# Designing a Sub-Set of the UMLS Knowledge Base Applied to a Clinical Domain : Methods and Evaluation

A. Burgun[1], M.D., D. Delamarre[1], G. Botti[2], M.D., B. Lukacs[3], M.D, D. Mayeux[4], M.D., M. Bremond[5], M.D., F. Kohler[4], M.D., M. Fieschi[2], M.D., Ph.D., P. Le Beux[1], M.D., Ph.D.
1 Laboratoire d'Informatique Médicale, C.H.U., F-35033 Rennes, France. 2 Service de l'Information Médicale, C.H.U. la Timone, F-13385 Marseille , France. 3 Service Urologie, Hôpital Tenon F-75020 Paris, France. 4 Département d'Information Médicale, C.H.U., F-54035 Nancy, France. 5 Groupe Image, Ecole Nationale de la Santé Publique, F-94410 St Maurice, France

*The UMLS is a complex collection of interconnected biomedical concepts derived from standard nomenclatures. Designing a specific subset of the UMLS knowledge base relevant to a medical domain is a prerequisite for the development of specialized applications based on UMLS. We have developed a method based on the selection of the appropriate terms in original nomenclatures and the capture of a set of UMLS terms that are linked to them in the network to a certain degree. We have experimented it as the foundation for a concept base applied to urology. Results depend on the exhaustiveness of the relationships between the Metal concepts. A preliminary analysis of the sub-base reveales that some adaptations of vocabulary and ontology are required for clinical applications.*

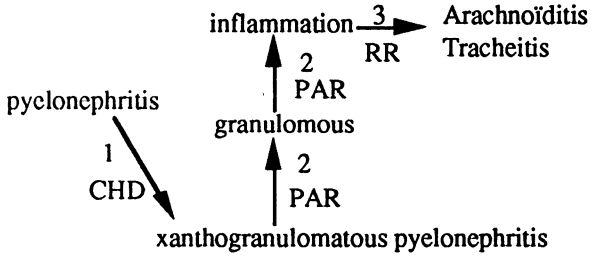
## INTRODUCTION

The UMLS is a vast repository of semantically categorized terms [1]. Designing a specific subset of the UMLS knowledge base relevant to a medical domain is a prerequisite for the development of specialized applications based on UMLS. We have developed a method to assist the researcher in that task.

## METHODS / RESULTS

The method is based on the scanning of nomenclatures and the exploitation of the UMLS interconcept links . The first step consists in selecting the appropriate terms within some of the nomenclatures that have been incorporated in UMLS. During the second phase, the initial set of concepts is extended in order to build a consistent subset of the UMLS knowledge base. The process starts by locating in the Metal the concepts corresponding to the selected nomenclature codes. This operation enables one to integrate the information - stemming from any source of the UMLS - related to those concepts, and to free oneself from the structure of the original nomenclature in order to exploit the fulness of UMLS. Then the different kinds of links within the UMLS network are explored as follows : (1) addition of all the children -all i.e. at all levels- of the concepts (2) addition of all the parents (3) addition of the related concepts (4) addition of the parents of the related concepts. The programs run on UNIX platforms as C-Shell procedures that operate on the ASCII files of the CD-ROM UMLS Sources distribution.

That algorithm was applied to 47 initial concepts selected from MeSH and SNOMEDII in order to build a base of urology concepts. The final vocabulary includes 984 concepts, 1905 terms and 3030 interconcept relationships. The wealth of the resulting sub-base is a function of the exhaustiveness of the relationships within Metal. That is the main limitation of our method. Noise is generated by the capture of terms from other medical domains. Among the 984 concepts of our urology sub-base, 681 (69%) were found to be relevant whether they were specific or not. The below figure shows how the selection of *Pyelonephritis* at the first step brings out *Arachnoïditis*.




## DISCUSSION

About the method : (1) The consistency of the sub-base is guaranteed since it represents the projection of the UMLS knowledge base on a specific domain. (2) It is well-adapted for the implementation of successive editions of the UMLS sources. (3) The sub-base is built up easily. (4) Further developments are required to reduce the noise.

About the resulting sub-base : (1) Because of the intended use of UMLS to facilitate information retrieval, the vocabulary may require some adaptations for clinical applications : clinical distinctions are missing *-Cryptorchism* and *Ectopic testis* are synonymous in Metal. (2) Ontology may be adapted; for example a category *surgical approaches* is needed for the description of surgical procedures.

## AKNOWLEDGEMENTS

The authors thank the N.L.M. for having provided the UMLS Knowledge Sources on CD-ROM.

Reference

[1]. Lindberg D.A.B., Humphreys B.L., Mc Cray A.T. The Unified Medical Language System, Methods of Information in Medicine, 1993; 32:281-91