An Ontological Analysis of the UMLS Metathesaurus™

Domenico M. Pisanelli, Aldo Gangemi, Geri Steve
Istituto di Tecnologie Biomediche - CNR Roma, Italy

Paper-based terminology systems cannot satisfy anymore the new desiderata of healthcare information systems: the demand for re-use and sharing of patient data, their transmission and the need of semantic-based criteria for purposive statistical aggregation. The unambiguous communication of complex and detailed medical concepts is now a crucial feature of medical information systems. Ontologies can support a more effective data and knowledge sharing in medicine. In this paper we briefly survey our ontological analysis and integration of various top-levels of terminologies and we report the main results of the ontological analysis of the UMLS Metathesaurus™.

INTRODUCTION

Physicians developed their language in order to reach an efficient way to store and communicate general medical knowledge and patient-related information. This language was appropriate for the only support available for archiving, processing and transmitting knowledge: the paper.
Paper-based terminology systems cannot satisfy anymore the new desiderata of healthcare information systems, such as the demand for re-use and sharing of patient data, their transmission and the need of semantic-based criteria for purposive statistical aggregation (i.e. different aggregation criteria for different purposes). The unambiguous communication of complex and detailed medical concepts (possibly expressed in different languages) is now a crucial feature of medical information systems.

Unfortunately such a task is not an easy one to be achieved and a deep analysis of the structure and the concepts of medical terminologies is needed. Such analyses can be performed by adopting an *ontological* approach for representing medical terminology systems and for integrating them in a medical ontology.
The role of ontologies for allowing a more effective data and knowledge sharing is widely recognized (see for instance Neches[1] and Guarino[2]).
Apart from its definition in a philosophical context - where it refers to the subject of existence - ontology in our context is "a partial specification of a conceptualization"[3]. Recently Sowa proposed the following definition influenced by Leibniz[4]:

> «The subject of ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called an ontology, is a catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D. [...] »

Actually there is some disagreement on what is an ontology. Some admit informal descriptions and hierarchies, only aimed at organizing some uses of natural language; others require that an ontology be a *theory,* i.e. a formal vocabulary with axioms defined on such vocabulary, possibly with the help of some axiom schema, as in description logics (for a position see Hayes[5]).
In our perspective, an ontology is a formal theory which partially specifies the conceptualization (i.e. the intended meaning) of a lexical item as it is used in a certain domain. Since lexical items are often used with more than one conceptualization in the same domain (they are "polysemous"), such different conceptualizations have to be specified and segregated within different formal contexts, or conceptualizations must have assigned distinct names within the same context. A “context” is a theory which serves as a module within a system which allows a partial ordering among its component theories.
The procedure by which the lexical items from a terminology system are conceptually analyzed and their conceptualizations are (partially) specified within a context hierarchy is what we call the "ontological analysis" of a terminology.
The sources of the ontological analysis in our project are medical terminology systems. Our analyses aim at explicitating the implicit relationships among the conceptualizations of the lexical items (“terms”) included in the sources, and maintaining the reference of such relationships to a set generic theories.
In this paper we give some hints about the ontological analysis that we perfomed on some top-levels of medical terminology systems and we report the current main results of the ontological analysis that we perfomed on the UMLS Metathesaurus™[6].

TOP-LEVELS INTEGRATION

Past experiences show that an explicit conceptualization of a terminology needs to be philosophically and linguistically grounded[7-8-9]. Not everyone recognizes the relevance of generic (domain-independent) theories to the development of ontologies (see various papers in Guarino[2]). Examples of generic theories include: "mereology" or theory of parts, "topology" or theory of wholes and connexity, "morphology", or theory of form and congruence, "localization" theory, "time" theory,

"actors" theory, etc.
Our position is that generic theories are essential to the development of ontologies and to a rigorous conceptual integration of heterogeneous terminologies. Generic theories should not necessarily be formal, nevertheless a formal theory is more easily discussed, specially if it is accompanied by a rich informal documentation. Currently there are sophisticated systems which provide services, such as formal contexts and concept classification, which greatly help the development of domain theories by specializing generic theories.

We developed ONIONS[9], a methodology for integrating domain terminologies by exploiting a “library” of generic theories. ONIONS was defined in order to build the core model of a medical terminology server in the context of the GALEN Project[10]. Later this methodology was revised with the goal of building and re-using a library of generic theories by formalizing ontologies from the literature in AI, philosophy, linguistics, cognitive science. Such theories have been represented in the Ontolingua language and are partly available on our web site (http://saussure.irmkant.rm.cnr.it).
ONIONS methodology has been applied for analyzing and integrating the following top-levels of medical terminology systems: the UMLS Semantic Network[6] (1997 edition: 135 'semantic types', 91 'relations', and 412 'templates'), the SNOMED-III[11] top-level (510 'terms' and 25 'links'), GMN[12] top-level (708 'terms'), the ICD10[13] top-level (185 'terms'), and the GALEN Core Model[10] (2730 'entities', 413 'attributes' and 1692 terminological axioms).
Conceptual integration in ONIONS has been carried out as follows: all terms, templates, and axioms have been formally represented. When available, natural language glosses have been axiomatized; such intermediate products have finally been integrated by means of a set of generic theories (e.g. "topology", "mereology"). We experimented a web-based tool for cooperative modelling; different modellers could experiment and face each other about the effects of ontological analysis on terminology integration[14]. For a deeper explanation of the problems, considerations and methods used in the integration, see[15]. For a complete presentation of the methodology, see[9].
An example of the outcome of such integration activity is the formalization of the concept "Body-Region" resulting from the UMLS Semantic Network, GALEN Core Model and generic theories "topology", "meronymy", "localization". Figure 1 reports such formalization expressed in the Loom language[16]. In its definitional axioms (i.e. `:is-primitive`) the formula states that a "Body-Region" is a "Region" whose location is a "Body-Part" or a "Tissue" and it is portion of an "Organism". There follow some implicational axioms (i.e. `:implies`) which are not classified, but only semantically checked (for a detailed discussion on this example, see[15]).

```
(defconcept Body-Region
  :is-primitive (:and Region
                 (:some whole-location-of
                  (:or Body-Part Tissue))
                 (:some portion Organism))
  :implies (:and (:some connected Body-Region)
                 (:some component (:or Body-System Body-Part))
                 (:all near (:or Body-Region Body-Space Body-Part))
                 (:all context-of (:or Biologic-Function Injury Poisoning))
                 (:all crosses-through Body-Region))
  :context anatomy)
```

Figure 1.
The concept "Body-Region" formalized in Loom.

```
(defconcept Fibromyalgia
  "UMLS-CUI C0016053"
  :is-primitive (:and Disorders-of-the-muscles-ligaments-fasciae-and-other-soft-tissues
                      Muscle-functions-and-symptoms
                      Myalgias/Myopathy
                      Muscular-Diseases
                      Rheumatic-Diseases
                      Disease-or-Syndrome))
```

Figure 2.
An example of the formalization of the concept "Fibromyalgia" in Loom.

THE ONTOLOGICAL ANALYSIS OF THE UMLS METATHESAURUS™

Apart from the generic theories and the top-levels of medical terminologies, our activity is aimed also at integrating some relevant "bottom-level" medical terminologies.
We started from the Metathesaurus™[6], developed in the context of the Unified Medical Language System (UMLS) project by the U.S. National Library of Medicine[17]. It is a significant starting point, since it collects millions of terms belonging to the most significant nomenclatures and terminologies defined in the United States and in other countries too. Such feature makes it a proper object of analysis and reuse, being it probably the largest repository of terminological knowledge in medicine. However we do not plan to limit ourselves to its ontologization, since other important sources are worthwhile to be integrated in our ontology. To this aim, it is very relevant the work carried out by Spackman and colleagues concerning the formal representation of SNOMED in a description logic based formalism[18]. We envisage an integration between the two representations. Being both SNOMED-RT and our ontology expressed in similar formalisms, such integration would profit from a set of formal tools for automatic consistency verification.

When we started our work, the 1998 edition of the Metathesaurus was not yet available. However, even if the analysis is being updated, the methodology adopted and most of the results are still valid.
Among the various sources, the National Library singled out about 330,000 “Concept Unique Identifiers” (CUIs) chosen as representative of a set of synonyms and lexical variants (only in English at the beginning of the project, but currently including Spanish, French, German, Russian and Portuguese). Campbell and co-workers report some cases of over-normalization (i.e. not true synonyms under the same CUI) and point out that a CUI has only an extensional meaning, whose referents are the terms taken from the source terminologies of the UMLS[19]. This means that if a term or set of terms is polysemous, the CUI is polysemous as well. However, CUIs provide official codes and preferred names which allow to restrict the medical lexicon usually without losing specificity.

Starting from the Metathesaurus we built a database featuring:

1) the preferred names of the CUIs (e.g. "Fibromyalgia");
2) the instances of IS_A relations between different CUIs that UMLS mutuated from its sources (e.g. "Fibromyalgia" IS_A "Muscular-diseases");
3) the instances of IS_A relations between a CUI and its "semantic types" (e.g. "Fibromyalgia" IS_A "Disease-or-Syndrome");
4) the definition of the CUIs in plain text, as reported in authoritative sources such as medical dictionaries.

The database is implemented in MS Access and re-arranges the original tables provided by the NLM in four tables:

1) *constr*, featuring 331,756 records with the CUI and its preferred term;
2) *conpar*, featuring 51,814 records with the couples CUI-child CUI-parent and the terminology source of the parenthood;
3) *consty*, featuring 443,770 records reporting the semantic type or types for every CUI;
4) *condef*, featuring 21,895 records reporting the definitions of CUIs if available;

It should be pointed out that UMLS defined a parent CUI only for a minority of CUIs, usually mutuating the parents from the titles of classification sections (e.g. "Bronchial-Diseases"). On the contrary every CUI has one or more semantic types, therefore about 443,000 pairs of CUI - semantic type have been defined.
A computer program generated an expression in the Loom representation language for each CUI in the database. Apart from the natural language definitions, all the other information in the database were represented and subsequently classified. As an example, figure 2 reports the outcome for the concept "Fibromyalgia".
The syntax of such expression is typical of the applications running on top of the Lisp language. The semantics is quite intuitive to capture, it simply states that the IS_A relation holds for every pair "Fibromyalgia" – concept belonging to the list reported in the expression. The semantics of the IS_A relation is that of the inclusion of classes, as used in description logic. In other words when we say that "Fibromyalgia" IS_A "Muscular-Diseases", we say that the former class is included in the latter class. Such a relation is analogous to that holding between subclasses and superclasses in the object-oriented paradigm.

Having all the 331,756 CUIs available in the appropriate formalism, the Loom toolkit was employed for classifying them. Classifiers are able to organize and possibly re-arrange hierarchies of concepts by reasoning about their definitions. .
As an example of the re-arrangement performed, figure 3 shows the concept "Fibromyalgia" after classification.

```
? (pc 'fibromyalgia)

(defconcept Fibromyalgia
  "UMLS-CUI C0016053"
  :is-primitive (:and Rheumatic-Diseases
                      Muscular-Diseases
                      Myalgias/Myopathy
                      Muscle-Functions-And-
                                  Symptoms))
```

Figure 3.
The concept "Fibromyalgia" after classification.

Some "ancestors" have disappeared now. In its inherent parsimony, the classifier omits those concepts for which the fact of being more generic can be transitively found, leaving only the immediate "parents".
Classification allowed also to detect several cycles (about 100) in the source definitions. For example UMLS correctly states that "simple goiter" has "goiter" as a parent CUI, but elsewhere it states also that "goiter" has "simple goiter" as a parent CUI in the context of an enumeration of the different kinds of goiters.
In other places, cycles are due to the presence of partial concept overlapping (for example: "eczema" and "dermatitis"). In such cases, the choice of preferred terms was evidently uncertain. Ontological modelling helps distinguishing the cases in which overlapping concepts can be merged from the cases in which the definitions have to be kept disjoint.

Another problem concerns the mis-use of some terminological hierarchies to express generic term association or partonomy instead of inclusion. For example, "infertility" has "fertility" as a parent CUI (this actually is a generic association), and "social isolation" has "sociology" as a parent CUI (this actually is an "issue-in" relation). This is a major point, and currently ONIONS methodology is being applied to make explicit the relations underlying such pseudo-parenthood.
Beyond such quite evident mis-use of terminological hierarchies, which is recognized even by UMLS authors themselves (in the introduction to the documentation[6]), there are more subtle issues which originate from the *polysemous* use of medical terms and constitute the main focus of ontological analysis. In the next paragraph we report how the classifier facilities support the ontological analysis of the Metathesaurus.

DIVIDE ET IMPERA: PATTERN OF SEMANTIC TYPES

Concept Unique Identifiers (CUIs) with more than one semantic type are very frequent in the Metathesaurus. About 90,000 of them (almost one third of the corpus) exhibit this feature and have a number of semantic types ranging from two to five. This may be due to the inherent polysemy of these terms, whose real meaning is determined by the context in which they are used. For example: "Salmonella-Choleraesuis" is classified both under "Disease-Or-Syndrome" and "Bacterium"; "Onychotillomania" is "Sign-Or-Symptom", "Individual-Behavior" and "Mental-Or-Behavioral-Dysfunction". There are about 900 different combinations (*patterns*) of semantic types which occur in the Metathesaurus, 1997 edition (table 1). They have been singled out by means of the classifier's facilities.

Disease-Or-Syndrome	24610
Disease-Or-Syndrome Acquired-Abnormality	606
Disease-Or-Syndrome Anatomical-Abnormality	352
Disease-Or-Syndrome Classification	15
Disease-Or-Syndrome Congenital-Abnormality	1169
Disease-Or-Syndrome Finding	379
Disease-Or-Syndrome Injury-Or-Poisoning	827

Table 1.
Some patterns of semantic types occurring in the Metathesaurus and number of concepts pertaining them.

The individuation of such patterns induces a partition in the Metathesaurus and facilitates its ontological analysis. This is a sort of *divide et impera* approach (i.e. "divide and rule"), since CUIs sharing the same pattern of semantic types are supposed to have similar features and can be analyzed and formalized together.

For example, we analyzed the pattern of semantic types "Finding" and "Injury-or-Poisoning". One of its CUIs has the preferred name "fractures, ununited". The classifier allows to detect all the IS_A relationships between "fractures, ununited" and its parents and semantic types.
Being ontological analysis and integration aimed at supporting clear identity criteria, such graph puts in evidence several ontological problems. Is it ontologically acceptable that "fractures, ununited" is classified both under "Natural Phenomenon" and under "Injury or Poisoning", which is not a "Natural Phenomenon"? Is ontologically acceptable a concept which is classified both under "Phenomenon" and "Conceptual Entity"?
One may simply conclude that hierarchical assignments here have been decided with disregard of logical semantics. On the other hand, this would be a superficial judgment. In fact, UMLS assignments try to cover some possible polysemous senses of "fractures, ununited" without creating ad-hoc distinctions (e.g. "fractures, ununited-1", "fractures, ununited-2", "fractures, ununited-3", etc.).
An advantage provided by ontological analysis and integration is the possibility of treating such polysemy without multiplying the ad-hoc distinctions.

```
(defconcept fractures-ununited
  :is-primitive (:and fracture
                      (:some morphology-of (:and bone
                                                 (:or (:some embodies malunion)
                                                      (:not integral))))
                      (:some follows fracture-event)
                      (:all interpretant-of clinical-situation))
```

Figure 4.
Part of the Loom formalization of the concept "fractures, ununited" after an ontological analyis.

After the application of ontological analysis, "fractures, ununited" will be conceptualized as: "a fracture of a bone which (1) necessarily bears a malunion (a pathology causing a morphological imprecision) or a nonunion (a lacking of integrity), which (2) necessarily follows a primary fracture event, and which (3) contingently may be a sign of something else.
Therefore, such conceptualization shows only one classification (under "fracture") and three definitional axioms which provide the identity criteria for the instances of "fractures, ununited" (figure 4).

CONCLUSIONS

The need for standardization in health care was perceived independently from computer use, but telecommunications and networking are dramatically changing the scenario of knowledge management and data sharing.
Traditional terminology systems are not appropriate anymore to satisfy the demand for re-use of data, unambiguous transmission and statistical aggregation. An ontological approach to the description of terminology systems will allow a better integration and reuse of these systems.
One may wonder if UMLS can be considered a "super source" to be preferred to its component sources (SNOMED, ICD, etc.). We believe that it should not replace the original sources, because they often embed more information than that incorporated by UMLS. We started our ontological analysis from the Metathesaurus because it has normalized the lexical variants and most synonyms and it has related the CUIs to a great amount of additional information: natural language definitions, IS_A relations (about 42,000 in the 1998 edition), domain specific relations to other CUIs (about 50,000), unspecified relations to qualifiers and associable CUIs (about 627,000), explicit mappings between the component sources (about 94,000).
The ontological analysis of UMLS yielded a partition of CUIs according to the original UMLS semantic types. Such a partition allowed us to define contexts which refer to medical sub-domains and include a library of generic theories.
However we do not plan to limit our ontological analysis to UMLS. It is a good starting point, but it will be followed by analyses of other important sources like SNOMED, for which the top-levels have already been analyzed.

References

1. Neches R et al., "Enabling Technology for Knowledge Sharing", *AI Magazine*; 3, 1991.
2. Guarino N (ed.), *Formal Ontology in Information Systems*, Amsterdam, IOS-Press, 1998.
3. Guarino N, *Formal Ontology and Information Systems*, in 2.
4. Sowa J, communication to the *ontology-std* mailing list, 1997.
5. Hayes P, note on the meaning of "ontology", http://ksl-web.stanford.edu/email-archives/srkb.messages/647.html.
6. National Library of Medicine, *UMLS Knowledge Sources*, 1997 edition, available from the NLM, Bethesda, Maryland.
7. Guarino N, Carrara M, Giaretta P "An Ontology of Meta-Level Categories" In J Doyle, E Sandewall and P Torasso (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of KR94*. San Mateo, CA, Morgan Kaufmann, 1994.
8. Sowa JF "Top-Level Ontological Categories" *International Journal of Human-Computer Studies*, 43, 1996.
9. Steve G, Gangemi A, Pisanelli DM, "Integrating Medical Terminologies with ONIONS Methodology", in Kangassalo H, Charrel JP (eds.) *Information Modelling and Knowledge Bases VIII*, Amsterdam, IOS Press 1997.
10. Rector A, Solomon WD, Nowlan WA, Rush T, "A Terminology Server for Medical Language and Medical Information Systems", *Methods of Information in Medicine*, 34, 1995.
11. Coté RA, Rothwell DJ, Brochu L, eds. *SNOMED International* (3rd ed.), Northfield, Ill, College of American Pathologists, 1994.
12. Gabrieli E, "A New Electronic Medical Nomenclature", *J. Medical Systems*, 3, 1989.
13. WHO, *International Classification of Diseases* (10th revision), Geneva, WHO, 1994.
14. Pisanelli DM, Gangemi A, Steve G, "WWW-available Conceptual Integration of Medical Terminologies: the ONIONS Experience", *Proceedings of AMIA 97 Conference*, 1997.
15. Gangemi A, Pisanelli DM, Steve G, "Ontology Integration: Experiences with Medical Terminologies", in 2.
16. MacGregor RM, "A Description Classifier for the Predicate Calculus" *Proceedings of AAAI 94, Conference*, 1994.
17. Humphreys BL, Lindberg DA, "The Unified Medical Language System Project", *Proceedings of MEDINFO 92.*, Amsterdam, Elsevier, 1992.
18. Spackman KA, Campbell KE, Coté RA, "SNOMED RT: A reference terminology for health care", *Proceedings of AMIA 97 Conference*, 1997.
19. Campbell KE, Oliver DE, Shortliffe EH, "The Unified Medical Language System: Toward a collaborative approach for solving terminologic problems", *JAMIA*, 5(1), 1998.