Ontology Integration: Experiences with Medical Terminologies

Aldo Gangemi, Domenico M. Pisanelli, Geri Steve
ITBM-CNR, V. Marx 15, 00137, Roma, Italy
{aldo,nico}@color.irmkant.rm.cnr.it steve@itbm.rm.cnr.it

ONIONS is a methodology for ontology analysis and integration. It has been applied to some relevant and very large medical terminologies (e.g. the 332,000 medical concepts singled out by the National Library of Medicine in the UMLS project).
Current results include the alignment of the top-level ontologies of the terminologies considered. The paper reviews the formal and conceptual tools employed in this task, presents the most significant results obtained, and discusses two case studies.

1. Introduction

ONIONS methodology for ontology integration has been developed since 1992 to account for the problem of conceptual heterogeneity [12][37][38]. It also addresses some problems encountered in the context of the European project GALEN [11] and the Italian projects SOLMC (Ontologic and Linguistic Tools for Conceptual Modelling) [16] and ONTOINT (see the URL: http://saussure.irmkant.rm.cnr.it/onto/ontoint.html).

It is being applied in a project which has achieved some results in ontology integration of medical terminologies [31]. Aims of ONIONS include:

1) developing a well-tuned set of generic ontologies to support the integration of relevant domain ontologies in medicine. In fact, current medical ontologies lack axiomatization, or semantic precision, or ontological cleverness;
2) the integration of a set of relevant domain ontologies in a formally and conceptually satisfactory ontology to support many tasks, including information retrieval, natural language processing, computerized guidelines generation, data bases integration, etc.
3) providing an explicit tracing of concept mappings, constraints and choices in ontology building, in order to allow extensions and/or updating.

The tools of ONIONS include a set of formalisms, a set of computational tools which implement and support the use of the formalisms, and a set of generic ontologies, taken from the literature in either formal or informal status and translated or adapted to our formalisms.

The main products of ONIONS are: the ON9 library of generic ontologies; the IMO (Integrated Medical Ontology, that represents the integration of five medical top-levels of relevant terminologies, and the relative mappings); a formalized representation of some medical repositories (mainly the UMLS Metathesaurus™ defined by the U.S. National Library of Medicine) with their classification within the IMO.

2. Basic definitions

The current debate about ontology integration and related notions such as "alignment", "compatibility", "equivalence", etc., is growing enough to prove that the disagreement persists at a deep, "ontological" level. Since a precise expression of our opinions on the matter would occupy too much space here (for a discussion see [39]), we present our intended meaning of such terms in a very pragmatic way, with some references only to underlying problematics.

In order to situate our intended meaning of ontology, we refer to a John Sowa's definition proposal on the ontology-std mailing list, which is manifestly influenced by Leibniz:

«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. The types in the ontology represent the predicates, word senses, or concept and relation types of the language L when used to discuss topics in the domain D.»

In our opinion, this means that we conceptualize for some purpose (task, goal) and in order to do it we have to use a language as a pre-existing tool. To call it L means that it has been stated which elements (words) are accepted in it; to fix a domain D helps restricting the possible interpretations (semantics) of the language L.

Words have at least one meaning, generally constituted by the informal and approximate intuition that we apply when we read or hear them in some particular conditions. This refers to the importance of the cultural context: if in our ontology we sensibly violate these cultural context conventions, the intuition goes too far from our intended meaning and we would probably be misunderstood.

Divergent views about context relativity and linguistic conventionality originate the current debate concerning the definition of notions such as "ontology integration", "compatibility of ontologies", "equivalence of ontologies", "ontology unification", and in the search for a so-called "reference ontology", which should provide context-neutral and language-neutral categories.

Following our assumptions, the goal of building context-neutral or language-neutral categories in the framework of a neutral or definitive "reference ontology", seems unrealistic compared to the complexity of real world and the hetereogeneity and creativity of cultures.

We are rather concerned with the goal of an ontology adequate enough to an actual task: we believe that the contingency of such an ontology is a benefit if it means openness and updatability to further ontological investigations.

For this and other reasons, we maintain that ontology building needs to explicit both task-dependecies and general-model references.

It is a matter of fact that all conceptualizations and their linguistic interpretations, in spite of their hetereogeneity, show some similarities; these quasi-invariants can be due to the real world structures or to the cognitive attitudes of humans or to our widespread culture, and to their interrelations.

In our opinion, to explicit the general-model references of our ontology means to explicitly refer our conceptual distinctions to some articulated model, accounting for general notions such as part, connection, localization, cause, form, granularity, judgment, etc. (see §6. for details). Such general models have to be extracted from authoritative literature about those notions and should not have to be intended as ultimate; on the contrary, conflicting models should serve the purpose of being exchanged, criticized, and negotiated among ontological engineers, in a way similar to the validation and evolution of alternative scientific models.

From the viewpoint of information systems, general models can be seen as a basis on which an interlingua for ontology integration can be built. The use of such an interlingua involves complex methodological and design issues, which cannot be easily summarized here. Thus we only include in the following some definitions in order to have a

framework in which to discuss if formal ontology can be, and it actually is, successfully used in the integration of terminologies.

2.1 Kinds of ontologies

The following is a classification of ontologies according to the level of explicitness and formalization:

- *catalog of normalized terms*, e.g. a list of terms used in the reports from a laboratory: no inclusion order, no axioms, no glosses;
- *glossed catalog*, e.g. a dictionary of medicine: a catalog with natural language glosses;
- *taxonomy*, e.g. the SNOMED taxonomy [6] or the UMLS Metathesaurus [28]: a collection of concepts with a partial order induced by inclusion;
- *axiomatized taxonomy*, e.g. the GALEN Core Model [11]: a taxonomy with axioms;
- *context* (or *ontology*) *library*, e.g. the CYC encyclopaedia [22]: a set of axiomatized taxonomies with relations among them (*inclusion* of a context into another one, or *use* of a concept from a context in another one).

2.2 Kinds of ontology modules

The following is a classification of ontology modules (formal contexts) according to generality (an elaboration of, among others, Guarino [15] and Van Heijst [42]:

- *representation* ontologies specify the conceptualizations that underly knowledge representation formalisms (see theory: *metaontology* in §5);
- *top-level* ontologies are a particular recipe of generic and intermediate ontology concepts either on top of a domain ontology, or stand-alone with claims of domain-independence. For example, the UMLS Semantic Network [19]: is a typical domain top-level, CYC top-level stays on top of a maximally comprehensive set of ontologies while the one of Guarino [17] is a stand-alone top-level.
- *generic* ontologies concern the general, foundational aspects of a conceptualization. See below the section on 'Conceptual tools' for examples.
- *intermediate* ontologies concern the general aspects of the conceptualization of a domain. For example, the GALEN Core Model is an extended top-level for medicine, but with characteristics more typical of intermediate ontologies than generic ones: loose axiomatization of most general concepts, no reference to generic ontologies, etc.
- *domain* ontologies specialize a subset of generic ontologies in a domain or subdomain. For example, the SNOMED taxonomy is an ontology of the medical domain, while our ontology of surgical procedures is a sub-domain ontology.

References to our ontologies (or 'theories' or formal 'contexts') quoted in this paper are intended to the following URL: **http://saussure.irmkant.rm.cnr.it**, in which a large part of our ontology library (§5.) is available for browsing.

2.3 Kinds of ontology design

The following is a classification of ontological systems according to design choices:

- *ad-hoc* ontologies: no explicit motivation of ontology design;
- *explicitly motivated* ontologies: explicit design motivation, possibly derived from the adoption of one or more generic ontologies.

2.4 Kinds of ontology integration

From our point of view, we could consider the level of integration as a continous quantity, however, in order to facilitate communication about this level, here we refer to three levels proposed by Sowa:

«**Integration:** the process of finding commonalities between two different ontologies A and B and deriving a new ontology C that facilitates interoperability

between computer systems that are based on the A and B ontologies. The new ontology C may replace A or B, or it may be used only as an intermediary between a system based on A and a system based on B. Depending on the amount of change necessary to derive C from A and B, different levels of integration can be distinguished:

• *Alignment* is the weakest form of integration: it requires minimal change, but it can only support limited kinds of interoperability. It is useful for classification and information retrieval, but it does not support deep inferences and computations.
• *Partial compatibility* requires more changes in order to support more extensive interoperability, even though there may be some concepts or relations in one system or the other that could create obstacles to full interoperability.
• *Unification* or total compatibility may require extensive changes or major reorganizations of A and B, but it can result in the most complete interoperability: everything that can be done with one can be done in an exactly equivalent way with the other.»

The distinctive criterion adopted by Sowa is that of *interoperability*. One may intend the interoperability as a computational property only, while someone else may assume that interoperability has to be grounded on conceptual integration at the ontological level; this means that any level of interoperability must be consequent to a level of conceptual integration and can provide an indirect, operative measure of it.

In the first approach, the interoperability between heterogeneous ontologies can be obtained by building computational, ad-hoc integrations; in the second approach it may be reached by analysing their concepts and building sharp and coherent definitions inside the framework of general knowledge: integrated concepts have to be referred to such definitons and consequently to the general knowledge.

Notice that in this paper, "ontology", "theory" and (formal) "context" are basically synonyms. Philosophers are often reluctant to use "ontology" in such a way, whereas some others are unwilling to use "theory". Moreover, a confusion is possible between formal and ontological "contexts": a *formal* context is a module in an ontology library, while an *ontological* context should be a generic concept axiomatized in a dedicated theory which covers notions such as "situation", "physical region", "temporal interval", "text", etc.

Consequently, we adopted a flexible approach: in this section on definitions derived mainly from AI literature, we use "ontology", as usual in this area; in §5., when our library is introduced, we use 'theory', which is the term used in Ontolingua (the language used in the library currently browsable from our site); in §6., when case studies are treated, we use 'context', which is the term used in Loom (the language used for the description logic version of our library).

3. Current status of the project

ONIONS methodology has been applied to the integration of the following medical terminologies: the UMLS top-level (1997 edition: 135 'semantic types', 91 'relations', and 412 'templates'), the SNOMED-III top-level (510 'terms' and 25 'links'), GMN [10] top-level (708 'terms'), the ICD10 [44] top-level (185 'terms'), and the GALEN Core Model [33] (2730 'entities', 413 'attributes' and 1692 terminological axioms). ONIONS has also been applied to the integration of various sub-domain catalogs and taxonomies.

Ontology integration in ONIONS has been carried out as follows: all concepts and axioms have been formally represented (see §4.); when available, natural language glosses have been axiomatized; such intermediate products have finally been ontologically integrated by means of a set of generic ontologies (see §5.).

For a practical explanation of the problems, considerations, and methods used in the integration, see §6. For a complete presentation of the methodology, see[38].

3.1 The UMLS Metathesaurus™ investigation

A special investigation is being made on the corpus of concepts from the UMLS Metathesaurus™. The National Library of Medicine (NLM) in the United States has collected several millions of medical terms from various sources and has singled out about 332,000 preferred terms in English in the context of the Unified Medical Language System (UMLS) project [28]. Preferred terms are chosen among synonyms and lexical variants and are labelled by NLM "concepts". Although such a definition of "concept" may seem too much granular, for the sake of our ontologization we will mantain the original definition.

Starting from the rough public-domain UMLS sources (made available on CD-ROM by the NLM) we built a database featuring:

1) the "concepts" (e.g. "acute bronchitis");

2) the instances of IS_A relationships between different concepts that UMLS mutuated from its sources (e.g. "acute bronchitis" IS_A "bronchial diseases");

3) the instances of IS_A relationships between a concept and its "semantic types" (e.g. "acute bronchitis" IS_A "disease or syndrome").

It should be pointed out that UMLS defined a parent concept only for a minority of concepts, usually mutuating the parents from the titles of classification sections (e.g. "bronchial diseases"). About 43,000 instances of IS_A relationships have been defined.

On the contrary, every concept has one or more semantic types, therefore about 443,000 pairs of concept - semantic type have been defined.

Starting from the database - which systematizes the UMLS definitions without further assumptions - for each concept we generated an expression in a formalism suitable for automatic classification (description logic). Here follows an example written in Loom (§4.):

```
(defconcept Acute-bronchitis-NOS
   "UMLS-CUI C0149514"
   :is-primitive (:and Acute-bronchitis-and-bronchiolitis
          Acute-lower-respiratory-tract-infection
          Disease-of-bronchus-NOS
          Bronchial-Diseases
          Respiratory-Tract-Infections
          Disease-or-Syndrome))
```

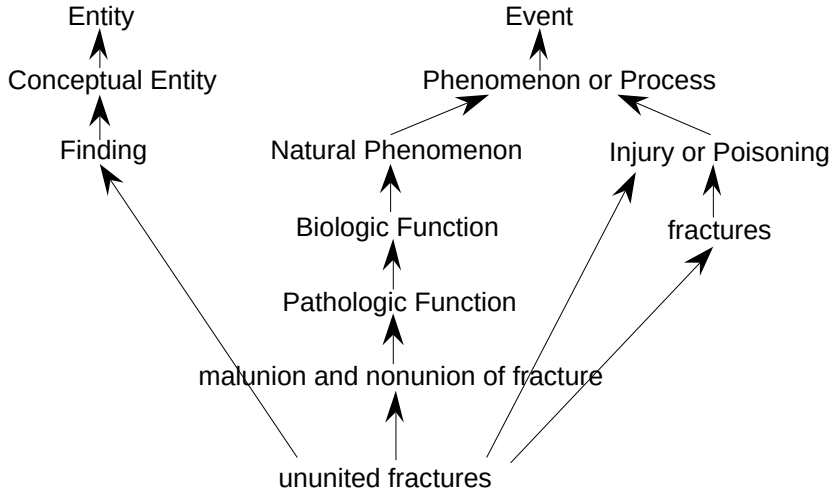
Such 331,776 expressions were automatically classified by Loom and put in evidence several problems with the source. It is worthwhile to point out that several cycles (about 100) were discovered in UMLS. For example UMLS correctly states that "simple goiter" has "goiter" as a parent concept, but elsewhere it states also that "goiter" has "simple goiter" as a parent concept in the context of an enumeration of the different kinds of goiters. Such an improper use of parent relationship is evident and the formal consistency of the classification is easily restored by cancelling the latter instance of the relationship.

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 concept hierarchies to express generic concept association or partonomy instead of inclusion. For example, "infertility" has "fertility" as a parent concept (this actually is a generic association), and "social isolation" has "sociology" as a parent concept (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 concept hierarchies, which is recognized even by UMLS authors themselves (in the introduction of the UMLS documentation [28]), there are more subtle and more ontologically interesting issues, which originate from the *polysemous* use of medical terms.

For example, the concept "ununited fractures" has the semantic types "Finding" and "Injury or Poisoning" (semantic types are denoted by capital letters), and the parents: "fractures" (whose semantic type is "Injury or Poisoning") and "malunion and nonunion of fracture" (whose semantic type is "Pathologic Function"). The following graph results (arrows mean IS_A):



Such graph puts in evidence several ontological problems, at least if ontological analysis and integration are aimed at support clear identity criteria (see also §6.). Is it ontologically acceptable that "ununited fractures" 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 "ununited fractures" without creating ad-hoc distinctions (e.g. "ununited fractures-1", "ununited fractures-2", "ununited fractures-3", etc.).

An advantage provided by ontological analysis and integration is the possibility of treating such polysemy without multiplying the ad-hoc distinctions. For example, after the application of ontological analysis, "ununited fractures" would be conceptualized as: "a fracture of a bone which (1) necessarily bears a malunion (a morphological imprecision) or a nonunion (a lacking of connectedness) at a given time after the primary fracture event, and which (2) contingently is a sign of something else. Therefore, such conceptualization shows only one classification (under "fracture") and two axioms which provide the identity criteria for the instances of "ununited fractures" (for details on formal and conceptual tools used in ontological analysis, see §5. and §6).

During our investigation, we discovered that polysemous concepts (i.e. concepts with more than one semantic type or parent) are very frequent. Just to give a hint, the allowed combinations of different semantic types (ranging from two to five), are about 600 and they account for about 100,000 Metathesaurus concepts (almost one third of the corpus).

For a full report about the UMLS ontology integration, see [13].

3.2 Evaluation of terminological sources

More generally, medical terminological sources showed either one or more of the following issues (see §6. for detailed examples):

- *lack of axioms*: for example, ICD10 shows nude taxonomies, without axioms or even a natural language gloss;
- *semantic imprecision* (cycles, relation range violation, etc.): for example, the semantic network used as the top-level of the UMLS Metathesaurus includes a set of templates for its taxonomy, but the semantics of such templates is unknown: after careful analysis, the most that can be done is considering UMLS templates as default axioms (see Case Study 1 in §6.);
- *ontological opaqueness* (lack of motivation for choosing a certain predicate, or lack of reference to an explicit, axiomatized generic ontology, or at least to a generic informal theory): for example, in the GALEN Core Model nearly all concepts and relations in the top-level are non-axiomatized and undocumented: they have been chosen with disregard of formal ontology: no trace of mereological, topological, localistic, dependence notions is retrievable (see Case Study 2 in §6.);
- *linguistic awkwardness in naming policy*: for example, in the GALEN Core Model, purely formal architecture considerations have originated a lot of redundancy and curious relation and concept names (see Case Study 2 in §6.).

It should be pointed out that even top-level ontologies 'on the market' for general purpose, such as CYC, Pangloss, etc., do not seem to be satisfactorily applicable to ontology integration. In fact Guarino and co-workers report that general-purpose top-levels are: 1) complicated, or 2) difficult to understand, or 3) confused (either in intended meanings, aspects, or meta-level categories) [17].

3.3 Current products of ONIONS

Thus, on the one hand, we wanted to integrate medical ontologies with a methodology which supports extensive axiomatization, clear semantics, and ontological cleverness; on the other hand, we collected an ontology *library* which includes, to our best knowledge, an optimal choice of generic ontologies available to support the modelling of the medical section of the library. This sometimes includes multiple choices among partially incompatible ontologies. Finally, we provided a "metaontology" which states the semantics of the meta-level categories that we adopted to distinguish among the concepts in our library (see §5.).

Current products of ONIONS include:

• the ON9 ontology library, v. 9.1, including a set of about 50 generic ontologies used in the integration of medical terminologies, and the medical intermediate ontologies resulting from the integration;

• the IMO (Integrated Medical Ontology), a library which represents the unification of the five medical top-levels mentioned above. IMO allows an extended interoperability among the integrated top-levels;

• the formal translation of a set of medical repositories, including the 332,000-concept UMLS Metathesaurus™, which already allows a limited alignment of several large terminologies under a small top-level. The formal translation of UMLS, coupled with the IMO, allowed the classification of such a very large corpus, and the inheritance of axioms defined within the IMO. The hard work now concerns the distribution of the corpus in a large set of sub-domain ontologies and the definition of specialized axioms, which should be conducted in collaboration with standard development committees.

4. Formal tools

To represent our ontologies, we assume a first-order logic with identity, which also supports the definition of meta-level categories, which are not given in real second-order logic, since the set of our predicates is finite (at a given state of the library).

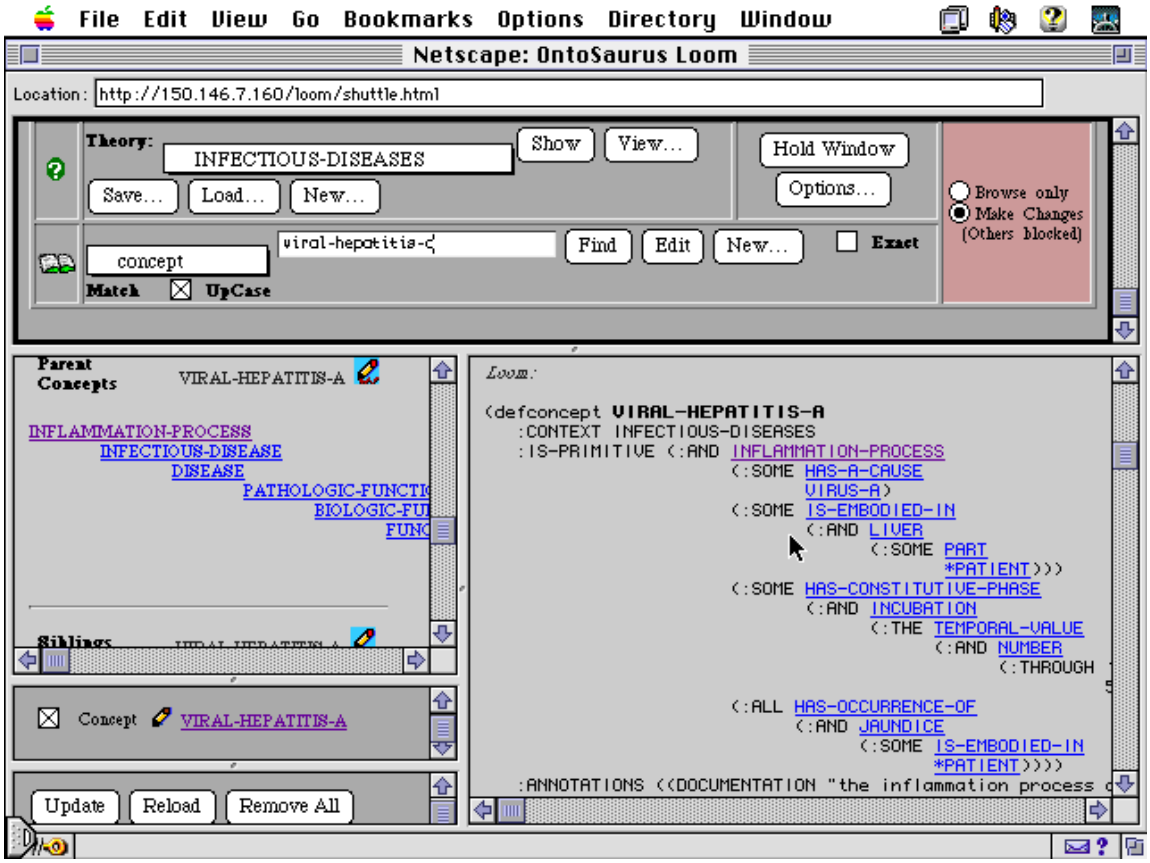


Figure 1: The Loom taxonomy and definition for "viral-hepatitis-A" by means of Ontosaurus.

We use Ontolingua [8], which is a very expressive language derived from KIF [29], and the Loom knowledge representation system [25], which provides TBox classifier services for a subset of our axioms, and a confortable ABox service to express (without classifier) the remainder (for TBox vs. ABox distinction in description logics see [30]. Implementations of both languages allow HTML translation and browsing facilities. In particular, Ontosaurus [40], an interface to Loom through the CL-HTTP server [26], is particularly appropriate for allowing collaborative development of ontologies. We found it to be a crucial point in domain ontology design. Examples of Loom and Ontolingua definitions accessed via Ontosaurus are shown in Figures 1. and 2. Other examples of Loom in §6.

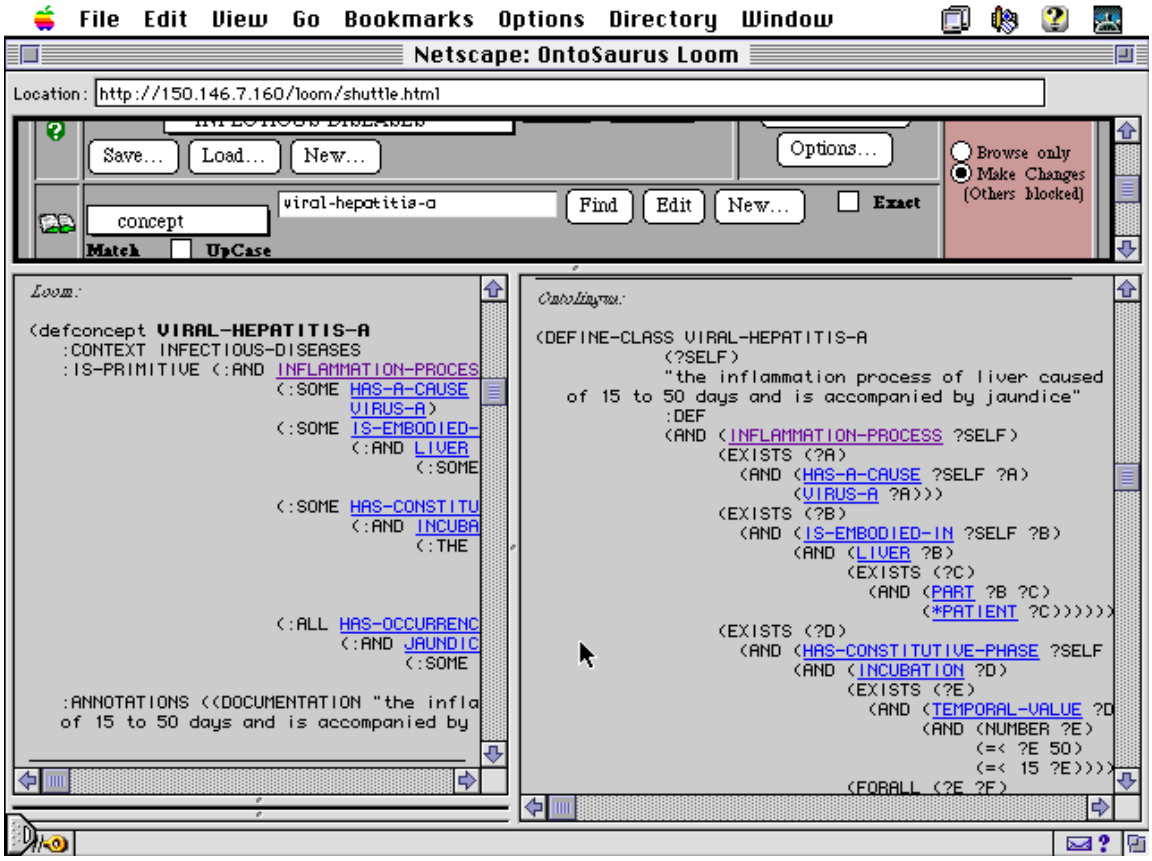


Figure 2: The Ontosaurus entry for "viral-hepatitis-A" in both Loom and Ontolingua.

5. Conceptual tools

To populate the generic section of our library, we looked at relevant ontological theories. In the following we provide some summaries for such ontologies; they are available on-line at our WWW site: http://saussure.irmkant.rm.cnr.it.

1) "formal ontology" theories: theory of parts (*mereology*), of wholes (*topology*), of *identity*, and of *dependence*. This is the philosophical sense of "formal ontology" (according to Husserl), i.e. the study of the fundamental categories of reality, shared by any conceptualization. Thus, it has a meaning different from that more or less accepted in AI, where it means: "formalized, or semantically explicit, theory".

In particular, the theories of parts and of wholes resulted essential in the axiomatization of other relevant generic theories: *localization*, *morphology*, *actants* (participation roles in processes or activities), *time* ontologies, etc.

A 'formal' ontology concerns the possibility of existence (formal existence), not the actuality of existence. In the sense of physics, nothing in common sense exists; while in Quine's sense everything exists which we talk about: existence is existence in a universe of discourse (the domain of quantification). But this also concerns the fundamental relations underlying entities (formal connections among entities, independently from 'material' or 'domain' properties: physical, biological, chemical, etc.).

Theory: *dependence* is an introductory set of dependence relations, as defined in philosophical work of Simons [35], Varzi [43], etc. A (formal-) ontologically relevant dependence may be: causal, physiological, psychologic, functional (physical), and proper (substance/accident). The proper dependence is such when something cannot exist without something else.

Theory: *mereology* presents a version of classical extensional mereology which should be compliant with both Leonard-Goodman calculus of individuals [23] and Tarski's axioms [41].

The focus of theory: *topology* is the 'whole', while mereology focuses on parts. The basic relation is 'connected'. Most axioms are a reinterpretation of axiomatizations given by Varzi [43] and by Asher and Vieu [2].

In theory: *equality* we state our position about *identity*. Due to our conventionalistic approach to identity criteria, no particular grounding theory is provided. We only found that some relations are needed for the task of ontological engineering. In particular, we distinguish between (a) equality and difference applied to a numerical universe, (b) equality and difference applied to function values ranging on non-numerical strings, (c) equality and difference as partial identity with explicit neutralization of some property (space, time, morphology, etc.), and (d) mereological sameness defined as reciprocal parthood.

2) "stratificational" theories: theory of *layers*, and of *granularity*. Stratificational notions help organizing entities of a domain according to the 'life form' they are about (see the Wittgenstein's notion of meaning as basically dependent on the form of life that is producing it [45]). For example, the same object (say, a spleen) has very different identity criteria when it is considered at a molecular biological level, or from the macroscopic, human-fittest viewpoint.

Theory: *layers* implements the so-called 'strata' [18]: Material, Biological, Psychological, Social; this theory also specializes strata according to some scientific granularities [3]. The basic intuition is that reality (in the most neutral way we could figure it out) is 'layered', and such layers have a complex inter-dependence.

Theory: *granularity* implements Sowa's adaptation [36] of Searle's ontology of intentionality [34], which makes a fundamental distinction between 'epistemic' and 'ontological' facts, also recognizing an 'intentional' level which pertains (mainly) to the 'semiotic' aspects of the description of the world in the human (or another organism's) form of life.

3) "individuation" theories: theory of *localization*, of *'absolute' time*, of *morphology*, of *contexts*. Theory: *localization* axiomatizes regions and some special relations: Exactly-Located, Generically-Located, Partly-Located, Wholly-Located [5].

Theory: *morphology* contains some basic and anatomical morphology notions: substance composition, form, morphological properties, etc.

Our theory of contexts, defined in theory: *peak-level*, claims that "ontological" contexts have to be distinguished from "formal" contexts to which a given language or system commit, such as Ontolingua 'theories', Loom 'contexts', CYC 'microtheories', etc. There is no necessary correspondence between an ontological context and a formal context: the former can be modelled in one or more formal contexts, while the latter can represent one, many, or no ontological contexts at all.

In fact, ontological contexts help implementing the notion of localization, as well as situations, states of affairs, beliefs, etc. For example, situations are special ontological contexts for constructing the notion of localization of heterogeneous entities with interrelationships among them (e.g. a state of affairs including an object, a process embodied in it, one or more regions at which they are located, a domain in which this state of affairs may occur, a time span of occurrence, etc.).

4) "applied formal ontology" theories: *meronymy*, theory of *positions*, time theories. There are three different ontologies of time in this library:

a) the*temporal-mereology* [1] see also an adaptation in [36]) uses mereological concepts in its definitions; its relations apply directly to intervals, thus it is difficult to state the common sense notion by which we talk of processes and situations as intrinsically intervallistic. Anyway, the basic ideas in all these theories are founded on some version of the Allen's ontology;

b) our theory: *unrestricted-time* aims at representing the common sense notion mentioned above, plus Kamp's parallel time lines (platforms) [20];

c) the *simple-time* theory follows the temporal mereology approach, and defines useful notions for dealing with "absolute" time expressions.

Theory: *position* is a domain application of localization theory: related positions and coordinated positions. It is inspired by the common sense use of prepositions and some cognitive semantics models [4][24].

Theory: *meronymy* is provided to account for special notions of whole and part, widely used in domain ontologies: societies, collections, etc. The basic move is relaxing mereological definitions (through concept like 'weak part' and 'weak whole') and then creating specialized notions. Some specializations come from the work of Gerstl [14] and others.

5) "participation" theories. Theory: *actants* describes the dimension of actantial concepts, which deals with roles and participation played in situations, scripts, scenes, narratives, etc. This theory contains the minimal relations to account for domain models encountered to now and is inspired by narratology [32] and linguistics [9].

Theory: *process&participants* is a summary of various linguistic ontologies concerning event structure. The main distinction is between a temporal analysis of processes into subprocesses (see also theory: *peak-level*), and a spatial analysis into participants. Most of this theory comes from Sowa [36], which stresses the relevance of Aristotle's 'aitiai' and Whitehead's 'nexus' for modelling processes, actors, scenes, etc.

6) Theory: *assessment* includes various relations pertaining to the 'epistemic' aspects of ontology: notions of interpretation, representation, belief, relevance, conventionality, typicality, etc. These are very challenging notions to axiomatize from a strict ontological viewpoint, also because the work done is very limited. Current definitions are still in progress.

7) portions of some mathematical theories (*abstract algebra, set theory, geometry*)

8) standard quantities (dimensions, units), in theories: *units, standard-units* and *standard-dimensions*.

9) some physical concepts in theory: *physical-concepts*.

10) "meta-level" theories. Theory: *structuring-concepts* introduces the general categories for the dimensions which structure (allow the axiomatization of) the concepts within a conceptualization; they are defined as (meta) classes whose instances are either unary, binary, or ternary relations. An extension of this will be a theory of semantic fields (e.g. [21]).

As far as 'representation ontology' is concerned, we have defined a theory: *metaontology*, which axiomatizes some meta-level categories on the basis of the work of Guarino [15] and some cognitive literature. It is aimed at giving an explicit semantics to usually intuitive or merely formal notions such as 'category', 'type', 'property', 'relation', 'role', 'reified property'.

Other literature used as a conceptual tool includes cognitive semantics "schemas", linguistics notions, some mathematical and engineering theories.

6. Case Studies

In order to illustrate the methods of ontology integration employed in ONIONS, we provide herewith two case studies. The first one concerns a concept from the UMLS top level semantic network, the second one regards a concept from the GALEN Core Model.

(1) "body-location-or-region" in the UMLS top-level

Our aim is to get an ontologically motivated definition. The original definition from UMLS top-level is firstly translated to Loom: default semantics is applied to bypass inconsistency with inherited axioms, recursive axioms, etc., found in the source ontology:

```
(defconcept body-location-or-region
   :ANNOTATIONS ((DOCUMENTATION "An area, subdivision, or region of the body
    demarcated for the purpose of topographical description."))
   :is-primitive spatial-concept
   :default (:and (:all has-conceptual-part body-location-or-region)
                  (:all traversed-by body-location-or-region)
                  (:all traverses body-location-or-region)
                  (:all has-location
                   (:or body-location-or-region
                    body-part-or-organ-or-organ-component))
                  (:all adjacent-to
                   (:or body-location-or-region
                    body-part-or-organ-or-organ-component
                    body-space-or-junction))
                  (:all connected-to body-location-or-region)
                  (:all location-of
                   (:or acquired-abnormality
                    tissue biologic-function body-location-or-region
                    body-part-or-organ-or-organ-component
                    injury-or-poisoning congenital-abnormality))
                  (:all conceptual-part-of
                   (:or fully-formed-anatomical-structure
                    body-system body-location-or-region)))
  :context umls-sn)
```

The formula states that a `body-location-or-region` IS_A `spatial-concept` which, by default, may have some relations with other concepts, for example, that it maybe `traversed-by` another `body-location-or-region`.

Secondly, a consistent and correctly quantified definition is built: we used open-world semantics, the distinction between definitional (i.e. `:is-primitive`) and implicational (i.e. `:implies`) axioms, and the distinction between "essential" axioms (the `:some` clauses) and merely "contingent" axioms (the `:all` clauses):

```
(defconcept Body-Location-Or-Region
  :is-primitive (:and Spatial-Concept
                 (:some Conceptual-Part-Of
                  (:or Body-System Fully-Formed-Anatomical-Structure)))
  :implies (:and (:all Result-Of Mental-Process)
                 (:some Conceptual-Part-Of
                  (:or Body-System Fully-Formed-Anatomical-Structure))
                 (:all Adjacent-To
                  (:or Body-Location-Or-Region Body-Space-Or-Junction
                       Body-Part-Or-Organ-Or-Organ-Component))
                 (:some Location-Of
                  (:or Body-Location-Or-Region
                       Acquired-Abnormality Congenital-Abnormality
                       Injury-Or-Poisoning Biologic-Function Tissue
                       Body-Part-Or-Organ-Or-Organ-Component))
                 (:all Traverses Body-Location-Or-Region)
                 (:some Connected-To Body-Location-Or-Region))
  :context consistent-umls-sn
  :annotations ((DOCUMENTATION "An area, subdivision, or region of the body
    demarcated for the purpose of topographical description.")))
```

Finally, an ontological definition with the correct identity criteria from generic theories is developed (another intermediate step, bypassed here, is the re-use and axiomatization of the information available from the natural language definition).

To do this, we have to solve a main ontological issue: what is the primary identity criterion of body regions? Are they body-parts (first class objects, which have location and time as primitive dimensions) or regions (objects whose identity criterion is their essential dependence on another object whatsoever: location of something)? Since they can be touched, cut, filled, etc., the intuition goes to the first class interpretation, but one could think of a special metonymy of medical language: when a body region is at hand, a body part located at that region is at hand, and which one is evident from the operations carried out by physicians and (usually) shared by them, or simply from the functions involved in the parts located at the region.

On the other hand, if we adopt the regional interpretation, we could have hard times in axiomatizing it: a body region can only exist within an organism ('rigidly-depends' on it), but cannot be part of it (in fact, UMLS has it as 'conceptual-part'), otherwise it would be a 'body-part'.

Currently, we adopt the regional interpretation and axiomatize it by (1) restricting the kind of objects which can be located at body regions and (2) restricting the part relations applied to 'body-part' (component) and 'body-region' (portion) (axiomatized in theory: *meronymy*). The result is:

```
(defconcept Body-Region
  :is-primitive (:and Region
                 (:some whole-location-of
                  (:or Body-Part Tissue))
                 (:some portion organism))
  :implies (:and (:some whole-location-of
                  (:or Body-Part Tissue Body-Region))
                 (: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)
```

(2) "myopathy" in GALEN

The original definition of "myopathy" in GALEN (here translated to Loom) features correct TBox semantics, but lacks ontological clarity or any gloss to interpret it:

```
(defconcept myopathy
   :context galen
   :is (:and clinical-situation
             (:some shows
                    (:and presence
                          (:some is-existence-of
                                 (:and muscle
                                       (:some has-pathological-status
                                              pathological)))))))
```

If taken literally, and having no further hints from the overall structure of the model, this says that: a myopathy is a clinical situation which shows `'the presence which is existence of'` muscle which have a pathological 'pathological status'. Apart obscurity and linguistic bizarreness, since neither 'presence', 'existence', nor 'pathological-status' have an axiomatization in the model, one is at odds in justifying their inclusion to state the simple paraphrase of myopathy as "any disease of a muscle", as can be found in a medical dictionary like the Dorland's: [7].

For example, in ON9.1 we could define myopathy straightforwardly as:

```
(defconcept myopathy
   :context pathologic-functions
   :is (:and pathologic-function
             (:some embodied-in muscle)))
```

by using the process taxonomy (process function biologic-function physiologic-function pathologic-function) and the ontology of participants, by which a process has to be 'embodied in' some object. Both process taxonomy and participants are axiomatized in dedicated theories (§5.).

Actually, the above GALEN definition implicitly states another assumption: that a myopathy is not simply a disease, but a 'clinical-situation' characterized by that disease: the use of presence, existence, showing, etc. might have been motivated by that assumption. If accepted in an ontological framework, this is a quite radical move: all disease concepts would become contexts rather than processes, and their identity criterion would be essentially changed. Such a choice is ambivalent even in the GALEN Core Model, where a 'clinical-situation' is a 'psychosocial-construct', while the "pathological" value of 'pathological-status' makes a concept classify under 'pathological-condition' which is a primitive concept just under the top concept.

Incidentally, such an understatement of ontological choices is typical of many terminologies and ontologies, and even of some top-levels, as shown in [17].

However, within the ONIONS methodology framework, no choice should remain intrinsically ambivalent: it must be explicited and - in case of conflict - segregated in a specialized context. A treatment of disease as a situation is possible, although such conceptualization should be separated from that of disease as a process (as well as from another alternative: disease as a diagnosis); for example:

```
(defconcept myopathy
   :context clinical-situations
   :is (:and clinical-situation
             (:some context-of
                     (:and pathologic-function
                                 (:some embodied-in muscle)))))
```

which makes use of the ontology of contexts (§5.) (on their turn, 'clinical-situation', 'patient', and various healthcare structures are axiomatized elsewhere).

Case Study 2 shows the importance of formal ontology and methodology to avoid obscurity and linguistic awkwardness. On the other hand, if the task at hand is having GALEN Core Model completely integrated with other ontologies (say: 'unified'), even redundant relations and concepts must find a place in the unified ontology, or at least special 'mapping rules' are to introduced to get complete interoperability. But the integration of the intended meanings ('partial integration'?) should be sufficient to solve most integration-based problems or at least be preliminary to solve them.

7. Conclusions

Our experience has proved that the ontologies produced by means of the ONIONS methodology support:

- formal upgrading of terminology systems: term classification and definitions are now available in a common, expressive formal language;
- conceptual explicitness of terminology systems: (local) term definitions are now available, even though the source does not include them explicitly;
- conceptual upgrading of terminology systems: term classification and definitions are translated so that they can be included in an ontology library which has a subset constituted of motivated generic ontologies;
- ontological comparability, since pre-existing ontology libraries pertaining to different fields are largely employed.

There are intrinsic factors hampering automatic ontology integration, mainly due to the necessity of off-line human intervention in the search, choice, and formalization of generic ontologies. For example, the formalization of system theory (the usual configuration of component-state-event-process), widely available in the engineering domain does not fit the medical domain. If we want to understand the basic principles motivating the conceptualization of terminology in domains such as medicine, it is therefore necessary to adopt an approach like the one presented here, i.e. to refer to the theories provided by linguistics, philosophy, and cognitive science.

Acknowledgements

This paper has benefited from the useful comments and suggestions of Nicola Guarino, Bruno Felluga, Fabrizio Giacomelli and the anonymous referees. Our research is partly supported by the Italian National Research Council Special Project ONTOINT (ONTOlogical tools for Information iNTegration) and by the EU funded Concerted Action PROGUIDE.

References

[1] Allen J, Hayes P. "A Common-Sense Theory of Time" in: *Proceedings of IJCAI85,* 1985.

[2] Asher N, Vieu L, "Towards a Geometry of Common Sense" in *Proceedings of IJCAI95*, 1995.

[3] Blois M, *Information and Medicine*, Berkeley, University of California Press, 1980.

[4] Cardona GR, *I sei lati del mondo*, Bari, Laterza, 1985.

[5] Casati R, Varzi A, "The Structure of Spatial Localization", *Philosophical Studies*, 82, 1996.

[6] Coté RA, Rothwell DJ, Brochu L, eds. *SNOMED International* (3rd ed.), Northfield, Ill, College of American Pathologists, 1994.

[7] Dorland's Illustrated Medical Dictionary, 27th edition, 1991.

[8] Farquhar A, Fikes R, Rice J, "The Ontolingua Server: a Tool for Collaborative Ontology Construction", *Proceedings of Knowledge Acquisition Workshop,* Banff, participants edition, 1996.

[9] Fillmore CJ, "Frames and the Semantics of Understanding" *Quaderni di Semantica*, 6, 2, 1985.

[10] Gabrieli E, "A New Electronic Medical Nomenclature", *J. Medical Systems*, 3, 1989.

[11] GALEN Project, documentation available at the URL: http://www.cs.man.ac.uk/mig/galen

[12] Gangemi A , Steve G, Giacomelli F, "ONIONS: An Ontological Methodology for Taxonomic Knowledge Integration" In P. van der Vet (ed.) *Proceedings of the Workshop on Ontological Engineering, ECAI96*, 1996.

[13] Gangemi A, Pisanelli DM, Steve G, "Ontologizing the UMLS Metathesaurus", ITBM-CNR Technical Report 0198A, 1998.

[14] Gerstl P, Pribbenow S, "Midwinters, Endgames, and Body Parts: A Classification of Part-Whole Relations". *International Journal of Human-Computer Studies*, 43, 1996.

[15] 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 Kaufmannn, 1994.

[16] Guarino N, SOLMC Project, documentation available at the CNR, 1996.

[17] Guarino N, Masolo C, Carrara M, "Top-Level Ontological Categories" unpublished draft, 1997.

[18] Hartmann N, *Zur Grundlegung der Ontologie*, Berlin, de Gruyter, 1966.

[19] Humphreys BL, Lindberg DA, "The Unified Medical Language System Project" in Lun KC et al. (eds): *Proceedings of MedInfo92,* Amsterdam: Elsevier Science Publishers, 1992.

[20] Kamp H, "Events, Instants and Temporal Reference", in: *Meaning, Use and Interpretation of Language*, Berlin, de Gruyter, 1979.

[21] Kittay EF, *Metaphor*, Oxford University Press, 1991.

[22] Lenat DB, Guha RV, *Building Large Knowledge-based Systems: Representation and Inference in the CYC Project*, Menlo Park, Addison-Wesley, 1990.

[23] Leonard HS, Goodman N, "The Calculus of Individuals and its Uses", *Journal of Symbolic Logic*, 5, 1940.

[24] Levinson S, "Primer for the Field Investigation of Spatial Description and Conception", *Pragmatics*, 2/1, 1992, 5-47.

[25] MacGregor RM, "A Description Classifier for the Predicate Calculus" *Proc.edings of AAAI 94, Conference* 1994.

[26] Mallery JC, "A Common LISP Hypermedia Server", *Proc. WWW 94*, 1994.

[27] McCarthy J, Buvac S, "Formalizing Context", Stanford Un. Tech. Note STAN-CS-TN-94-13, 1994.

[28] National Library of Medicine, *UMLS Knowledge Sources*, 1997 edition, available from the NLM, Bethesda, Maryland.

[29] Neches R et al., "Enabling Technology for Knowledge Sharing", *AI Magazine*; 3, 1991.

[30] Patel-Schneider PF, Swartout B, "Draft of the Description Logic Specification from the KRSS group of the DARPA Knowledge Sharing Effort", 1993.

[31] Pisanelli DM, Gangemi A, Steve G, "WWW-available Conceptual Integration of Medical Terminologies: the ONIONS Experience", in: *Proceedings of AMIA97,* Philadelphia, Hanley&Belfus, 1997.

[32] Prince G. *Narratology,* Berlin, de Gruyter, 1982.

[33] Rector A, Gangemi A, Galeazzi E, Glowinski A, Rossi-Mori A, "The GALEN CORE Model Schemata for Anatomy: Towards a Re-Usable Application-Independent Model of Medical Concepts",*Proceedings of Medical Informatics EuropeMIE94,* 1994.

[34] Searle JR, *The Construction of Social Reality,* New York, Free Press, 1995.

[35] Simons P, "Parts: a Study in Ontology", Clarendon Press, Oxford (1987).

[36] Sowa J, "Knowledge Representation: Logical, Philosophical and Computational Foundations", Boston, PWS, in press.

[37] Steve G, Gangemi A, "ONIONS Methodology and Ontological Commitment of Medical Ontology ON8.5", *Proceedings of Knowledge Acquisition Workshop,* Banff, participants edition, 1996.

[38] 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.

[39] Steve G, Gangemi A, Rossi-Mori A, "Knowledge Integration of Medical Terminological Sources", *Proceedings of FLAIRS,* 1996.

[40] Swartout B, Patil R, Knight K, Russ T, "Toward Distributed Use of Large-Scale Ontologies", *Proceedings of Knowledge Acquisition Workshop,* Banff, participants edition, 1996.

[41] Tarski A, *Logic, Semantics, Metamathematics,* Oxford, Clarendon, 1956.

[42] Van Heijst G, Schreiber ATh, Wielinga BG, "Using Explicit Ontologies in KBS Development", *Int. Journal of Human-Computer Studies,* 1997.

[43] Varzi A, "Le strutture dell'ordinario", in: *Logos, teorie dell'essere, teorie della norma,* Milano, Giuffre', 1996.

[44] WHO, *International Classification of Diseases* (10th revision), Geneva, WHO, 1994.

[45] Wittgenstein L, *On Certainty,* Oxford, Blackwell, 1969.