

# ONIONS: An Ontological Methodology for Taxonomic Knowledge Integration

Aldo Gangemi, Geri Steve, Fabrizio Giacomelli

Reparto Informatica Medica, Istituto Tecnologie Biomediche, CNR, Roma, Italy  
email: steve@relay.itbm.rm.cnr.it; aldo@color.irmkant.rm.cnr.it; fabrizio@psc2.irmkant.rm.cnr.it

## Abstract

We describe *ONIONS*, a methodology for integrating ontologically-heterogeneous taxonomic knowledge and its current application to medical domain. Some clarification is given of our intended meaning of ontology and related notions, then main problems of ontology design are addressed, with a short comparison with alternative approaches. The methodology is described as a sequence of phases. The top-level of the current integrated ontology of heterogeneous medical taxonomies is presented in an order-sorted logic.

*ONIONS* includes no claim of global objectivity (it performs an integration of explicit—or explicited—ontologies of given taxonomic sources), but provides a feasible solution to the problems of modelling stopover and cognitive basicity. *ONIONS* has been defined in order to be applied to sources within the same domain, nevertheless it has been applied to a very wide and inherently heterogeneous domain like medicine, so complex that it can be considered in itself an integration of subdomains.

## 1. Introduction

The craft of ontology design could take many routes, depending on what task is supposed to be accomplished by means of an “ontology”.

Philosophers may prefer top-down creation of domain-independent top-models, domain modellers may have a penchant towards bottom-up induction of local rules. Our approach drastically rejects both: we do not want to state the rightest, general purpose categories of “reality” or “human culture”, nor do we want to abstract ad-hoc decisions on the basis of local domains.

We will outline our *hybrid* approach based on the following assumptions:

- any domain knowledge (DK) shares **high-level theories**, usually implicit, which are motivated by both external world and cognitive attitudes;
- DK shares **operational**, complex **knowledge** which is hardly reducible to linguistic descriptions; on the contrary, it includes various sensori-motor routines and abstract reasoning developed in years of practice;
- DK usually shares a **language**, retailed on natural language features, which is a special means for activating, time by time, the relevant knowledge to communicate;
- often, DK is further organized through linguistic **repositories**, which conventionalize some relevant knowledge for various purposes.

On these assumptions, we defined a methodology for the *integration of medical taxonomic knowledge by means of source comparison and abstraction through general and domain theories*.

In this document, we will briefly situate our methodology in the context of some current disciplines: formal ontology, cognitive semantics, knowledge bases integration, etc. Then, we will detail the methodology itself, as applied to the medical domain (the current output of which is presented in the appendix in an order sorted logic). Finally, we will give some information about implementation issues for both the methodology and the output.

## 2. The *Ontology* semantic field

### 2.1 Representation and content

A basic distinction has been made by Gruber<sup>11</sup> between Representation ontology and Content ontology: «Ontologies like the frame ontology ... in KIF are called representation ontologies. [They] provide a framework, but do not offer a guidance about how to represent the world. Content ontologies make claims about how the world (or a conceptualization of it) should be described». Actually, the choice of proper mappings between content structure and representation structure is extremely important for avoiding the confusion provoked by formal issues when they are not distinguished from ontologic issues. Important issues (such as the proper intended meaning of a role, a concept, a relation, etc.) have been treated in various works by Guarino<sup>30, 32, 33</sup>. The ontologies resident in the Ontolingua package<sup>11</sup> are another step in the direction of clear distinctions.

As far as we are concerned, in this paper we present a sample mapping among the general objects of the knowledge included in our top-level, and a minimal set of formal categories (sort, relation, etc.).

The *ontological integration* envisaged here is at a deeper level than *representational integration* (Fig.1). In fact, representational integration concerns heterogeneity of *formal languages* (e.g.<sup>17</sup>), or heterogeneity of *data base schemata* (e.g.<sup>25,44</sup>). Ontological integration concerns the heterogeneity among *conceptualizations*. In knowledge representation, ontological integration is often bypassed, while task-oriented systems usually bypass representational integration as well.

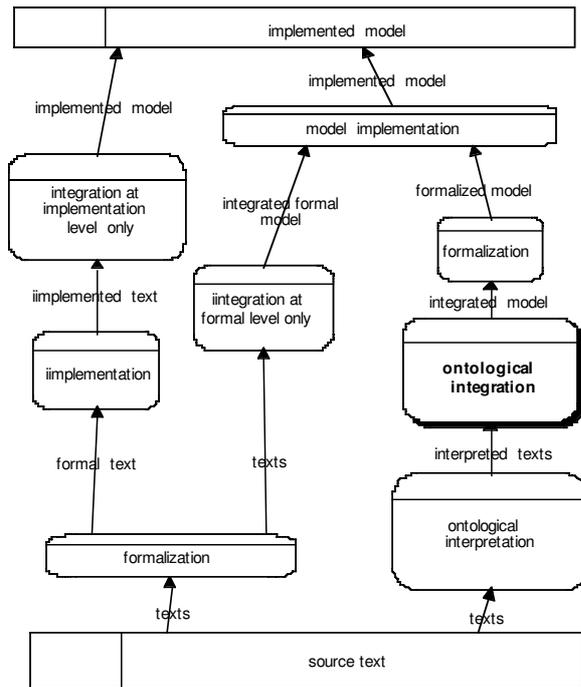


Figure 1: Alternative integrations are featured from texts to model implementation. Representational integration usually bypass content ontology. Implementational integration usually bypass representation ontology as well.

## 2.2 What's a content ontology

Content ontologies as well do not share a common intended meaning for the term itself. Some philosophers mean: "the conditions of the possibility of the object in general and the individuation of the requirements that every object constitution has to satisfy" (Husserl), others mean "the categories of knowledge" (but the formers claim that this is *epistemology*). In logic, formal ontology may be directly defined as: «the systematic, formal, axiomatic development of all forms and modes of being»<sup>49</sup>. AI is more pragmatic: the abovementioned definition is intended as "making claims on how to describe something", and another<sup>46</sup> points out: the «schematic descriptions of the contents of domain knowledge».

A temptative systematization has been given by Guarino<sup>34</sup>:

- a) Ontology (1): a branch of philosophy which deals with the nature and the organization of reality.
- b) Conceptualization: an intensional semantic structure which encodes the implicit rules constraining the structure of a piece of reality.
- c) Ontology (2): a logical theory which gives an explicit, partial account of a conceptualization.
- d) Ontological theory: a set of formulas intended to be always true according to a certain conceptualization.
- e) Ontological commitment: a partial semantic account of the intended conceptualization of a logical theory (this is the 'representation ontology').

We substantially agree with Guarino's definitions, except that: *a conceptualization encodes the implicit rules constraining a piece of general or domain knowledge, which in its turn is motivated by both the structure of reality and the cognitive attitudes of humans*. Following this meaning of *conceptualization*, we use *formal ontology* in the sense of Guarino's Ontology (2), while by *non-formal ontology* we mean the pragmatic sense of AI: an explicit, schematic account of a conceptualization, not yet formalized in the logical sense.

The reasons for specializing this way the terminological field of ontology are both theoretical and methodological.

On one hand, we find extremely useful to assume *cognitive semantics* (rather than *philosophical ontology*) as our background philosophy (see section 2.4), on the other hand, our methodological choices restrict the currently feasible development of formal ontologies to the integration of *explicit, task-oriented expert knowledge* (see below sections about ONIONS methodology).

## 2.3 General and domain ontologies

Obviously, a distinction is assumed between a general non-formal ontology (simply: "ontology" in the following, as opposed to "formal ontology") and a domain one. An operational definition is provided:

- A general ontology can be extracted from general purpose encyclopaedias and dictionaries, common sense physics, philosophical categorizations (such as Plato's, Aristotle's, Lull's, Kant's, Peirce's (and Sowa's<sup>21</sup>), Hartmann's<sup>35</sup>, etc., see<sup>29</sup>), top-levels of various computational large KBs (Penman<sup>27</sup>, CYC<sup>37</sup>, Roget's thesaurus, etc.), and even through introspection.
- A domain ontology can be extracted from special purpose encyclopaedias, dictionaries, nomenclatures, taxonomies, handbooks, scientific special languages (say, chemical formulas), specialized KBs, and from experts.

#### 2.4 Some excerpts from semiology

One of the main farsighted views of the linguist Saussure<sup>42</sup> was his distinction between *praesentia* and *absentia*. Given a meaningful linguistic chunk—a word, a phrase, a morpheme, a sentence—which he called *signe*, such a *signe* could be *in praesentia* if it is considered as part of an actual text we are dealing with, while it could be *in absentia* if it is part of possible texts that can be associated to another *signe*.

A similar account to Saussure's, though independently developed, was contributed by philosopher Peirce, whose notion of *interpretant* constitutes the medium between his *sign* and *external world*<sup>39</sup>. An interpretant is any sign or group of signs which is used to *interpret* a given sign and/or a state of affairs or phenomenon in the world.

The *interpretant* is a quite similar notion to Saussure's *signe in absentia*; only, it is considered dynamically, as dependent on the human active involvement in real world processes and human attempts to interact with environment by interpreting it through the inferential process of *abduction*. In other words, a context faces a perceiving individual with an interpretive need, which is accomplished through his/her navigation within the ideal knowledge (either background knowledge or operational knowledge): a non-finite net of interpretants, which, among others, feature linguistic signs.

Certain regions within ideal knowledge seem more appropriate to the context, where appropriateness is judged through abductive inference, which provides judgement for further intentional behaviour inside the context.

The interplay of interpretants in the ideal knowledge creates the situation of a *global*, non-finite conceptualization which motivates the mutual comprehension of humans. Only a *local* context can motivate a given ontology which schematizes a conceptualization.

#### 2.5. Knowledge sources as conventional contexts

The only alternative to a dynamic context is a local structured source of knowledge. That is the reason why we decided to integrate ontologies from taxonomic knowledge sources in a domain (medicine).

The above theory, used for taxonomic sources, implies the following: the reason for the dissimilarity of intended meanings is that ideal knowledge of, say, *viral hepatitis* is not in any non-trivial sense correspondent to the phrase 'viral hepatitis' nor to a viral hepatitis 'out there'. Phrases in different taxonomic sources or communicative situations are links to ideal knowledge; a source organization or a context, depending on the particular task they are for, make them 'emerge' some relevant issues in the context of the ideal knowledge.

For example, a minimal ontological definition of *viral hepatitis* is:

“inflammation of liver caused by virus”

But *inflammation* may mean (in different—or within the same— taxonomic source:

- a *physiological function* performing segregation of external agents;
- a *portion of a body part* which embodies that physiological function;
- a specific *abnormal morphology* (texture, color, shape, various abnormalities) of a portion of a body part.

All these *inflammations* are equally acceptable. Only, one cannot figure out, from the organization of a single source, all the valid usages of that phrase in the world out there. Thus taxonomic sources and contexts select some issues of the ideal knowledge, like pointing the finger to a site in the mind of an expert.

#### 2.6. Global and local again

The notions of ideal knowledge, context, and interpretation are connected to the ontologic task. In fact, a set of heterogeneous texts containing more or less structured linguistic data (a reference universe for an ontological task) have to be considered semantically if we relate them to a global space, such as the *ideal knowledge*. And moving, *navigating* in those data would equal to navigate in a local region of ideal knowledge, i.e. a *context* [cf. <sup>40, 50</sup>]. But how to interpret a context inside the ideal knowledge, without modelling it in its totality? Here comes the need for a grounding of signs in their cognitive and ontological dependences.

Another remark has to be made on what a context is meant to be. In fact, a given spatial environment including a perceiving individual is a context in an *objective* sense, while a local region of the ideal knowledge net, which can be the objective context as perceived by individuals, is a context in a *subjective* sense. This is quite relevant for the distinction between the cognitive and the external conditions of ideal knowledge structure.

The notion of *ideal knowledge* and its global, distributed, and super-individual nature, does not prevent us from investigating the motivation which makes up the ideal knowledge as it develops. According to current research, the motivation seems to reside in two powerful influences: *external world structure*, and *cognitive attitudes* shared by human individuals.

A source of constraint for ideal knowledge is the actual structure of the world<sup>40</sup>, at least the structure that directly interacts with humans: the ordinary, or 'common sense' structure<sup>23, 53</sup>: *universals*, or *invariants*, of cognitive perception, such as *wholeness* and *parthood* of objects [cf. <sup>43</sup>], *connexity*, *gestaltic* properties, *strata* of reality (material, biologic, psychological, socio-cultural) [cf. <sup>35</sup>].

Cognitive semantics<sup>36,51,52</sup> maintains that *cognitive schematization* intervenes to impose form on perception data, as well as upon syntactic and grammatical structure, and even upon interpretation paradigms. Especially *kinaesthetic image schemata*, such as *up/down, front/back, containment, configuration, path, link, force dynamics*, pave the way for perceiving external world, for ordering words and understanding syntactic relations. In other words, schemata should be a means for constraining the complex structure of ideal knowledge, thus performing a restriction on possible interpretation (explicitation of an ontology as understanding).

### 3. Some approaches to ontology design and related problems

We discuss here main problems of ontology design: the assessment of *stopover* and *relevance* in ontology explicitation.

#### 3.1 The stopover problem

How to stop detailing the explicitation of a conceptualization, i.e., how do we stop refining an ontology? This problem has appeared in different forms in the past. The classic distinction between *terminologic* and *assertional* knowledge in KL-ONE languages tried to overimpose a formal criterion on an ontological problem, but what is terminological? How to state the border? Analogous forms of the stopover problem are the linguistics debate between *dictionary-type* and *encyclopaedic-type* definitions, or the logical riddle between *analytic* and *synthetic* categories.

Marconi<sup>38</sup> gives an account of the problem in terms of a plausibility metrics. Other accounts have been given by evoking contextual solutions (for instance, the contextual triggering of CYC<sup>37</sup>, and contextual logics<sup>14, 54</sup>).

The interest of such approaches notwithstanding, one still lacks an ontological criterion. This is quite obvious, because the stopover is dependent on *task*, and tasks can be modelled only for specialized, very limited, and conventional protocols of planning and acting.

Our solution is to integrate knowledge sources which have been developed by experts for given tasks with consequent stopovers. Obviously, not all domains allow such a solution. We tried with medicine and the results seem encouraging.

#### 3.2 The relevance problem

Another ontological problem is more logically-oriented: how much ontology has to be encoded (formalized) as *a priori sorts* rather than as *roles* (domain or range of a relation, usually encoded by means of *lambda-abstraction*)? Although this problem seems exquisitely representational, it constitutes an interface between

ontology and ontological commitment (representation ontology). In our practice, any source seems to activate only certain sorts, but when ontology is made explicit, immediately the problems arises about considering some sorts as absolutely relevant and other sorts as only useful abstractions to be secondarily retrieved from a special query. In terms of ontological commitment, this problem has also taken the form of distinction between *property* and *role* in terms of permanency, rigidity, etc. (see<sup>32,33</sup>).

Our solution also adopted the notion of cognitive basicity<sup>a 41, 36</sup> of a sort for a given sub-domain, but there is the danger of loosing this criterion when integration takes into account more and more sub-domains.

#### 3.3 Capture and coding of ontologies

A well-focused approach to ontology design *from scratch* is the "Procedure for ontology capture", introduced by<sup>45</sup>. They distinguish *capture* from *coding*, which should be the phase of formalization according to a representation ontology.

The capturing phase is structured in *scoping* (brainstorming and rough arrangement of terms); *definition* (identification of basic terms and their necessary and sufficient definition in natural language text, possibly addressing for clarity); *review*; *meta-ontology* (the set of devised terms and their definitions as requirement specification).

We do not commit to this method since its *from-scratch* attitude, which provides no solution to the *stopover* and the *relevance* problems. In fact, though the notion of basicity is well-grounded in linguistic, philosophical, and psychologic theories, it provides no Ockham's razor until basic terms are listed. The task is even more guessing-like in scientific domains, where basicity is weakened by conventional modularity and granularity: different modules within the same domain do not share the same basic terms, as well as many scientific domains have various granularities, which also affect the basicity evaluation.

We will commit to the capture with a quite different procedure, which scope is somehow more restricted, but which can account for stopover, probably relevance, and basicity in a sound, domain-oriented way.

#### 3.4 Automatic agent mediation

A fascinating approach is that of<sup>55</sup>, which figures out an agent which mediates among different ontologies. The problem is the dynamics that an agent should perform to mediate: either an enormous ontological knowledge base should be loaded in it, or very tentative decisions are to

---

<sup>a</sup> Basic terms are those closer to direct experience, not abstracted out, nor overspecialized (*chair* towards *furniture* and *arm-chair*).

be taken in a situation about heterogeneous conceptualizations. In fact ontological merging has been performed to now by hand<sup>56</sup>.

#### 4. The ONIONS methodology

Integration of large knowledge bases is a main issue<sup>5,11,12,15,17</sup>, which embraces taxonomic knowledge integration as well. In order to analyse and integrate domain ontologies of large KBs, we developed the

ONIONS (ontologic integration on naïve sources) methodology (fig. 2)<sup>9,22,60</sup>.

It creates a common framework to interpret the definitions that are used to organize a set of terminological sources. In other words, it allows to work out coherently a domain ontology for each source, which can be then compared with the others and mapped to an *integrated model*.

Our domain is medicine. In fact, our research (besides the GALEN project, which we take part in) is related to other efforts —CEN/TC251/pt003<sup>2</sup>, and CANON<sup>5</sup>.

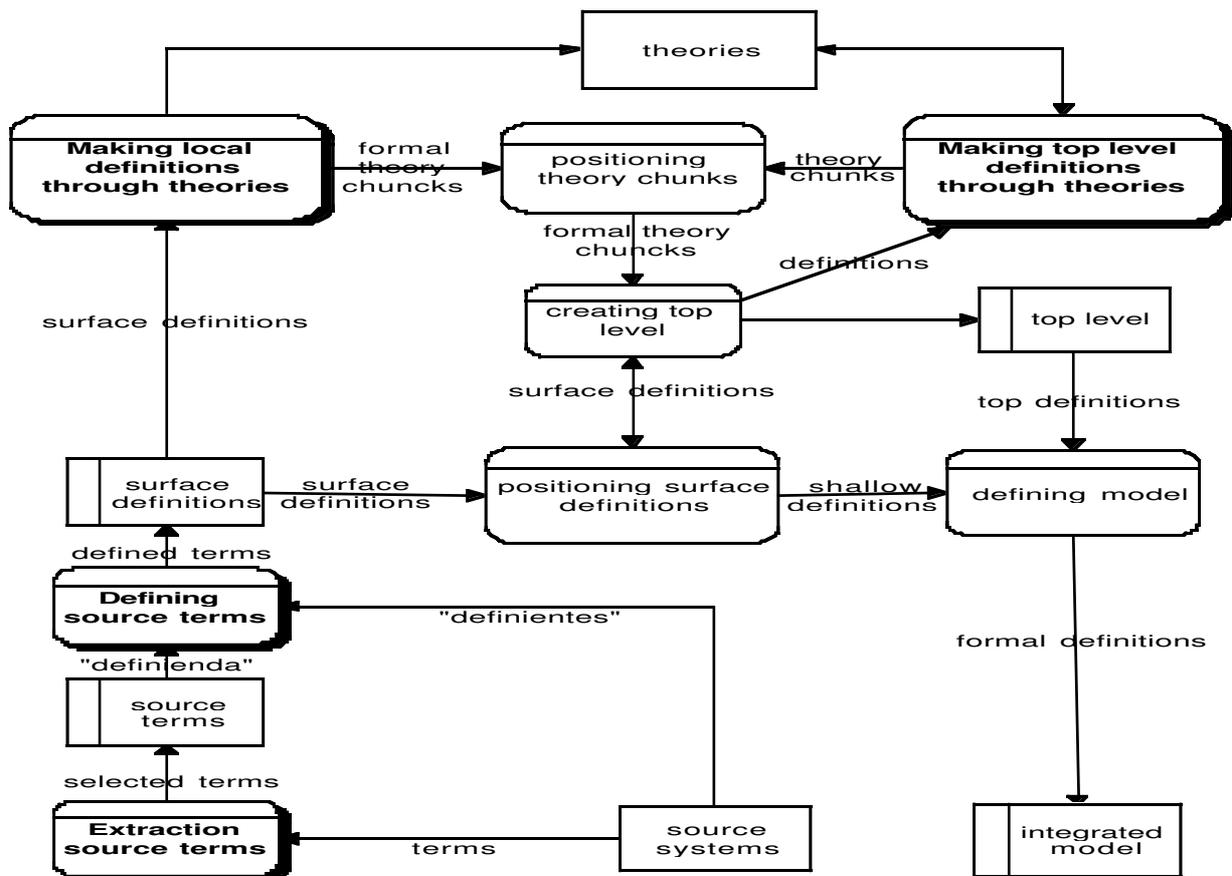


Figure 2: A data flow diagram of the procedure of ontological integration: from heterogeneous terminological sources to an integrated model.

Our work has two main goals: a) generalizing a framework to integrate terminological knowledge from various medical sources by analysing their domain ontologies, and b) defining a new and open domain ontology which merges a general ontology with various domain ones.

The first aim of this ontological integration has been to develop a CORE model of medical concepts which supports a conceptual convergence among different terminological systems or repositories.

Taxonomic sources in medicine —such as classifications, nomenclatures, semantic networks— are

dependent on ontologies, usually implicit, but coherent with specific tasks (epidemiology, indexing, retrieval, acquisition, expert systems)<sup>20,4</sup>.

Current efforts are mainly addressed to extending the ontological knowledge base to map larger parts of the sources and to apply ONIONS to microdomain integration, where a proficuous collaboration has been activated with standardizer physicians in the microdomains of *medical devices* and *vital signs*.

#### 4.1 Phase I: Extraction of source terms

Selecting relevant sets of terms from terminological sources (*source terms*): code definitions or key-words from classifications, nomenclatures, coding systems, thesauri. This phase has hooks to corpora formation techniques and textual types definition and acquisition (not investigated here).

When a given corpus is collected, the order of terms contained in each single source is inferred. The sources used to now for medical ontology are mostly taxonomies (only once a semantic network, sometimes flat lists).

The order is exploited to identify the top-level categories in the source, and then top-level categories are used to choose a depth limit in the hierarchy (for example, we chose to truncate any *body part* hierarchy—in the *vessel* branching (if any)—to kinds of vessel, not including instances of arteries, veins, etc. Since our main scope was to integrate general medical taxonomic knowledge, the detailed taxa for anatomy are excluded. This seems to be sound to the extent that a specialized microdomain integration can be done in a further phase.

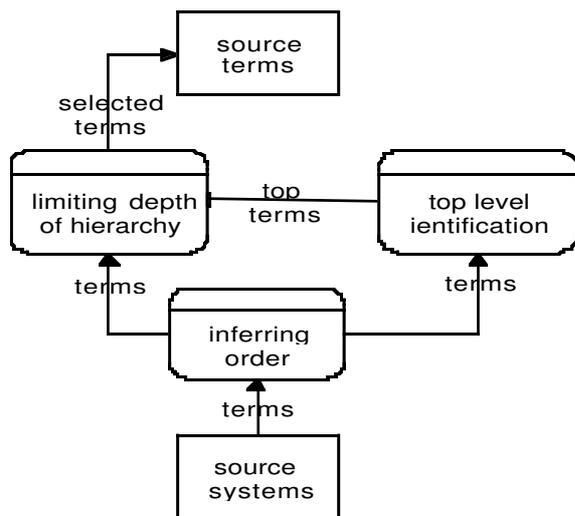


Fig. 3. Phase I: source terms extraction

#### 4.2. The medical sources

Up to now, our research has taken into account 5 sources (terminological repositories): the UMLS semantic network<sup>13</sup> (all ~170 semantic types and relations, and their defined combinations), SNOMED-III<sup>3</sup> (~600 most general concepts) and GMN<sup>7</sup> (~700 most general concepts) nomenclatures, ICD10<sup>24</sup> classification (~250 most general concepts), and the CORE model previously developed by the GALEN project (version 5g<sup>8,19</sup>, non-ontologically oriented, all ~2000 items).

UMLS has a hierarchical structure, includes relations, provides free-text definitions and combinations of types

and relations. It has a browser but does not allow to create new assertions. It uses the MeSH thesaurus and other nomenclatures as its ‘bottom level’.

SNOMED and GMN have some general axes (partially hierarchical), do not apply relations, are homogeneous between top and bottom parts. ICD is hierarchical, has no relations, is homogeneous between top and bottom parts. CORE model v.5g is hierarchical, applies relations, is homogeneous between top and bottom parts, and has a terminological engine which allows to compose concepts and relations—with some degrees of validity—into canonical forms, and provides tools to debug and browse large models.

The analysis allowed us to explicitly formalize different viewpoints on important issues which have heterogeneous ontologies in the sources:

- diagnosis as *process* or *outcome of a process*. For example, SNOMED or ICD identify disease and diagnosis because they are oriented to statistics, records and reports, and a diagnosis is considered as a *diagnostic outcome*. UMLS is oriented towards scientific knowledge, and diagnosis is considered as an inferential process;
- morphology as *form* vs. *outcome of a function* vs. *structure in a given condition*, in particular when a structure is the outcome of a surgical procedure or a pathological process;
- regions, spaces, holes<sup>23</sup> as *conceptual arrangements of structures* (e.g. in UMLS) vs. *structures themselves* (*immaterial objects*).

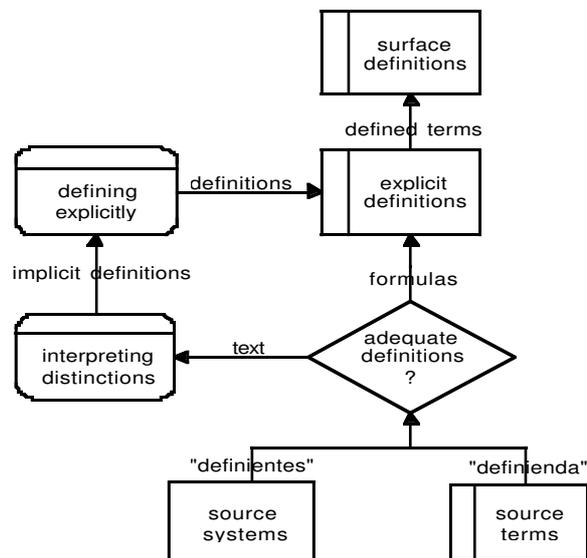


Fig. 4: Phase II: source terms definition.

#### 4.3 Phase II): Local (or “surface”) definition of terms

Once we get a relevant set of terms for each source, we focus on the criteria of classification, i.e., the *local definitions* of terms. Put differently, we should answer

the “definitional” question: which is the difference within a group of homogeneous concepts from the same source, typically between two children-concepts of the same parent concept? (cf. also <sup>28</sup>).

From a definitional viewpoint, terms to be defined are “definienda”, and defining terms are “definientes”. The problem is that very often sources have informal, or poor definitions, and sometimes they lack at all.

When definitions are lacking (when they are implicit), we ask the abovementioned definitional question, and create a sound explicit definition.

Implicit definitions can be inferred with different reconstruction processes, depending on the structure of the terminological source:

— *local definition is implicit, but inferable from extension*, as in SNOMED. These criteria are extensionally inferable because they are suggested by their children and their position in the hierarchy:

```

Topography
  Cardiovascular System
    Heart and Pericardium
    Blood Vessels
  
```

— *local definition is thoroughly implicit*, as in natural language texts: these can be only reconstructed:

... respiratory failure ... respiratory dysfunction resulting in abnormalities of oxygenation or CO<sub>2</sub> elimination severe enough to impair or threaten function of vital organs...<sup>26</sup>

When definitions are explicit in the source, two different strategies are applied depending on which kind of explicitness is shown in the source:

— *local definitions are dictionary-like explicit*, as in textual definitions of UMLS:

Activity: An operation or series of operations that an organism or machine carries out or participates in

— *local definitions are formally explicit*, as in GALEN 5g, which exploits the GRAIL formal subsumption <sup>18</sup>:

```

(BodyProcess which hasOutcome
AreaOfPolyposis) name PolyposisProcess
  
```

which states that a *polyposis* is a process of the body and is defined by its usual outcome: an *area of polyposis*.

#### 4.4 Phase III): Multi-local (or “shallow”) definition of terms: triggering theories related to distinctions made in local definitions

Local definitions imply an ontology for each source. We state that these ontologies are *surface* ontologies. Our purpose is the enrichment of surfaces by triggering general (global) ontologies.

Such an enrichment is not an arbitrary choice: it is made in order to connect heterogeneous local definitions within an ontological *slot* (see below the *ontological frame of a state of affairs*).

It is like filling the lexical gaps among different languages: where English has *wood*, Italian has *legno* (*as matter*), *bosco* (*as aggregate of trees*), *foresta* (*as wide, heterogeneous aggregate of trees*). Not that English lack the Italian ontology: it only does not let it emerge in the lexicon of words (but one can paraphrase it). Filling gaps is a finite, strictly decidable work, thus the stopover problem does not come back in the form: “who knows how much of a general theory should I include in the integrated model?”.

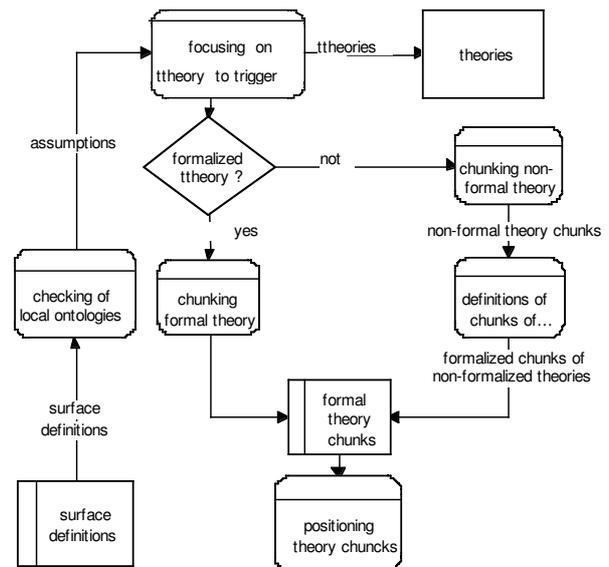


Fig. 5: Phase 3: enriching local definitions according to general theories

This amounts to make an analysis of the contextual framework of local definitions. The setting also implies investigating a large heuristics (with several degrees of formality): experts' knowledge, general theories (either formal or non-formal ontological statements), cognitive constraints which structure our knowledge: *spatial orientations*, *kinaesthetic schemata*, *naïve force dynamics*, etc. This heuristics interacts with the ontological frame of a state of affairs (see below) in order to locate appropriate theories and appropriate chunks of them which may allow to fill the gap.

A state of affairs may include:	ON8 top-level includes:	ON8 maps content categories to:
Structure	: Structure : ...	sorts
Process	: Process : ...	sorts
Actantial roles	ActantialRoles	binary predicates
Agent	actsOn	=
Performer	performs	=
Instrument	isInstrumentFor	=
Patient	—	—
Affected	embodies	=
Receiver	receives	=
Cause	isACauseOf ...	=
Outcome	isOutcomeOf ...	=
Goal	isGoalOf	=
Interpretation	—	
Signs	: Sign : ...	sorts
Typicality	hasTypicality ...	unary predicates
Assessment	hasAssessment ...	binary to ternary predicates
Cognitive Schemas	—	—
Space	hasSpatialDisposition	predicates
Topology	hasConnectedness ...	unary to ternary predicates
Mereology	hasParthood ...	binary predicates
Position	hasPosition ...	binary to ternary predicates
Direction	hasDirection ...	unary predicates
Time	hasTemporalDisposition	predicates
Topology	hasTemporalConnectedness ...	binary predicates
Value	hasTimetable ...	=
Dynamics	hasDynamicDisposition	unary predicate
Balance	hasBalance	=
Quantitative Schemas	—	—
Scales	hasScale ...	unary predicates
Dimensional Patterns	hasDimensionalPattern ...	=
Physical concepts	hasMaterialDisposition	predicates
Physical State	hasPhysicalState ...	unary predicates
Physical Category (or “Quality”)	hasPhysicalCategory ...	unary to binary predicates
Morphology	hasMorphology ...	binary predicates
Composition	hasComposition ...	unary to binary predicates
Ontologic Layers	hasLayer ...	unary predicates
Context	: Context	sorts
Situation	: Situation	=
SpatialRegion	: Region	=
Domain	: Domain	=
Text	: Text	=

Tab. 1: The ontological frame of a state of affairs with current mappings in ON8 top-level and logical categories used. Indentation is used for sub-sorts and sub-relations. The name of a sort is preceded by “:”.

#### 4.5 Phase IV): Multi-local (or “shallow”) definition of terms: triggering theories for top-level categories design

A procedure analogous to Phase III is made at Phase IV for deciding on a top-level which is super-imposed to the integrated formal ontology model we want to obtain. This is a subjective work, depending on the “taste” of the ontological engineer. It should also be proposed as a hypothetic and easily modifiable taxonomy. Our current top-level has explicit mappings to a frame of a state of affairs as envisageable in the medical domain (see Tab. 1).

#### 4.6 A guide in the general ontology forest: the ontological frame of a State of Affairs.

We apply here the distinction sketched by Lehmann<sup>57</sup> between Objects (such as a thing or an event) and Determinables (such as color, or height), where a determinable is a second-order predicate describing first-order predicates; moreover, it can take other determinables as values, generating a hierarchy (a third-order structure).

We mapped Objects to sorts in an order-sorted logic, and Determinables to n-ary predicates. Additional (“less relevant”) Objects derive from the application of Determinables to sorted Objects: these last are meant to be *roles*.

A state of affairs has —as its canonical formal semantics— a tuple

$$\langle S, P, C, R \rangle,$$

where  $S$  is the domain of *structures*,  $P$  is the domain of *processes*,  $C$  is the set of possible *contexts* arising as complements of the sum of others intensional entities in a state of affairs, and  $R$  is the set of intensional “relations”, composed by  $\{ a, i, sch, ph, l \}$ , where  $a$  is the set of *actantial roles*,  $i$  the set of *relations of interpretation*,  $sch$  the set of *schematic relations*,  $ph$  the set of *relations expressing physical concepts*, and  $l$  the set of *ontological layers*. In Fig. 6 the top-level of Objects ( $S$ ,  $P$ , or  $C$ ) is shown.

#### 4.7 Phase V): Multi-local (or “shallow”) definition of terms: merging local definitions and top-level categories

Phase V practically merges shallow definitions and top-level: finding direct correspondences among local items and elements of the theory chunks triggered for top level, or emending/enlarging theory chunks to allow local items to have room according to top level.

#### 4.8. Phase VI)

##### Formalization of an Integrated Model

The model is formalized —and eventually implemented— according to the *syntactic* and *pragmatic constraints* of the logic formalism and/or of the computational tool employed. Our current model of medical taxonomies integration, ON8, uses an order-sorted logic syntax and semantics<sup>47, 48</sup> (the top-level presented in the Appendix shows sorts definitions through quantification on sub-sorts, and hierarchically defined relations).

#### 4.9. The current state of the implemented medical ontology as an output of ONIONS

The results of the application of ONIONS to medical taxonomies currently are:

- i) a mapping among ~4000 heterogeneous terms from the 5 terminological sources;
- ii) ON8: a formal domain ontology including the definition of ~2800 high-level items. ON8 is the evolution of ON7<sup>22</sup>, which was modelled in the GRAIL terminological language<sup>8</sup> (as part of the ontological foundation of the CORE model within the GALEN project, cf.<sup>19</sup>).

As abovementioned, ON8 is currently modelled in an order-sorted logic.

- iii) an in-progress implementation of ON8 in the SNePS<sup>59</sup> semantic network propositional system. There also exists a computer-based model of ON7, implemented as a GRAIL application<sup>8</sup> (CORE model v.6f, running on Smalltalk). It includes ~800 computational definitions and allows the

computational mapping of ~1000 heterogeneous items from UMLS, SNOMED, and GRAIL5g, to ON7<sup>9</sup>.

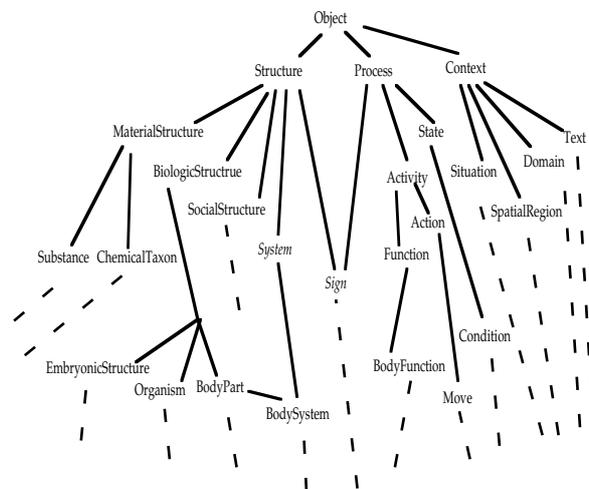


Fig. 6: A top-level sort hierarchy of Objects in ON8

The implementation of ON8 will satisfy the mapping for the overall heterogeneous terms considered, as well as a microdomain ontology for medical procedures (in collaboration with CEN<sup>2</sup>).

Notice that such an integrational work can be seen as the categorial framework for context translation as well. One could talk of, say, *viral hepatitis* in the context of UMLS, or in the context of SNOMED, and then rearrange everything with a formalism such as *context logic* (see<sup>14</sup>, which provides an example of database assumptions integration in the domain of engine parts warehouses). But we have assumed that formal issues come *after* (or at least *accompanying*) ontological integration (cf. also<sup>1</sup>): context logic can be extremely useful for formalizing an ontology which has *already* been integrated. Otherwise, formalizing contexts would result in alternative modules, which are not integrated at all.

On the other hand, we could take advantage of such an approach in case of irreducible ontological incompatibility among sources (which seems uncommon in western scientific communities<sup>b</sup>): this would provide *segregated* pieces of ontologies.

<sup>b</sup> This seems uncommon, but it is possible to envisage some cases, though hopefully not in given taxonomic sources. For instance, the decision about administering a *hysterosalpingography* (a somehow invasive gynecological diagnostic procedure) has very different ontologies in France or in USA (cf.<sup>58</sup>). Notice that this is a problem of inferential ontology, as correctly distinguished in<sup>4</sup>.

## Conclusion

We have presented some topics concerning the difficulties of conceptual convergence among different taxonomical sources, with an application to medicine: health care operators accomplish convergence of different intended meanings through shared high-level theories, context and their operational knowledge. Though, computational integration needs a *formal description of explicit intended meanings*.

The integration task of ONIONS has been accomplished by defining the specific ontologies of five taxonomic sources within the framework of ON8, which holds them together by referring to general ontology notions.

The comprehensive medical domain ontology of ON8 acts as a library (cf.<sup>11,6</sup>) of the more specific domain ontologies which form the base for each source.

The library paradigm is often connected to a claim of *modularity*. ON8 is potentially open to further modules—even if partially not compatible—provided that they can be analysed within a more general ontology framework (e.g., a module on "traditional Chinese" medical record cannot be integrated, because its underlying ontology is not compatible with our "western" general ontology). The modularity hypothesis is being tested through an ongoing experiment which analyzes the categorial structures<sup>2</sup> of concept systems produced by the European Standardization Body (CEN) on *drugs, surgical procedures, laboratory quantities, and medical devices* (in cooperation with CEN/TC251/PT015).

The most interesting point is to verify if and how much of ON8 is stable as far as its iterative applications to other sub-domains of medicine are concerned.

## Acknowledgements

The work has been partially funded by the EC project GALEN (AIM-2012) and by the Italian CNR special project SOLMC: 'Strumenti ontologici e linguistici per la modellazione concettuale' (Ontological and Linguistic Tools for Conceptual Modelling).

## References

- [1] Bernauer J. Modelling Formal Subsumption and Part-Whole Relation for Medical Concept Descriptions. In Workshop on Parts and Wholes: Conceptual Part-Whole Relations and Formal Mereology, ECAI '94, 69-79
- [2] CEN/TC251/PT003. Model for representation of terminologies and coding systems in medicine. in: Proceedings of the Seminar "Opportunities for European and U.S. Cooperation in Standardization in Health Care Informatics", 1992
- [3] Coté RA, Rothwell DJ, Brochu L (eds). SNOMED International, 3rd ed., 4 vols. Northfield, Ill: College of American Pathologists, 1994
- [4] EPISTOL Core Group. Knowledge Processing for Decision Support in the Health Sector. in Barahona P & Christensen JP (eds.): Knowledge and Decisions in Health Telematics, IOS Press, 1994
- [5] Evans DA, Cimino JJ, Huff SM, Bell DS for the CANON Group. Toward a Medical-Concept Representation Language. Journal of the American Medical Informatics Association 1994; 1:207-17
- [6] Falasconi S Stefanelli M. A Library of Medical Ontologies. in Workshop on Comparison of Implemented Ontologies, ECAI 94
- [7] Gabrieli E. A New Electronic Medical Nomenclature. Journal of Medical Systems 1989; 3:6
- [8] GALEN Project. Documentation available from the main contractor Rector AL, Medical Informatics Group, Dept. Computer Science, Univ. Manchester, Manchester M13 9 PL, UK, 1992-1994
- [9] Gangemi A, Steve G, Rossi Mori A. Cognitive Design for Sharing Medical Knowledge Models. in MEDINFO-95
- [10] Gangemi A, Poli R, Steve G. General, Regional, and Domain Ontologies: An Outline. Roma, CNR-ITBM: Technical Report 95-05-01 (1995).
- [11] Gruber T. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 1993; 5:188-220
- [12] Guarino N. Formal Ontology, Conceptual Analysis and Knowledge Representation. In N Guarino & R Poli (eds.) Formal Ontology in Conceptual Analysis and Knowledge Representation. Dordrecht: Kluwer 1995
- [13] Humphreys BL, Lindberg DA. The Unified Medical Language System Project. in Lun KC et al. (eds.) MEDINFO 92. Amsterdam: Elsevier Science Publishers, 1992
- [14] McCarthy J, Buvac S. Formalizing Context. Stanford Un. Tech. Note STAN-CS-TN-94-13, 1994
- [15] Musen M. Dimensions of Knowledge Sharing and Reuse. Computers and Biomedical Research 1992; 25:435-67
- [16] National Library of Medicine. MeSH Medical Subject Headings. Bethesda Maryland: NLM (yearly)
- [17] Neches R et al. Enabling Technology for Knowledge Sharing. AI Magazine 1991; fall:35-56
- [18] Rector A. Compositional Models of Medical Concepts: Towards Re-usable Application-Independent Medical Terminologies. in Barahona P & Christensen JP (eds.). Knowledge and Decisions in Health Telematics, IOS Press 1994
- [19] 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. in Proceedings of 12th International Congress of the European Federation for Medical Informatics (MIE94), 1994
- [20] Rossi Mori A, Gangemi A, Galanti M. The Coding Cage. in Proceedings of 11th International Congress of the European Federation for Medical Informatics (MIE93), Freund Publishing House 1993 466-71

- [21] Sowa JF. Top-Level Ontological Categories. In N Guarino & R Poli (eds.) *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Dordrecht: Kluwer 1995
- [22] Steve G, Gangemi A. Modelling a Sharable Medical Concept System: Ontological Foundation in GALEN. in *Artificial Intelligence in Medicine Europe, AIME95*
- [23] Varzi A, Casati R. *Holes and Other Superficialities*. Boston: MIT Press 1994
- [24] WHO. *International Classification of Diseases 10th revision*. Geneva: WHO 1994
- [25] Fankhauser P, Kracker M, Neuhold E. Semantic vs. Structural Resemblance of Classes. Special issue: *Semantic Issues in Multidatabase Systems, SIGMOD RECORD*, Vol. 20, No. 4, December 1991, pp. 59-63
- [26] Schroeder SA, Krupp MA & Tierny LM. *Current Medical Diagnosis and Treatment*. Prentice Hall, 1988
- [27] Bateman JA, Kasper RT, Moore JD, Whitney RA. *A General Organization of Knowledge for Natural Language Processing: The Penman Upper Model*. Tech. Rep., USC/Information Science Institute, Marina del Rey, CA, 1990.
- [28] Finin, Timothy W. - Constraining the Interpretation of Nominal Compounds in a Limited Context - in R.Grishman e R.Kittredge (Eds.): *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Hillsdale, L. Erlbaum Ass. (1986) 163-173.
- [29] Gangemi, A - Ricategorizzare la memoria. Le categorie tra semiotica e ontologia - in Negrini G (ed.): *Atti del Seminario su "Categorie e modelli della conoscenza"*, CNR, Roma (1995b).
- [30] Guarino N, Boldrin L - Concepts and Relations - in
- [31] Guarino N, Poli R: *Pre-Proceedings of the International Workshop on Formal Ontology*, Ladseb, Padova (1993a) 1-17.
- [32] Guarino, N. 1992. Concepts, Attributes and Arbitrary Relations: Some Linguistic and Ontological Criteria for Structuring Knowledge Bases. *Data & Knowledge Engineering*, 8: 249-261.
- [33] Guarino, N., Carrara, M., and Giaretta, P. 1994. An Ontology of Meta-Level Categories. In J. Doyle, E. Sandewall and P. Torasso (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference (KR94)*. Kaufmann, San Mateo.
- [34] Guarino, Nicola - Ontologies and Knowledge Bases: Towards a Terminological Clarification- in *Proceedings of 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases* (1995).
- [35] Hartmann N. *Zur Grundlegung der Ontologie*. Berlin, de Gruyter, 1966
- [36] Lakoff, George - The Invariance Hypothesis: is Abstract Reason Based on Image Schemas? - *Cognitive Linguistics*, 1, 1 (1990) 39-74.
- [37] Lenat DB, Guha RV. *Building Large Knowledge-based Systems: Representation and Inference in the CYC Project*. Menlo Park, Addison-Wesley, 1990
- [38] Marconi D. *A Metrics for Isolating Terminological Knowledge*. AI\*IA Notizie, 1994
- [39] Peirce, Charles Saunders - On Signs and the Categories - in *Semiotica. I fondamenti della semiotica cognitiva*, Torino, Einaudi (1980).
- [40] Petitot Jean & Smith, Barry - *New Foundations for Qualitative Physics* - in JE Tiles, GJ McKee, GC Dean (eds): *Evolving Knowledge in Natural Science and Artificial Intelligence*, Pitman, London (1991).
- [41] Rosch, E., C. B. Mervis, W. D. Gray, D. M. Johnson & P. Boyes-Braem - Basic Objects in Natural Categories, *Cognitive Psychology* 8, pp 382-439 (1976).
- [42] Saussure, Ferdinand de - *Cours de linguistique générale* - Payot, Lausanne (1906/11) tr. it. *Corso di linguistica generale*, con intr. e comm. di T. De Mauro, Bari, Laterza (1970).
- [43] Simons, P. 1987. *Parts: a Study in Ontology*. Clarendon Press, Oxford.
- [44] Sujansky W, Altman R. *Bridging the Representational Heterogeneity of Clinical Databases*. Stanford Un. Knowledge Systems Laboratory Report KSL-94-07.
- [45] Uschold M, King M. *Towards a Methodology for Building Ontologies*. IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing.
- [46] van Heijst G, Schreiber Ath, Wielinga BG. *Using Explicit Ontologies in KBS Development*. *International Journal of Human-Computer Studies*, to appear
- [47] Oberschelp A. *Order Sorted Predicate Logic*. in Bläsius KH & al.: *Sorts and Types in Artificial Intelligence*, Springer Verlag, 1989
- [48] Cohn AG. *Taxonomic Reasoning with Many-Sorted Logic*. *Artificial Intelligence Review*, 3, 1989
- [49] Cocchiarella NB - *Formal Ontology* - in Burkhardt H, Smith B (eds.): *Handbook of Metaphysics and Ontology* - Munich, Philosophia Verlag (1991).
- [50] Petitot, Jean - *Syntaxe topologique et grammaire cognitive - Langages*, 103 (1991) 97-128 .
- [51] Langacker, Ronald W. - *Concept, Image, and Symbol. The Cognitive Basis of Grammar* - Berlin, Mouton De Gruyter (1991).
- [52] Talmy L. *The Cognitive Culture System*. *The Monist*, 78, 1995
- [53] Poli, R. *Ontologia Formale*. Marietti 1992
- [54] Bouquet P, Giunchiglia P. Reasoning about theory adequacy. A new solution to the qualification problem. In "Fundamenta Informaticae", vol. 23, n. 2-3-4, June-July-August 1995, Don Perlis (ed.).
- [55] Campbell AE, Shapiro SC. *Ontologic Mediation: An Overview*. IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing.
- [56] Knight K, Luk SK. *Building a Large Scale Knowledge Base for Machine Translation*. *Proceedings of AAIL*, Seattle, 1994
- [57] Lehmann F. *Combining Ontological Hierarchies*. in Guarino N, Poli R: *Pre-Proceedings of the International Workshop on Formal Ontology*, Ladseb, Padova, 1993
- [58] Payer L. *Medicine and Culture*. Gollancz, 1988
- [59] Shapiro SC, Rapaport WJ. *The SNePS Family*. In F.Lehmann (ed.): *Semantic Networks in Artificial Intelligence*, Pergamon, Oxford (1992) 243-275.
- [60] Steve G, Gangemi A, Rossi Mori A. *Knowledge Integration of Medical terminological Sources: An Ontologic mediation*. In S.Ali(ed.): *FLAIRS 96 Track on Information Interchange* (1996).