to structured vocabularies. Medical data dictionaries are necessary for uniquely describing findings like diagnoses, procedures or functions. Therefore we decided to locally install a version of the Unified Medical Language System (UMLS) of the U.S. National Library of Medicine as a repository for defining entries of a medical multimedia database. Because of the requirement to extend the vocabulary in concepts and relations between existing concepts a graphical tool for appending new items to the database has been developed: Although the database is an instance of a semantic network the focus on single entries offers the opportunity of reducing the net to a tree within this detail. Based on the graph theorem, there are definitions of nodes of concepts and nodes of knowledge. The UMLS additionally offers the specification of sub-relations, which can be represented, too. Using this view it is possible to manage these 1:n-Relations in a simple tree view.

On this background an explorer like graphical user interface has been realised to add new concepts and define new relationships between those and existing entries for adapting the UMLS for specific purposes such as describing medical multimedia objects.

1. Background

Knowledge may be described as a set of facts which have to be seen in a certain context. Those facts are represented as causal relationships between -elements, procedures and attributes which gain their validity in a certain context. One example in medicine may be, that an increase of blood pressure to 180/100 mmHg is pathologic for a resting person but physiologic for a sportsman during his training period.

Therefore many knowledge bases are based on relations between single items, and one favorite model is the semantic network. [1] The basic version of such a network consists of three relational tables. One table consists of the items, which are generally called concepts or meanings (concepts-table) and define the basic facts. The second table defines a set of possible relations (relations-table), and the third table finally comprises the knowledge statements in quadruples of concept, relation, concept and context (relationship-table). For example: "cell" - "is child of" - "tissue' - "anatomy".

The primary realization of such a knowledge base is programmed in relational databases, although in the recent time object-oriented approaches have been tested, too. The benefit of the object-oriented design is in the quick access of one single concept and its complete "environment", but there is no advantage in accessing multiple concepts [2].

In the relational version the relationship-table is a specialty of the entity-attribute-value-model. This model is widely used in programming techniques. In the relationship-table of the semantic network the entities show a recursive definition:

relationship-table = $\{(a,b,c,d) \mid a,c,d \in$ concepts-table; $b \in$ relation-table$\}$.

Using the EAV-version in database design it provides a high flexibility. There are no limits on the number of attributes for an entity and the number of relations is not restricted. Additionally it provides a simple physical data format. [3]

One of the main problems of this representation is the difference between logical and physical schemata. In normalized relational databases the information can be easily retrieved through tables with fixed counts of columns. But in the EAV-model the structure of database

is inverted. One approach for a user friendly data-access might be, that the "user interface ... (is) creating the illusions of conventional data organization" [3].

Another possibility for describing the semantic network is the graph theory [4]. The knowledge is represented as a graph where the nodes are corresponding to the concepts and the arcs to the relations between them. In this point of view a generic semantic network can be presented as a labeled, directed and unilaterally connected simple graph: The net is reduced to its nodes and arcs.

An example for a semantic network is the metathesaurus of the Unified Medical Language System (UMLS) of the U.S. National Library of Medicine [5]. In its most recent version it contains 626,893 individual concepts and 1,362,823 different concept names. The sources for the concepts are about 50 different vocabularies and classifications. The relationship-table provides the knowledge of 7,614,416 relationships.

Although there is a huge amount of information entries in this database, investigations have and must have shown, that there is the need for further entries [6]. Planning an image and multimedia database [7] we found missing data, which are useful for describing the multimedia files, for example "enlargement" of images or image-formats.

On one hand the NLM offers the use of a knowledge server, which enables retrieving concepts, on the other there is the need for locally defined entries. The licence agreement only forbids altering existing entries but not adding new ones. Therefore we developed a tool for graphically navigating through the UMLS entries on a locally installed database and for adding new concepts and relations in this instance.

2. Methods

The UMLS concepts and relations have been stored in an Oracle8i database using the database scheme of the UMLS text-files MRCON, which contains the main concepts, and MRREL, which is a file of relationships. The MRCON table has been extended with one additional column, which contains the textual representation of the entries in a slightly standardised manner: all entries are in minor letters and German special letters have been changed (e.g. 'Ä' → 'ae'). The purpose of this pre-processed standardisation was to provide the possibility of indexing and searching in a single column without the need of an intra-search-time modification of each entry. This uniformed column was additionally indexed with the Oracle8i capabilities of full-text search. Using this index it is possible to use for example a fuzzy scheme for finding misspelled words (e.g. appedicitis → appendicitis).

The UMLS is using three independent indexes, one for the words (SUI), one for modifications of an entry (LUI) and one for the concept (CUI). These indexes are managed by the program depending on the action of the user. To avoid problems of compatibility with new indexes of the NLM local entries are marked with a locally defined leading prefix ('M' for 'Muenster'). For example, the original index term for concepts of the UMLS is "C1234567", the local one "M"+"C123456". Therefore 999,999 individual concepts could be created. On this database design the following methods have been defined and implemented within a graphical user interface:

- "Dropping" an existing entry (a) on another (b) enables the definition of a new relation, selected by the user, and the inverse relation between them: a ↔ b, context 'USER'
- "Dropping" an existing entry (a) on any existing relationship of another (b) produces: a ↔ b, context 'USER'
- "Dropping" a new entry (a) on another (b) defines:
  CUI(a) = CUI(b), LUI(a) = LUI(b), SUI(a) = new SUI()

- "Dropping" a new entry (a) on the relation ("Idents"), which comprises all synonyms of the chosen entry (b) defines:
  CUI(a) = CUI(b), LUI(a) = new LUI(), SUI(a) = new SUI()

The database connection is realised with an ODBC bridge. The program itself is written with Delphi5 from Borland. This programming language is object-oriented, supports database-access and enables a rapid application development for the windows platform.

table 1: example of relationships.

data node	knowledge node		data node
concept	**relationship**	**attribute**	**concept**
'Cell'	'is child of'	'no other specification'	,Topographic Regions and cellular structure'
'Cell '	'is child of'	'no other specification'	'Anatomical structure'
'Cell '	'is child of'	'no other specification'	'Cellular and subcellular structure'
'Cell '	'is child of'	'no other specification'	'tissue'
'Cell '	'is parent of'	'constitutes'	'Organoid'
'Cell '	'is parent of'	'contains'	'Cellular inclusion'
'Cell '	'is parent of'	'has part'	'Fraction, subcellular'
'Cell '	'is parent of'	'has part'	'Micronucleus'
'Cell '	'is parent of'	'has part'	'Cytoplasm'
'Cell '	'is parent of'	'has part'	'Cell nucleus'
'Cell '	'is parent of'	'has part'	'Cell membranes'
'Cell '	'is parent of'	'inverse is a'	'pericyte'
...			

The UMLS semantic network offers the possibility to further specify simple relations, as it is illustrated in table 1, but only for a minority of the entries those relationship attributes have been defined. In the cases, where no relationship attribute is maintained by the UMLS, the attribute 'no other specification' = 'NOS' has been inserted for our local implementation.

Since this table-design, which is the result of a query within a relational table, lacks in clarity in order to understand the hierarchies in this semantic network, the "node and arc"-design seems to be adequate. One very well known implementation of nodes and arcs is the model of the explorer established within the windows file-system.

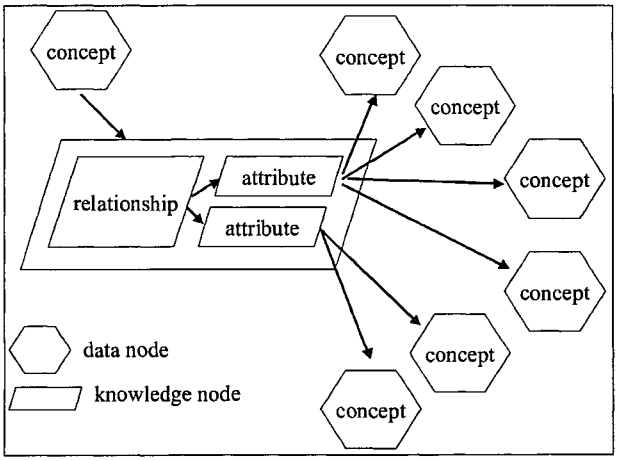

fig. 1: graph design

The graphical user interface modifies the former graph theory on semantic networks [4]. In the former view all entries are nodes and all relations are represented by arcs. This works well for all relations which are unique for a single node. But in a large data repository there are many 1:n pairs, having the same relationship, thus increasing the complexity to visualise those relationships.

For this purpose the basic graph theorem can be expanded. Normally each node represents a data entry, which in UMLS is called concept. The arcs display the relations. To enable a clearly arranged view there is a benefit, if the relations are nodes of themselves like a set of knowledge nodes. The knowledge nodes consists of two types: the main relation and its modifier or attribute (see fig. 1)

This design allows the user to select groups of arcs for a directed navigation. On the level of the node it can be seen as the root of a navigation tree: the central node has been moved to the left upper corner. If the user has started with the concept 'cell' it is possible to select the main knowledge node, for example 'PAR' for all 'is parent of'-entries of 'cell'. This knowledge node offers seven entries, the specialised (e.g. 'constitutes') and one, which contains the other not nearer specified entries ('NOS'). Selecting the 'has part' subsequent node, these concept nodes are opened. For example the user can now further 'explore' the knowledge nodes of 'Cytoplasm'. A screen shot of the test implementation is shown in fig. 2.

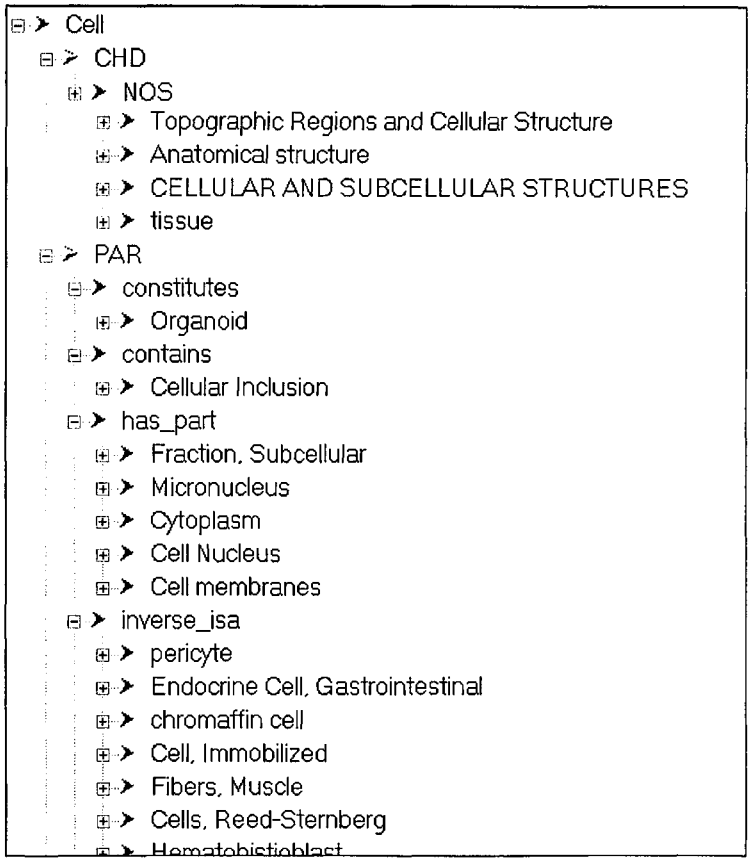

fig. 2: screenshot of the prototype

3. Discussion

In many medical applications a controlled and inter-application-wide valid data dictionary is required. The main problem of such a dictionary is its maintenance. Therefore one should integrate nationally and internationally agreed vocabularies in local systems as much as possible. One such vocabulary is the UMLS of the U.S. National Library of Medicine, which combines and maps many concepts from different data dictionaries, classification systems and nomenclatures. This database is not yet commonly used within medical applications in Germany, which may be due to the fact, that the majority of terms within the UMLS are English (even though since a few years other languages have been added and the Mesh-Terms for example are included in a German translation as well). A second reason for not applying the UMLS may be the handling of such a large concept repository with the need of adapting it to local circumstances. This always involves the integration of locally defined concepts.

In our project however, we have shown, that with the use of a graphical user interface it is possible to adapt such a database for the needs of a single institution and for specific tasks, such as the description of medical images. The tree view representation which has been implemented in our project offers the impression of interconnections between concepts on a commonly accepted design pattern. Drag and drop functions for adding new concepts and linking them to existing ones are intuitive and fast even for an untrained user.

One problem which however has not yet been solved is the central definition of concepts which might interfere with locally defined terms. Updates will always enforce the local user to recheck his entries but because of the well defined indexes the database administrator can update his 'old' indexes with the new one and inform the depending tables of the changes. In our opinion this might still be less work compared to the effort to completely create and fill a new medical data dictionary.

Acknowledgement

This work has been funded by the medical faculty of the university of Muenster, Germany, within the project "innovative researches in medicine" (Project Fi-1-2-II/96-17), within the HSP-III program of the "Ministry for schools, higher education, science and research of the state of North Rhine-Westphalia" and by SUN Microsystems.

References

[1] H.U. Prokosch, F. Amiri, D. Krause, G. Neek, J. Dudeck: A semantic network model for the medical record of a rheumatology clinic, Medinfo 8 (1995) 240-244

[2] H. Gu, Y. Perl, J. Geller, M. Halper, L. Liu, J.J. Cimino: Modeling the UMLS Using an OODB, Proc AMIA Symp. 1-2 (1999) 82-6.

[3] P.M. Nadkarni, L. Marenco, R. Chen, E. Skoufos, G. Shepherd, P. Miller: Organization of Heterogenous Scientific Data Using the EAV/CR Representation, JAMIA 6 (1999) 478-493

[4] W. Ruan, T. Bürkle, J. Dudeck: An object-oriented design for automated navigation of semantic networks inside a medical data dictionary. Artif Intell Med 18 (2000): 83-103

[5] U.S. Department of Health and Human Services: National Library of Medicine - UMLS Knowledge sources, January 1999, 10. Edition

[6] H. Goldberg, D. Goldsmith, V. Law, K. Keck, M. Tuttle, C. Safran: An evaluation of UMLS as a controlled terminology for the Promle List Toolkit, Medinfo 9 (1998) 609-612

[7] T. Frankewitsch, U. Prokosch Multimedia explorer: image database, image proxy-server and search-engine. Proc AMIA Symp. 1-2 (1999) 765-769