An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies

Aldo Gangemi, Domenico M. Pisanelli, Geri Steve
ITBM-CNR, V. Marx 15, 00137, Roma, Italy
{gangemi,pisanelli,steve}@saussure.irmkant.rm.cnr.it

The paper presents a review of the ONIONS project.
ONIONS is committed to developing a large-scale ontology library for medical terminology. The developed methodology exploits a description logic-based design for the modules in the library and makes extended use of generic theories, thus creating a *stratification* of the modules. Terminological knowledge is acquired by *conceptual analysis* and *ontology integration* over a set of authoritative sources.
After addressing general issues about conceptual analysis and integration, the methodology is briefly described. The central part of the article presents the investigation we have made on the 476,000 medical concepts singled out by the National Library of Medicine as the Metathesaurus™ in the UMLS project. This is followed by several case studies concerning lexical polysemy, the interface between ontologies and lexicon, and other special problems encountered in the specification of the ontologies. A section describing the current structure of the library and the generic theories reused is provided.
Current results of our research include the integration of some top-level ontologies in the *ON9.2* ontology library, and the formalization of the terminological knowledge in the UMLS Metathesaurus.

1. Introduction

The overwhelming amount of information available in various data repositories - especially over the web - emphasizes the relevance of knowledge integration methodologies and techniques to facilitate data sharing. The need for such integration has been already perceived for several years, but telecommunications and networking are quickly and dramatically changing the scenario.

The ever-increasing demand of data sharing has to rely on a solid conceptual foundation in order to give a semantics to the terabytes available in different databases and eventually traveling over the networks. Very often, domains and applications deal with a lack of conceptual foundation.

For example, within the domain of molecular biology, Schulze-Kremer [49] reports an interesting case of the relevance of semantic mismatches. Even an - apparently - unambiguous concept like *gene* may be found conceptualized in different ways in different genome data banks. According to one (GDB), *gene* is a DNA fragment that can be transcribed and translated into a protein, whereas for others (Genbank and GSDB), it is a "DNA region of biological interest with a name and that carries a genetic trait or phenotype".

In the domain of physiology, semantic mismatches can be found even between the two most used terminological repositories: ICD10 [61] and Snomed-III [9]. For example, in ICD10 the terms for *inflammation* are classified as "inflammatory diseases", while Snomed-III has *inflammation* under a separate taxonomy ("morphology") containing properties or structures *produced by* an inflammatory disease.

A standalone application using a local databank or terminological repository may be able to accomplish its task without serious flaws. However, when it is integrated with another application, semantic mismatches constitute a serious obstacle for the agent or interface that is negotiating or sharing information.

The obvious solution to the mismatches would be having a unique, standardized conceptualization for any sense of any lexical item that is used in some domain. It is not an easy achievement and it would be constantly put into discussion by special needs from the users.

An alternative solution is to have a domain-independent, solid conceptual foundation that helps each application, databank, repository, etc. to be represented unambiguously.

In fact, the actual demand is not for a unique conceptualization, but for an unambiguous communication of complex and detailed concepts (possibly expressed in different languages), leaving each user free to make it explicit his/her conceptualization.

Often this task is not an easy one to be achieved, since a deep analysis of the structure and the concepts of terminologies is needed. Such analyses can be performed by adopting a *principled* ontological approach for representing terminology systems and for integrating them in a set of ontologies. The role of ontologies to allow a more effective data and knowledge sharing is widely recognized [17][18].

ONIONS (*ONtological Integration Of Naive Sources)* methodology for ontology integration [54] has been developed since the early 1990s to account for the problem of conceptual heterogeneity. It addresses some problems encountered in the context of the European project GALEN [14] and the Italian projects SOLMC (Ontological and Linguistic Tools for Conceptual Modeling) [21] and ONTOINT (Ontological Integration of Information) [26].

Aims of ONIONS include:

- Developing a well-tuned set of generic ontologies to support the conceptual integration of relevant domain ontologies in medicine. Most medical ontologies lack a semantic foundation, some axiomatization, or ontological depth.
- Integrating a set of relevant domain ontologies in a formally and conceptually satisfactory ontology library to support several tasks, including information access and retrieval, digital content integration, computerized guidelines generation, etc.
- Providing an explicit tracing of the procedure of building an ontology, in order to facilitate its maintenance (evaluation, extensions and/or updating, and intersubjective consensus).

ONIONS methodology exploits: a set of formalisms, a set of computational tools that 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 current main results of ONIONS are: the ON9.2 ontology library; 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.

Some projects are related to ours. Some are mentioned herewith:

- CYC subproject on anatomical microtheories [32] is defining a rich set of relations and axioms for anatomical terminology. CYC has already defined millions of axioms for general-purpose knowledge; thus domain ontologies can reuse a lot of work. On the other hand, its top-level theories are hardly modifiable with some flexibility to account for the needs of special domains. Moreover, due to its idiosyncratic naming policy and its tangleness, CYC top-level taxonomy is commonly known to be cognitively opaque.
- GALEN [48] is a European Community project that is developing a terminology server for medicine that is used to build multi-lingual applications. It also supports some mappings to medical coding systems. It mainly (and overtly) commits to the specification of domain concepts and relations, without much attention to generic theories.
- Snomed-RT [53] is defining a relational structure between the axes (top-level taxonomy branches) of the Snomed nomenclature, by exploiting a description logic. Snomed is commonly recognized as the best taxonomy for medical terminology and it is worldwidely employed in clinical environments.
- MED [63] (and related projects) is aimed at maintaining a controlled vocabulary and supports mapping between terminologies. Mappings are made with a bottom-up approach, which tries to optimize the results without reformulating the sources according to generic theories.

For a wide bibliography concerning the huge field of information integration, with a lot of references to biomedicine, see [24].
The article has the following structure: Section 2 presents the basic definitions of some ontology kinds. Section 3 presents a methodology for conceptual analysis and ontology integration; it firstly introduces a characterization of ontology integration from the conceptual, operational, and practical viewpoints, then the ONIONS methodology is outlined. Section 4 describes our investigation on the nearly half-million concepts singled-out by the National Library of Medicine as the Metathesaurus™ in the UMLS project. Section 5 is a collection of case studies in ontological analysis and integration: there are examples of polysemy, lexical realizations against ontology namespace, complex formal solutions to modeling issues, etc. Section 6 describes the current structure of the ON9.2 ontology library.

2. Kinds of ontologies: Some basic definitions

An ontology is "a partial and indirect specification of a conceptualization". This definition is related to the semantics of ontologies characterized in [19]. An ontology is a set of axioms that account for the intended meaning (the intended models) of a vocabulary. In general, a set of axioms can only approximate such intended models that on their turn can only approximate a conceptualization. A conceptualization is a set of *conceptual* relations that range over a domain and a set of relevant states of affairs (possible worlds) for that domain. Therefore, a precise definition of "ontology" (as used in AI) might be "a partial specification of the intended models of a logical language".

Then again, in a broad meaning, glossaries and vocabularies as well as formal theories that specify a terminology are all considered ontologies. However, when an ontology is not or poorly formalized, with no explicit semantics, its conceptualization is not simply 'approximate', but mostly implicit, since only few natural language cues are available to interpret the intended meaning of the vocabulary.

The degree and type of formalization is consequently a criterion to classify ontologies (in a broad sense):

- *Informal ontological repositories*:
  - *Catalog of normalized terms*, e.g. a list of terms used in the reports from a laboratory: no taxonomy, no axioms, and no glosses.
  - *Glossed catalog*, e.g. a dictionary of medicine: a catalog with natural language glosses.
  - *Taxonomy*, e.g. the SNOMED taxonomy [9] or the UMLS Metathesaurus [39]: a collection of concepts with a partial order induced by inclusion.
- *Axiomatized taxonomy*, e.g. the GALEN Core Model [14]: a taxonomy with axioms.
- *Ontology library*, e.g. the Ontolingua repository [11]: a set of axiomatized taxonomies with relations among them. Each element of the library is a *module*, which can be included into another one. Also, a concept from a module can be only *used* into another one. Ontology modules can be considered subdivisions of the namespace of a model. Modules can also be assigned a *context* semantics, e.g. in CYC 'microtheories' [33].

When ontologies are specified at the most refined formal level - i.e. as modules in a library – a further classification is needed which is based on the generality of the concepts and relations that are defined within a module. The following typology is an elaboration of, among others, Guarino [20] and Van Heijst [59]:

- *Representation* ontologies specify the conceptualizations that underlie knowledge representation formalisms (see theory: *metaontology* in §6.7). Concepts and relations defined in the other kinds of ontology modules are considered instantiations of concepts in the representation ontologies.
- *Generic* ontologies concern the general, foundational aspects of a conceptualization, such as "part", "cause", "participation", "representation". They are usually intended to be domain-independent. It also

seems a good design choice to support multiple alternative theories for the same topic. §6 contains many examples.

- *Intermediate* ontologies contain the general concepts and relations of a domain; e.g. in medicine, "body-part", "tissue", "congenital-abnormality", "treats". In an ideal design, they are used as an interface between domain ontologies and generic ones, but in fact many intermediate ontologies simply act as 'non-generic' domain top-levels. For example, the GALEN Core Model is a top-level for medicine, but with a loose axiomatization of the most general concepts and no reference to generic ontologies.
- *Top-level* ontologies are a particular recipe of generic and intermediate ontology concepts. They must be distinguished from a *lattice of top-level concepts* that is – by convention - any lattice of concepts (or relations), usually with a limited depth: 3 or 4, that contains the most general items of a taxonomy. Such concepts can even be sparse within several modules of a library. A *top-level ontology* is a special case of a lattice of top-level concepts. It is a unique module and is used to stay on top of a domain ontology, or to be a stand-alone, domain-independent theory. For example, the UMLS Semantic Network [28] is a typical domain top-level, the CYC top-level stays on top of a maximally comprehensive set of ontologies, the PENMAN top-level is used to organize a huge natural language thesaurus, etc.
- *Domain* ontologies contain the concepts of a domain or subdomain. For example, the SNOMED taxonomy could be considered an ontology of the medical domain. On the other hand, a more refined definition requires that a domain ontology specializes a subset of generic ontologies in a domain or subdomain, possibly through some intermediate module. For example, an ontology of "fractures" in the ON9.2 library (§6.) would specify a set of concepts related to bone ruptures by including intermediate ontologies such as "anatomy", "biologic-functions", "clinical-activities", etc. that on their turn specialize generic ontologies such as "mereology", "topology", "actors", etc.
- *Task* ontologies describe specific tasks or activities by reusing the vocabulary specified in generic, intermediate, and domain ontologies, e.g. an ontology of the "guidelines for the treatment of breast cancer".

In our opinion, the trademark of a good ontology design is the adoption of richly documented and formalized generic ontologies and a cognitively transparent top-level. Moreover, intermediate modules should contain the most general concepts of a domain, which are specified by *integrating* existing standardization proposals (if any), or experts' knowledge, accordingly to the generic ontologies and the top-level. This creates a *stratified* design of an ontology library.

Defining a domain ontology without this design provides what we call *ad-hoc* ontologies. An *ad-hoc* ontology can be useful for a given task, but is hardly suitable for being shared, reused, or integrated with other ontologies.

In the following section, we describe our methodology for a principled ontological integration of terminological sources.

3. A methodology for ontology integration

3.1 A principled ontology integration

Ontology integration is – generally speaking – the construction of an ontology C that formally specifies the union of the vocabularies of two other ontologies A and B.
Three aspects of an ontology are taken into account: (a) the intended models of the conceptualizations of its vocabulary (see §2.), (b) the domain of interest of such models, i.e. the union of the possible domains of the concepts in the ontology, and (c) the namespace of the ontology.
The most interesting case is when A and B are supposed to commit to the conceptualization of the same domain of interest or of two overlapping domains. In particular, A and B may be:

- *Alternative ontologies*: the intended models of the conceptualizations of A and B are different (they partially overlap or are completely disjoint) while the domain of interest is (mostly) the same. This is a typical case that requires integration: different descriptions of the same topic are to be integrated.
- *Truly overlapping ontologies*: both the intended models of the conceptualizations of A and B and their domains of interest have a substantial overlap. This is another frequent case of required integration: descriptions of strongly related topics are to be integrated.
- *Equivalent ontologies with vocabulary mismatches*: the intended models of the conceptualizations of A and B are the same, as well as the domain of interest, but the namespaces of A and B are overlapping or disjoint. This is the case of equivalent theories with alternative vocabularies.

Some interesting cases occur also when the domains of interest are supposedly disjoint:

- *Overlapping ontologies with disjoint domains*: the intended models of the conceptualizations of A and B overlap while the domains of interest are disjoint. This concerns overlapping theories with different extensions. Actually, it is often the case that some fragments from an ontology A can be reused as components in another ontology B that models a different topic.
- *Homonymically overlapping ontologies*: the intended models of the conceptualizations of A and B do not overlap, but A and B overlap. This is the case of two unrelated ontologies with a vocabulary intersection that – if preserved – generates polysemy: this is one reason to maintain ontology modules.

To be sure that A and B can be *integrated* at some level, C has to commit to both A's and B's conceptualizations. In other words, the intension of the concepts in A and B should be mapped to the intension of C's concepts.

Unfortunately, this cannot be realized using only the conceptual relations specified in A and B for local tasks (for a specific *context*). The methodological principle adopted here is that *generic ontologies* reused from the philosophical, linguistic, mathematical, AI literature must found the comparison of different intensions. Our approach may be called *principled* conceptual integration.

For example: the domain ontology A specifies a concept "body area" with the intended meaning of "loosely specified part of the body that can be cut, filled, etc.". The domain ontology B specifies "body region" with the intended meaning of "region of the body at which body parts are located". A and B approximately cover the same domain of interest; "body area" and "body region" roughly include the same subclasses. How to build an integrated ontology C based on of the given relations only? Do the two intended meanings overlap? What is the place of each one? Is there a preferred one? (for this object/region alternation, see also §5.3).

Formal ontology provides theories that can support integration at the generic level (cf. §3.4, §6.).

Linguistics provides some insights into the way cognitive processes use language, which sometimes prevent us from having the kind of transparency one expects in order to build a logical model. For example, a known mechanism at work in the two different conceptualizations given above is 'metonymy': the activation of a concept by referring to another concept within the same intended model [31][46][58]. Metonymy in our example acts on "body area", whose intended meaning concerns body parts located at some region, although they are denoted by referring to the region ("area") itself. Hence, the metonymic concept has to be distinguished from the plain concept, and correctly related to it.

The distinction between objects ("body parts") and regions, and a notion of localization relation holding between objects and regions are both necessary to make the metonymy clear, and cannot be found in the specifications given in A or B. They have to be found in some generic theory.

The reported example is a case of 'alternative concepts', i.e. concepts with the same domain but overlapping or disjoint intended models. Alternative concepts can also have the same (polysemous) name. Actually, the relationship between conceptual integration and lexical semantics [46] is quite complex (see §3.4, and §5.1, §5.6, §5.7 for related case studies).

3.2 Levels of interoperability

The *interoperability* between two computer systems that use two source ontologies A and B respectively is an important factor to ontology integration, however an integrated ontology C was built, i.e. in a principled way or not.

Interoperability deals with operational integration, not only conceptual integration. In fact, an ontology C' that is not derived from a conceptual integration is often built in order to help a mediation (*information brokering*) between a system based on an ontology A, and a system based on an ontology B.

For example, C' can allow the querying of two heterogeneous databases - based on A and B respectively - by giving the illusion of a common query language [3]. In such a case, the schemata of A and B are generically mapped (mostly 'nearly' mapped) to some concept in C'.
In our perspective, C' would be an *ad hoc* ontology, because it is not based on a conceptual integration. Moreover, conceptual integration would be anyway required, if a complete interoperability is wanted. Furthermore, we defend a *principled* conceptual integration - as the one outlined in §3. – since it is a procedure that allows easier maintenance, negotiation, reusability, and transparency of an integrated ontology.
Anyhow, depending on the amount of change necessary to the operational integration of A and B, different levels of interoperability can be distinguished (for a related discussion, see [56]):

- *Mediation*: it requires no changes to A and B, but only mapping relations that describe the equivalence (partial or total) of A's and B's elements to C's elements. This may result in weak interoperability, since usually the intended models of A and B overlap only: some concepts from A may not have a correspondent in B, and vice-versa. This is the design choice for some recent information management architectures [3][10]. However, such architectures, even recognizing the need of ontological mediation, have nonetheless a weak commitment towards a principled way of conceptual integration (as it is outlined in §3.1), possibly for its additional cost.
- *Alignment*: it requires some change to fill the biggest gaps of A and B respect to an ideal C that completely integrates A and B. Therefore, alignment requires at least a partial conceptual integration. It may support a limited interoperability; for example, deep inferences may be excluded.
- *Unification*: it may require a major reorganization of A and B, which are 'harmonized'. Unification intervenes on the inferential features of the systems, and consists in a complete operational integration: everything can be made in one system, can be made in the other. It results in the most complete interoperability but requires a complete conceptual integration as well. In other terms, from the conceptual viewpoint, unification consists in the adoption of C as a standard in the systems using A or B.

To sum up, an *ad hoc* ontology may be used to support weak interoperability (mediation). A stronger interoperability requires some kind of conceptual integration and the rearrangement of the operational capabilities of the source systems. A *principled* conceptual integration offers more added value to the integrated ontology.
A more complete characterization of ontology integration should take into account many practical issues when selecting sources, extracting terms, analyzing intended meanings, etc., beyond the conceptual and operational aspects mentioned here. In the next section, a list of such practical issues is listed which must be addressed by a methodology for conceptual analysis and integration.

3.3 Common issues in the conceptual analysis of terminologies

From the point of view of an ideal ontology, one suited to be easily integrated, shared, and collaboratively developed and maintained, existing terminologies present several issues. Each one requires a solution from a methodology for ontology integration:

- *Lack of hierarchies*: lists, glossaries, and most dictionaries are not organized taxonomically. Their subsumption hierarchy has to be guessed during conceptual analysis.
- *Ambiguous hierarchies*: the hierarchical link in some thesaurus-like repositories (e.g. MeSH) is multifarious; it may mean "subsumed by", "broader than", "narrower than", "associated to", "part of". The intended meaning of the link must be disambiguated during conceptual analysis.
- *Informality*: most medical repositories are currently informal or contain informal descriptions of terms. Conceptual analysis must deal with the representation and explicitation of informal intended meanings.
- *Lack of modularity*: most terminological ontologies are not modular, neither by task, nor by domain. Ontology integration should modularize the namespace of a domain and separate task-oriented knowledge from domain knowledge.
- *Polysemy*: many terms in poorly formalized repositories are polysemous; many relations are used polysemously – mostly by metonymy (see §5.3 and §5.7). Integration must 'unpack' polysemy, not simply by enumerating senses, but by creating explicit definitions, which often must be properly related one to each other.
- *Uncertain semantics*: 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 not defined at all. After careful analysis, one could consider the templates as default axioms.
- *Prototypical descriptions*: some term descriptions do not allow a clear-cut definition, since their conceptualization can be satisfied by different sets of axioms. These can be formalized by stating different sets of *sufficient* axioms (whereas usual concepts definitions have necessary or both necessary and sufficient axioms).
- *Ontological opaqueness*: lack of reference to an explicit, axiomatized generic ontology, or at least to a generic informal theory. For example, systems in which concepts and relations in the top-level part are non-axiomatized and undocumented are ontologically opaque. If the system is modular, reference to generic theories should lead to a *stratification* of modules.
- *Lack of a (minimal) set of axioms*, which makes it explicit the intended distinctions between siblings: for example, ICD10 shows naked taxonomies, without axioms or even a natural language gloss.
- *Confusing lexical clues*: this is related to the so-called "ontology-lexicon interface". Lexical realizations usually offer the correct conceptual insights to the ontological engineer, but sometimes they are confusing (see the case studies at §5.6 and §5.7).

- *Awkward naming policy*: some formal systems allow purely formal architecture considerations to originate a lot of redundancy and cryptic relation or concept names.
- *'Remainder' partitions*: some terminologies (e.g. Snomed, ICD) use a "NOS" (Not Otherly Specified) flag to talk of a 'remainder' subclass, i.e. a subclass C_n within C that contain all the class instances that are not classified in one of the other subclasses $C_1,\ldots,C_m$ within C.
- *'Exception' partitions*: some terminologies use an "except" flag to talk of an 'excluding' subclass, i.e. a subclass C_1 within C that is explicitly disjoint from another class C_2 (within C or within another class $D \neq C$).
- *Terminological cycles*: some terminologies contain recursive descriptions and even direct cycles. Many implementations of formal languages for ontology specification do not support recursion in concept descriptions (see §4.1).
- *Meta-level soup*: no distinction among kinds of concepts (§6.7). For example, much of the tangleness found in taxonomical repositories is due to the lack of distinction between "types" (like "arsenic") and "roles" (like "poison"). In addition, unary relations (like "abnormal") are usually represented as plain concepts. The adoption of meta-level distinctions greatly enhances the maintenance of large-scale ontologies.
- *Low maintenance capabilities*: difficult accessibility, lack of resources for cooperative maintenance.

Most of these issues are exemplified in §4 and §5. Some important features of ontologies and of their representation and implementation are listed herewith (for a related list, see [49]):

- An explicit taxonomy with subsumption among concepts.
- Semantic explicitness.
- Modularity of namespace.
- A stratified design of the modules.
- Absence of polysemy within a module.
- A proper interface between the ontology namespace and one or more sets of lexical realizations.
- Linguistically meaningful naming policy (cognitive transparency).
- Rich documentation.
- Some minimal axiomatization to detail the difference among sibling concepts.
- Explicit linkage to concepts and relations from generic theories.
- Meta-level assignments to distinguish among the formal primitives assigned to concepts.
- Languages and implementations that support the previous needs as well as the possibility of collaborative modeling.

In the next section, our methodology of ontology integration is outlined. It is supposed to be compliant with the requirements specified as far as here, i.e. from both the conceptual and the practical sides.

3.4 A summary of the ONIONS methodology

ONIONS (*ONtological Integration Of Naive Sources)* is a methodology for conceptual analysis and ontological integration and its products are supposed to support any level of interoperability, if used within appropriate systems (§3.2). For an abstract and comprehensive description of ONIONS, see [54]. The current implementation of the methodology employs Loom [36], a knowledge representation system that supports classification services based on of a quite expressive description logic. ONIONS implementation is meant to provide extensive axiomatization, clear semantics, and ontological depth to a domain terminology. Extensive axiomatization is obtained through a careful conceptual analysis of the terminological sources and their representation in logical languages with a rigorous semantics. Ontological depth is obtained by reusing a library of generic ontologies on which the axiomatization depends. Such library includes multiple choices among partially incompatible ontologies, and a 'metaontology' that states the semantics of the meta-level categories that we adopted to distinguish among the concepts in our library (§6.7). In particular, we suggest the importance of "mereology" or theory of parts, "topology" or theory of wholes and connexity, "morphology", or theory of form and congruence, "localization", or theory of regions, "time" theory, "actors", or theory of participants in a process, and "dependence" (see §6.).

Very briefly, ontology integration in ONIONS is carried out as follows:

- All concepts, relations, templates, rules, and axioms from a source ontology are represented in the ONIONS formalism, currently Loom (see §5.1 and §5.4 for examples of this activity, and §3.3 for a list of related issues).
- When available, plain text descriptions are analyzed and axiomatized (for extensive examples of axiomatization from informal descriptions, see [54]).
- Such intermediate products are integrated by means of a set of generic ontologies. This is the most characteristic activity in ONIONS, which can be briefly described as follows:
    - For any set of sibling concepts, the conceptual difference between each of them is inferred, and such difference is formalized by axioms that reuse - if available - the relations and concepts already in the library. If no concept is available to represent the difference, new concepts are added to the library.
    - For any set of polysemous senses of a term, different concepts are stated and placed within the library according to their domain and to the available modules. Polysemy occurs when two concepts with overlapping or disjoint intended models have the same name. A relevant subset of polysemous phenomena is described in §4.2.
    - Often, polysemous senses of a term - as well as different 'alternative' concepts - are metonymically related. These are called 'alternations' in lexical semantics. In [46], several kinds of them are described: process/outcome (as in the *inflammation* example in the Introduction), region/object (as in the *body region* example in §3.1), and others less frequent in the medical

domain. Alternations must be properly defined by making it explicit the relationship between them: e.g. "has-product" for *inflammation*, "location" for *body-region*. In fact, the conceptual analysis of alternations is already a case of conceptual integration.

- When stating new concepts, the links necessary to maintain the consistency with the existing concepts are created. If conflicts arise with existing theories, a more general theory is searched which is more comprehensive. If this is impracticable, an alternative theory is created.
- Relevant integration cases. Since ONIONS requires the use of generic theories to axiomatize alternative theories (§3.1), the integration of a concept C from an ontology O is performed by comparing C with the concepts $D_{1,...,n}$ already present in the evolving ontology library **L**, whose ontology set $M_{1,...,n}$ contains at least a significant subset of generic ontologies and the set of intermediate and domain ontologies at that state in the evolution of L. The following cases appear relevant to the methodology (see also §3.1):
  - C's name is polysemous in O. It means that, during the previous phases of the methodology, C has not been properly analyzed.
  - C's name is a homonym of a D_i. Homonymy occurs when both the intended models and the domains of two concepts with the same name are disjoint. Homonyms must be differentiated by modifying the name, or by preventing the homonyms to be included in the same module namespace. Languages supporting multiple name assumption can manage homonymy, though.
  - C's name is a synonym of a D_i. Synonymy is the converse of homonymy and occurs when two concepts with different names have both the same intended model and the same domain. Synonyms must be preserved, or included in the set of lexical realizations related to the concept.
  - C is a subset of some D_i in **L**., but has no total mapping on some D_j in **L**. The gap must be filled by adding C as a subconcept of D_i.
  - C is an intersection between two concepts D_i and D_j in **L**. Several subcases may occur; each one must be handled appropriately: see §4.2 for a typology and some solutions.
  - C has an alternative concept D_i in **L** (same domain, but overlapping or disjoint intended models). This case is non-trivial: its motivation should be analyzed:
    - if C metonymically depends on D_i, C is properly related to D_i (see above the treatment of 'alternations');
    - if C and D_i are different viewpoints on the same domain of interest, both concepts are kept, if the case, they are included in separate modules;
    - if the intended model of C is finer than D_i's, D_i is substituted with C;
    - if the intended model of C is coarser than D_i's, C is ignored (but track of it is kept for future mapping).

- C has an alternative concept D_i in **L** with the same name, i.e. the name of C is polysemous in O **L**. We follow the same procedure as for alternative concepts with different names, but names are managed appropriately to the system requirements.
- The library of generic, intermediate, and domain ontologies should be *stratified*, say domain modules should include intermediate modules - that should include generic modules - so that each set of modules can be plugged or unplugged from its more general set without affecting the coherence of the entire library (Fig. 1).

- The source ontologies are explicitly mapped to the integrated ontology, in order to allow (partial) interoperability. The only allowed mappings are *equivalent* and *coarser equivalent*. Formally (cf. §3.5 for concept semantics): for any source ontology SO and an ontology IO that is supposed to result (also) from the integration of SO, for any concept C_i in SO, there is a D_i in IO such that $C_i^I = D_i^I$ (equivalence of possible interpretations), or there is a disjunctive concept (*or* D_i D_j) in IO such that $C_i^I = D_i^I$ D_j^I (equivalence of possible interpretations to a disjunction of concepts – i.e. to a finer concept).
- Partial mappings must be already resolved through the methodology: if any, some step in the integration procedure must be iterated.

Moreover, two aspects seem critical in the development of integrated ontologies:

- Bottom-up modeling vs. top-down specification. Our project is involved in a twofold effort to define comprehensive and useful intermediate ontologies for medicine: the first effort is a top-down specification of medical concepts and relations by specializing generic theories. This effort receives further input from the second one: a bottom-up modeling of large domain terminologies, as the UMLS Metathesaurus (see §4.).
- Which generic theories? When developing domain ontologies, it is still unavoidable reusing a mixture of well-established, uncompleted, and home-cooked theories. In particular, each theory is at least partly 'customized' when it is formalized or translated into another formalism and enters the library. Both design and formal issues require such customization. The stability of the corpus of 'reference theories' should be appreciated by the future community of ontology users.

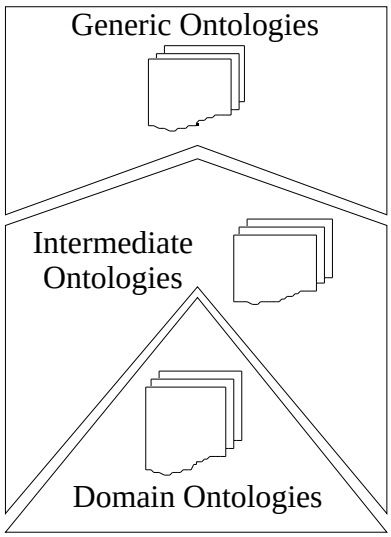

Figure 1
The stratified design of an ontology library after the application of the ONIONS methodology: domain ontologies are plugged into intermediate ontologies that are plugged into a set of generic ontologies. The 'plug-in' metaphor is a simplification, since each ontology module has relations of inclusion or use with several modules in the higher plug-ins (see also Fig.7).

3.5 The tools

Currently there are sophisticated systems that provide services, such as formal contexts [38], and concept classification [5], which greatly help the development of domain ontologies, especially if they are supposed to reuse generic theories.

In our research, we have used two languages:

- Ontolingua [11], derived from KIF [40], is principally aimed at *annotation* of ontologies in a very expressive syntax. It supports several translators to other languages, but does not have a real inferential capability.
- The Loom knowledge representation system [36] implements the Loom language, a description logic that supports *structural subsumption*, both TBox and ABox expressions, transitively closed roles, role hierarchy, implications (non definitional axioms), default axioms, a modular organization of the namespace, etc.

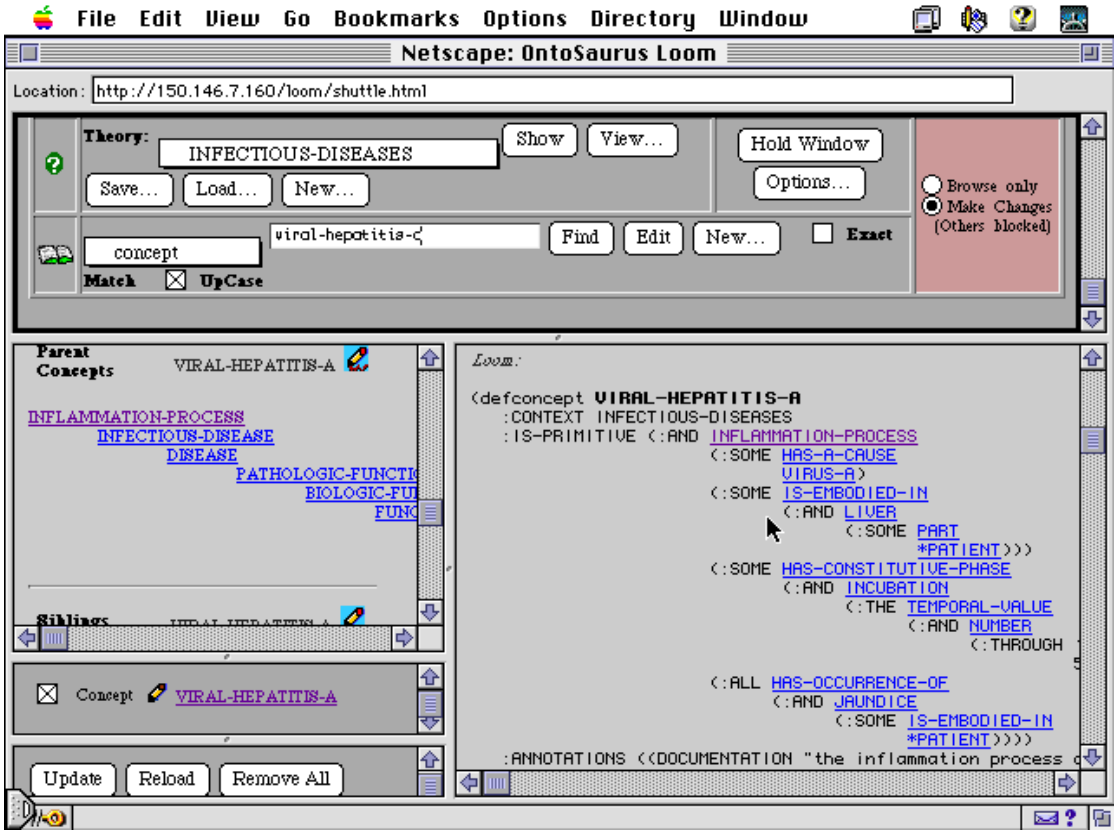

Figure 2: The Loom taxonomy and definition for "viral-hepatitis-A" by means of Ontosaurus.

The most used Loom constructs are summarized in Table 1. A semantic characterization of subsumption in description logics is the following: formally, a concept *A* is *subsumed* by a concept *B* relative to a set of terminological axioms and restrictions T, written $A \leq_T B$, if the interpretation of *A* is necessarily included in the interpretation of *B*; i.e., $A^I \quad B^I$ for all possible interpretations I that satisfy the restrictions specified by T.

From the viewpoint of maintenance and semantic validity, description logics seem particularly suited for ontology development, since they provide consistency checks and subsumption tests for concept constructions, although they are feasible only when the expressivity is equivalent to a fragment of first-order logic (cf. Tab. 1). Consequently, the principal issue in the choice and use of description logics is the trade-off between expressivity and tractability. Loom leans towards the expressive side, but we have employed it for classifying thousands of complex concept definitions with no computational flaws. For a review of description logics, see [5].Implementations of both Ontolingua and Loom languages allow HTML translation and browsing facilities. In particular, Ontosaurus [55], an interface to Loom through the CL-HTTP server [37], is particularly appropriate for allowing a cooperative development of ontologies. An example of a Loom definition accessed via Ontosaurus is shown in Figure 2. The examples in this article are given in the Loom language.

LOOM	SET-THEORETIC SEMANTICS
(and A B)	$A^I \cap B^I$
(or A B)	$A^I \quad B^I$
(not A)	$\Delta^I \setminus A^I$ (Δ^I is the domain of interpretation)
(all R A)	$\{i \ \Delta^I \mid j.(i,j) \quad R^I \quad j \quad A^I\}$
(some R A)	$\{i \ \Delta^I \mid j.(i,j) \quad R^I \quad j \quad A^I\}$
(exactly n R A)	$\{i \ \Delta^I \mid nj.(i,j) \quad R^I \quad j \quad A^I\}$
(filled-by R j)	$\{i \ \Delta^I \mid j.(i,j) \quad R^I\}$
(defconcept A :is-primitive B)	$A^I \quad B^I$
(defrelation R :is S)	$R^I = S^I$

Table 1.Some Loom language constructs and their set-theoretic semantics. A standard specification for description logics is reported in [41].

3.6 Current products of ONIONS

ONIONS methodology has been applied to the integration of the following medical terminologies:

- The UMLS top-level [39] (1998 edition: 132 "semantic types", 91 "relations", and 412 "templates"),
- The SNOMED-III [9] top-level (510 "terms" and 25 "links"),
- GMN [13] top-level (708 "terms"),
- The ICD10 [61] top-level (185 "terms"), and
- The GALEN Core Model [47] (2730 "entities", 413 "attributes" and 1692 terminological axioms).

ONIONS has also been applied to the integration of various sub-domain catalogs and taxonomies. Current products of ONIONS include:

- The ON9 ontology library, v. 9.2, including a set of 50 ontologies with about 1,500 concepts. The modules include generic theories used in the integration of medical terminologies, and the medical intermediate ontologies resulting from the integration. ON9 is available in both Ontolingua and Loom languages.
- The formal translation of a set of medical terminological repositories, including the 476,000-concept UMLS Metathesaurus™, which already allows a mediation between several large terminologies under a small top-level.
- The IMO (Integrated Medical Ontology), an evolving library that enables the alignment of the terminological sources. IMO supports an alignment of the integrated top-levels.

The formal translation of the UMLS, coupled with the ON9.2 library, allowed the classification of such a very large corpus, and the inheritance of axioms defined within ON9.2. The hard work now concerns the distribution of the corpus in a large set of sub-domain ontologies to populate the IMO, and the definition of more specialized axioms.

In the next section a description of the investigation made on the UMLS is presented.

4. Conceptual analysis of the UMLS Metathesaurus™

We are investigating the corpus of concepts from the UMLS Metathesaurus™ (results reported here are from the 1998 edition) [43]. The National Library of Medicine (NLM) in the United States has collected several millions of medical terms from various terminological sources (including Snomed, ICD, etc.), and has singled out more than 470,000 preferred terms in English in the context of the Unified Medical Language System (UMLS) project [39]. Preferred terms are chosen among the lexical variants of terms, and are labeled by NLM "concepts", each one having an alphanumeric "CUI" (Concept Unique Identifier).[1] The UMLS Metathesaurus is extensively used in various projects dedicated to Web sites retrieval, such as *Medical World Search* [27], to database intelligent querying, such as *Grateful Med* [23], to the development of middleware components for enterprise information management [57], etc. On the other hand, for best use in intelligent information integration, the Metathesaurus should have a formal and conceptually rigorous structure, which can be obtained only with the appropriate logical and ontological tools.

Heterogeneity of information in data base schemata or in other semi-formal information repositories is due to the different intended meanings of the terms that constitute the information in the repository. Such inherent polysemy of terminological information is coupled by polysemous taxonomical placements within existing medical terminologies. As we show herewith, polysemy is widespread in the UMLS Metathesaurus as well.

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

- The preferred names of the CUIs (e.g. "Fibromyalgia").
- The instances of IS_A relations between different CUIs that UMLS took from its sources (e.g. "Fibromyalgia" IS_A "Muscular-diseases").
- The relationships between CUIs that UMLS took from its sources (e.g. "COPAD protocol" USES "Asparaginase").
- The instances of IS_A relations between a CUI and its 'semantic types' (e.g. "Fibromyalgia" IS_A "Disease-or-Syndrome").
- The definition of the CUIs in plain text, as reported in authoritative sources such as medical dictionaries.

It should be pointed out that UMLS stated IS_A relations between CUIs only for a minority of CUIs (e.g. "Muscular-diseases"). About 43,000 instances of IS_A relationships have been explicitly stated in

[1] It should be pointed out that "concept" for NLM is not necessarily the same as "concept" in disciplines like logic, ontology, and conceptual modeling. In fact, a UMLS concept may have several conceptualizations, as we show in this section. Actually, the NLM "concept" means "preferred term".

the Metathesaurus, but we stated 318,385 more tuples as IS_A instances on the basis of an analysis of the available sources.
Moreover, UMLS assigned every CUI one or more semantic types, thus stating about 604,755 assignments of a semantic type to a CUI. The 132 semantic types are defined in a semi-formal top-level ontology called 'semantic network'.

Starting from the database - which systematizes the UMLS definitions without further assumptions - for each CUI, we generated a Loom expression. The following example states that "Acute bronchitis NOS" is subsumed by ("`:is-primitive`") certain other concepts:

```
(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))
```

The 476,307 Loom expressions generated from the 1998 UMLS sources concerning CUIs were automatically classified and this process has been helpful in the creation of a consistent model.

4.1 Cycle detection

Some subsumption cycles have been detected in the UMLS corpus. E.g., 523 cycles were found in the taxonomies defined by the UMLS sources: Read, Snomed, etc.
For example, in UMLS, "Adverse reaction to insulins and antidiabetic agent" both subsumes and is subsumed by "Adverse reaction to chlorpropamide", where "chlorpropamide" is a kind of "antidiabetic agent".
In this case, the solution is to maintain that "Adverse reaction to insulins and antidiabetic agents" subsumes "Adverse reaction to chlorpropamide", whereas the opposite is removed from the knowledge base. The reason for this is that, if we define:

```
(defconcept Adverse-reaction-to-chlorpropamide
  "UMLS-CUI C0413593"
  :is (and Adverse-reaction
       (some has-cause chlorpropamide)))

(defconcept Adverse-reaction-to-insulins-and-antidiabetic-agent
  "UMLS-CUI C0413590"
  :is (and Adverse-reaction
       (some has-cause (or insulin antidiabetoc-agent))))
```

and "chlorpropamide" is a kind of "antidiabetic agent", the automatic classifier (Loom in this case) infers that the first is a kind of the second and the inverse is false.
In some cases, the two CUIs are actually synonyms and fail to be normalized into one preferred term, e.g.: "Acinar cell tumor" and "Acinar cell neoplasms", or "Tonsil and other parts of mouth operations"

and "Other specified operations on tonsil or other parts of mouth". Possibly, the original motivation for such cycles is that one of the CUIs has an identifier whose lexical form is usually employed for a special classification purpose (for example, epidemiological classes vs. terms used in a patient record). From a strict ontological viewpoint, such motivation is uninfluential, although it may be relevant for the ontology-lexicon interface (cf. §5.6).

In other places, cycles are due to the presence of partial concept overlapping; for example, "Eczema" subsumes and is subsumed by "Dermatitis". A dermatitis is any inflammation of the skin, while an eczema may mean either dermatitis - but it is not an acceptable diagnostic term – or is an obsolete synonym of "atopic dermatitis", which is a kind of dermatitis.

In cases like this one, the subsumption is evidently uncertain. A possible solution is to distinguish the two meanings of "eczema", and to subsume both under "dermatitis", with some warnings in the documentation or with annotations that can handle the troublesome cases involving the ontology-lexicon interface.

4.2 Polysemous multi-classification of UMLS preferred terms

118,504 CUIs in the UMLS corpus are multi-typed, i.e. CUIs are assigned more than one semantic type. The allowed combinations of semantic types - we call them 'patterns' – result to be 1158, ranging in cardinality (i.e. number of semantic types pertaining to the pattern) from 1 to 6. Table 2 shows figures concerning such patterns.

Cardinality	CUIs	Distinct patterns	Average number of CUIs
1	357803	132	2711
2	108905	714	153
3	9262	358	26
4	331	84	4
5	4	1	4
6	2	1	2

Table 2.
UMLS patterns of semantic types: number of different semantic types in a pattern (Cardinality), number of CUIs pertaining to the patterns with such cardinality (CUIs), number of distinct patterns for that cardinality, and average number of CUIs per distinct pattern.

The individuation of such patterns induces a partition in the Metathesaurus and facilitates its ontological analysis. Some examples of patterns are shown in Table 3.

Pattern name	CUIs
Disease-Or-Syndrome	30601
Disease-Or-Syndrome & Acquired-Abnormality	606
Disease-Or-Syndrome & Anatomical-Abnormality	352
Disease-Or-Syndrome & Classification	15
Disease-Or-Syndrome & Congenital-Abnormality	1169
Disease-Or-Syndrome & Finding	379
Disease-Or-Syndrome & Injury-Or-Poisoning	827

Table 3.
Some patterns of 'semantic types' in the Metathesaurus and the number of CUIs pertaining to them.

We found that many multi-typed patterns are not referable to an actual conjunction of subsumptions (a logical AND). On the contrary, they are motivated by the polysemy of these terms, whose conceptualization can be disambiguated only by distinguishing the contexts in which they are used. Another example is "Onychotillomania" which is classified under "Sign-Or-Symptom", "Individual-Behavior", and "Mental-Or-Behavioral-Dysfunction".

A typology of multi-typing in the UMLS includes the following polysemy kinds:

- *The pattern includes a role*: a set of CUIs has a multi-typing including at least one *role-like concept* that shares a common super-concept with the other concepts composing the pattern. A role-like concept is a 'secondary' concept, whose definition includes transitory or functional features of entities [20][54]. For example, the pattern "Biologically-Active-Substance & Inorganic-Chemical" includes a 'primary' concept like "Inorganic-Chemical", and a secondary concept, "Biologically-Active-Substance", which includes the substances having the functional feature of being "biologically active". A good ontology library should be built according to a metaontology that specifies the meta-level primitives which concepts and relations are instance of. 'Role' is one of such primitives. The combination of roles with primary concepts is not a dangerous kind of polysemy. On the contrary, it should not be considered polysemy at all, since a pattern including a primary concept and a role does not shift from a sense to another, but preserves the primary concept sense in any situation, simply adding the role sense when it is the case. ONIONS methodology supports multi-typing as far as only one primary concept is included in the pattern.
- *The pattern includes two compatible sibling concepts*: a set of CUIs has a multi-typing including at least two compatible sibling concepts that are linked by a hidden relation. By *compatible* here we mean two concepts that can be defining elements in the same definition. For example, the CUIs having the pattern "Amino Acid, Peptide, or Protein & Carbohydrate", which is composed by two sibling sub-concepts of "Organic Chemical", have been analyzed and their pattern can be ontologized as "a protein which contains a carbohydrate". The analysis and integration procedure results in a Loom concept definition as follows:

```
(defconcept |Amino Acid, Peptide, or Protein & Carbohydrate|
  :annotations ((Suggested-Name "carbohydrate-containing-protein")
                (onto-status integrated))
  :is (and protein
       (some has-component carbohydrate))
  :context :substances)
```

  This is a weak form of polysemy that can be handled by making it explicit the hidden relationship holding between the components of the pattern, thus creating a new, more detailed concept.
- *The pattern includes two incongruous concepts*: a set of CUIs has a multi-typing including at least two concepts that are not compatible. For example, "Salmonella-Choleraesuis" is classified both

under "Disease-Or-Syndrome" and "Bacterium", although the salmonella is only the aetiology of a disease called "Salmonellosis". The polysemy originates from the metonymic use of the bacterium name, e.g. in sentences like "the patient is affected by salmonella". This polysemy has mainly a lexical import (the concept "Salmonellosis" may have "Salmonella-Choleraesuis" as a synonym), but other cases show that incongruous multi-typing often reveal new defining elements. For example, the pattern "Body Substance & Disease or Syndrome" is used to classify calculi. It mixes up a structural viewpoint (calculi are a body substance), and a functional viewpoint (calculi can be the product of some pathological function), as the following definition states:

```
(defconcept |Body Substance & Disease or Syndrome|
  :annotations ((onto-status integrated)
                (Suggested-Name "pathologic-calculus"))
  :is-primitive (and Body-Substance
                 (some product pathologic-function))
  :context :abnormalities)
```

As a summary, we list here the strategies used to reclassify polysemous multi-classifications in the UMLS Metathesaurus:

- Maintain multi-classification when one of the pattern conjuncts is a role.
- Integrate the pattern into ONE analytic concept definition when the pattern includes sibling conjuncts or otherly compatible conjuncts.
- Integrate the pattern into TWO or more analytic concept definitions when the conjuncts are neither compatible, nor different viewpoints of the same concept.
- Merge pattern into one of the conjuncts when the other(s) is redundant, or when the other is a subconcept of the first.
- Pattern instances are to be redistributed into various (existing) concepts when they are ontologically heterogeneous with no clear rationale to allow the creation of a new analytic concept.

5. Case studies

5.1 Treating polysemy axiomatically: "ununited fractures" in the Metathesaurus

Beyond multi-typing polysemy, more polysemous phenomena in the UMLS come from multiple subsumption relations among CUIs. For example, the concept "ununited fractures" has the semantic types "Finding" and "Injury or Poisoning", and the IS_A assignments: "fractures" (whose semantic type is "Injury or Poisoning") and "malunion and nonunion of fracture" (whose semantic type is "Pathologic Function"). The graph in Fig.3 results.

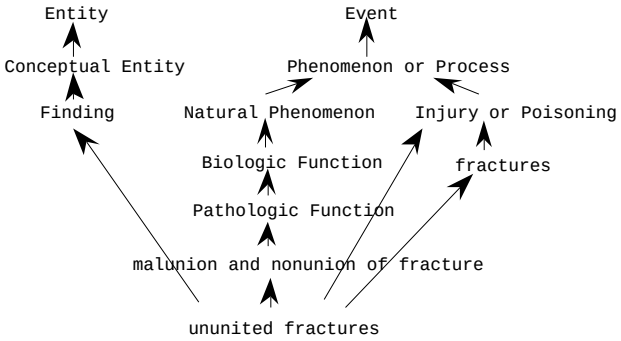

Figure 3.
Arrows mean IS_A, semantic types are denoted by capital letters.

Such graph puts in evidence several ontological problems, at least if ontological analysis and integration are aimed at supporting clear identity criteria [20]. 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", which in most top ontologies would be assumed as disjoint concepts (see §6.1)?
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", 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 follows:

- A fracture of a bone that necessarily *bears* a malunion (a pathology causing a morphological imprecision) or lacks integrity;
- It necessarily *depends on* and *postdates* a fracture resulted from a fracture event;
- It contingently may be an *interpretant* (a sign, see §6.1) *of* some clinical condition.

Therefore, such conceptualization shows only one classification (under "fracture") and three definitional axioms, which provide the identity criterion for the instances of "fractures, ununited".

```
(defconcept |fractures, ununited|
  :is-primitive (and fracture
                 (some morphology
                 (and bone
                   (or (some embodies malunion)
                    (not integral))))
                 (some dependently-postdates fracture)
                 (all interpretant clinical-condition)))
```

We already pointed out that an explicit conceptualization of a terminology needs to be well-founded. For example, in this definition of "ununited fracture", there is a subtle connection between "ununited fracture" and "fracture": an ununited fracture must *postdate* a previous fracture occurring in the same area of a bone,

which is complicated by a malunion. Moreover, an ununited fracture *depends on* that previous fracture.
According to the ONIONS methodology, postdating and dependence must have a definition in some generic theory, in order to be easily understood, reused, and maintained.
A "dependently-postdates" relation is actually defined in our ON9.2 (see §6.2) ontology module: "unrestricted-time", which contains the definitions of many temporal relations that hold for intervals, or processes, or entities in general (like "postdates"). Such distinctions in the domain and range restrictions of temporal relations are motivated by the different identity criteria that different kinds of entities have over time.
"Dependently-postdates" is a kind of "postdates" that also states that the second entity depends on the first for its existence. The definition of this relation makes use of the relation "strictly-depends-on", defined in the ON9.2 theory: "dependence".
A similar line of reasoning can be followed for the other axioms in the example definitions given above. "Embodies" - defined in theory: "actors" - is a special kind of actor relation, meaning that an object is the host of some process; "component" and "portion" – defined in theory: mereology - are two kinds of part relations, etc.

5.2 Rules for parts, locations, and embodiments.

The previous definition of "fracture" also hints a clear statement to an old problem of part-whole reasoning. In medical ontological engineering, it is sometimes mentioned the oddity deriving from the application of the sensible rule (assuming that "part" is transitive):

```
(implies (:composition embodied-in part) embodied-in)
```

From such rule, an "injury embodied in a part of an organ" is an "injury embodied in the organ". For example, "fracture of the phalanx of the thumb" would be "fracture of the arm", since a phalanx is a part of the thumb, which is a part of the arm.
Here the problem actually derives from the assumption of transitivity made on "part". If we use a non-transitive mereological relation, such as "component":

```
(implies (:composition embodied-in component) embodied-in),
```

the inference allowed by the rule will be such that a "fracture of the phalanx of the thumb" is a "fracture of the thumb", but not a "fracture of the arm". Indeed, this one-level inference through mereological relations is commonly accepted and used in everyday language, for example a "fist against the door knob" would be commonly accepted as a "fist against the door", but not as a "fist against the house".
[8] reminds us that transitive inference is used in medical classifications to talk of a "fracture of the phalanx of the thumb" as a kind of "fracture occurring in the upper limb".
Generic ontologies come into our aid to clarify the matter. "Fracture occurring in the upper limb", as well as "fracture of the arm", are metonymical terms that actually mean "fracture of a bone located at upper

limb" (or arm), since "upper limb" and "arm" are body regions, and not body parts. A contribution of a good anatomical ontology is to define relations and rules that can support such metonymy, for instance:

```
(implies (:composition component located) located)
(implies (:composition embodied-in located) located),
```

Such rules allow to infer that:

- if a thumb phalanx is a component of a thumb that is located at an arm, that thumb phalanx is located at that arm as well;
- if a fracture is embodied in a thumb phalanx, it is also located at the arm where the thumb phalanx is located; then
- if "fracture occurring in the upper limb" is defined as a fracture *located* at the upper limb, a "fracture of the phalanx of the thumb" would be classifiable under it; and
- if "fracture occurring in the upper limb" is defined as a fracture of a *bone* located at the upper limb, a "fracture of the phalanx of the thumb" would be classifiable under it as well.

5.3 Ontologizing informal sources: "body-location-or-region" in the UMLS top-level

Our aim is to get an ontologically motivated definition. The original definition from the UMLS top-level ('UMLS Semantic Network' [28]) is firstly translated to Loom. Default semantics is applied to bypass inconsistency with inherited templates, recursive templates, 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 may be "traversed-by" another "body-location-or-region".

Secondly, a consistent and correctly quantified definition is built (§3.3): we use the distinction between definitional (i.e. `:is-primitive`) and implicational (i.e. `:implies`) axioms, and the distinction between "necessary" axioms (the `'some'` clauses) and merely "contingent" axioms (the `'all'` clauses):

```
(defconcept Body-Location-Or-Region
  :annotations ((DOCUMENTATION "An area, subdivision, or region of the body
   demarcated for the purpose of topographical description."))
  :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)
```

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 should go to the first class interpretation. On the other hand, there is a metonymy in medical language by which, when a body region is at hand, also a body part located at that region is at hand. Which parts located at the region are implied results from the operations carried out by physicians, or simply from the functions involved in those parts.

On the other hand, if we adopt the regional interpretation, we should be careful in axiomatizing it. A body region can only exist within an organism ("strictly-depends" on it), but cannot be a generic "part" of it (by the way, 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 that can be located at body regions and (2) restricting the part relations applied to "body-part" (*component*) and "body-region" (*portion*) (both are axiomatized in theory: *meronymy*, see §6.2). The result is:

```
(defconcept Body-Region
  :is-primitive (and Region
                 (some localization^location Anatomical-Structure)
                 (some meronymy^portion Organism))
  :implies (and (some dependence^strictly-depends-on Organism)
                (some topology^connected Body-Region)
                (all localization^location
                 (or Anatomical-Structure
```

```
                  body-substance biologic-function body-region))
                (some meronymy^component
                 (or Body-System Body-Part))
                (all actors^bearer medical-procedures^medical-procedure)
                (all positions^near (:or Anatomical-Structure Body-Region))
                (all positions^crosses-through Body-Region))
   :context :anatomy)
```

5.4 Naming policy and ontologies: "myopathy" in GALEN

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

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

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, "presence", "existence", and "pathological-status" have no axiomatization in the model. Moreover, one is at odds in justifying their inclusion to merely state the simple paraphrase of myopathy as "any disease of a muscle", as can be found in a medical dictionary.

For example, in ON9.2 we could define myopathy straightforwardly as:

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

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

Actually, the above GALEN definition states also 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 [19].

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

Case Study 2 shows the importance of avoiding 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 be introduced to get complete interoperability. The integration of the intended meanings ('alignment') should be sufficient to solve most integration-based problems or at least be preliminary to solving them.

5.5 Coreference in conceptual models

Domain models should be designed to preserve the identity of objects through various manipulations. For example, the identity of an aortic valve replacement and the same valve inserted into or removed from the aorta needs to be preserved.
A solution is to use the same locative relation in all three situations, ignoring the difference between “into”, “in”, and “from”. Such solution is adopted e.g. in the Galen project.
Another solution that is compliant with linguistic usage is to distinguish the three situations, but also providing coreference in order to preserve identity. In first-order logic this is an easy task, while description logics are usually less flexible, because they are variable-less. For example, in Loom we use the following, which is an approximation:

```
(implies (and inserted-into installed-in)
          (same-as inserted-into installed-in))
(implies (and removed-from installed-in)
          (same-as removed-from installed-in))
```

5.6 Localization and anatomical lines: An issue of ontology-lexicon interface

A usual complain directed to 'language-neutral' ontologies from computational linguists is that particular lexicalizations contain conceptual problems that cannot be discovered independently of languages. This is mostly true, although the main issue is designing a good interface between the namespace of ontologies and their lexical realizations.
An issue in the GALEN project concerned the so-called anatomical "laterality". For example, "left nephrectomy” is expressed as “removal of the left kidney” with "left" referring to the anatomical part. But

since there is no anatomical part corresponding to a “bilateral kidney”, “bilateral nephrectomy” must be expressed as a “bilateral removal of kidneys”, where *bilateral* refers to the process “removal” rather than the anatomical object “kidney”.

A different solution is provided by ontological analysis. "Bilateral" simply means "both left and right". Therefore, a bilateral organ is an organ with two quasi-symmetric parts. If an operation is carried out on a pair of quasi-symmetric organs not having a lexical realization as a bilateral organ, this does not prevent us from using "bilateral" as defined. A "bilateral removal of lungs" or a "bilateral removal of kidneys" are both removals of both left and right organs, independently from the existence of an explicitly named "bilateral organ".

On the other hand, an intriguing problem of ontology-lexicon interface concerns the names of lateralized body parts, such as hands or lungs. For example, in "left hand", "left" may be used to describe the position related to the body plane used to distinguish laterality in a conventional presentation of the body (with arms lateral to the trunk, without crossing it), or to describe the position related to some plane, region or part of the body.

At the linguistic level we recognize a further difference, which actually depends on the context. In the case of "left hand", "left" is used in a stable lexicalized phrase (a 'term'), while in "injury at the left of the midline", "left" is used in a less stable proposition, which is usually found in patient records or descriptions. Nevertheless, although there is a different contextualization, the concept "left" has the same intended models in both contexts.

The real problem seems to be in the definition of "left hand", since it can apparently be the case that a left hand is at rest at the *right* of the median line of the body. ONIONS methodology suggests that in any case, when there are two conflicting intended meanings, some difference must be conceptualized and axiomatized.

In this case, the difference is that the naive definition of "left hand" makes a commitment to the left position of a left hand whatsoever, without taking into account the possibility of moving hands all around. This is exactly what we should avoid to solve the conflict. For example, we could use a special relation in the axiomatization:

```
(defconcept left-hand
   :is-primitive (and body-part
                      (some conventionally-located
                       (and body-region (some left-of anterior-median-line)))))
```

another solution is using a more general convention:

```
(defconcept left-hand
   :is-primitive (and body-part
                      (some wholly-located left-upper-limb)))
```

"left upper limb" is hardly to be found in ambiguous sentences, one that would sound like: "the left upper limb is at the right of the median line", at least if we do not take into account disarticulated bodies, which

anyway deserve a special modeling as dependent on some abnormality (medical ontologies should strongly pursue the representation of abnormality).

5.7 Using relation composition to disambiguate verb metonymy

In many domains, metonymy is widely exploited to obtain economy of lexicon and brief sentences. For example, physicians "treat" patients, patient groups, and conditions; therapies "treat" pathologies, abnormalities, and patients; devices "treat" abnormalities, etc.
"Treat" is not ambiguous in the experts' knowledge, but it is metonymically polysemous. Ontological theories should support the definition of relations that refer to the basic meaning of notions like "treat", but also they should reveal the relations 'implied' in the metonymies. Formally, this is ideally accomplished by relation composition.
For example, after fixing the basic meaning of "treats" as ranging over healthcare operators and health conditions, we defined "treatment-action" for activities performed by operators during treatment:

```
(defrelation treatment-action
  :annotations ((DOCUMENTATION "The relation for 'treats' when procedures
   used for treatment are the domain."))
  :is (and clinical-actor
       (:composition performed-by treats)
       (:domain activity)))
```

Similarly, we defined "treatment-method", "treatment-device", "treatment-resource", etc.
An extreme example of this design has been defined in the theory for clinical guidelines, which are special plans describing the method of a medical procedure. Guidelines usually focus on a "population group". The metonymy here is very complex (Figure 4), in fact a "group" is the "target population" of a guideline because it has "members" as parts that are "uniquely located" at some region, which is the location of some "health condition", which is the real target of the procedure that has the guideline as a method.

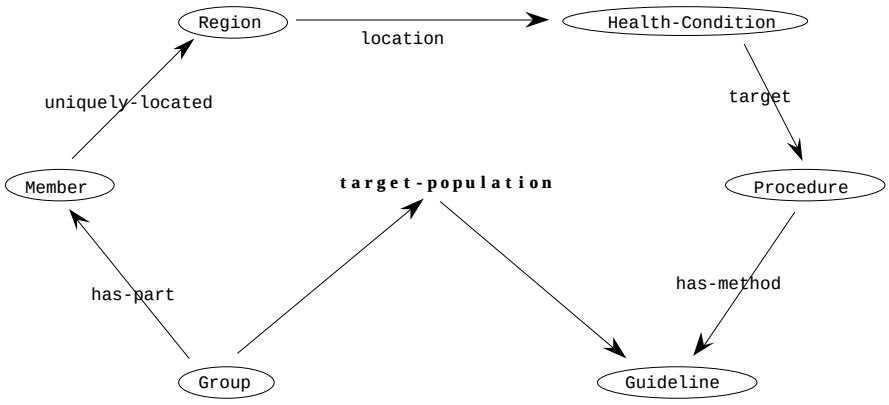

Figure 4
The definition of the relation "target-population" requires the composition of five relations.

5.8 Mixed conceptual, lexical, and formal issues: What is a morphology?

Within the *morphology* taxonomy, the SNOMED-III nomenclature [9] classifies some medical notions, denoting heterogeneous entities - often related to abnormal conditions – that concern some properties, forms and distributions of an organ within the body.

SNOMED-III taxonomy of morphologies is not very articulated: it only distinguishes between "normal" or "abnormal", and "congenital" or "acquired" morphologies. Nevertheless, our conceptual analysis - by exploiting a set of generic ontologies - offers a more detailed classification of SNOMED morphologies:

- a *property* ("color", "consistency", "thickness", "size", "number", "shape"),
- a *condition*:
  - a *topologically* relevant condition:
    - an alteration of *connection* (see Fig. 5):
      - that creates a *configuration* (a new property) in an object ("fracture", "wound"),
      - in the holey interior of an object ("obstruction"),
      - between several objects ("fusion"),
    - an alteration of the *boundary* between an object holey interior and the object complement:
      - creating a *configuration* in the boundary ("cavitation", "ulcer"),
      - producing a substance *flow* ("hemorrhage", "ulcer"),
  - an abnormal *placement* ("dislocation", "ectopia", "absence"),
  - a *form* alteration condition ("deformity", "hyperplasia", "hypoplasia"),
  - a condition involving the alteration of *several properties* ("inflammation", "eruption"),
- an abnormal, foreign *object* ("mass", "neoplasm", "calculus", "obstruction").

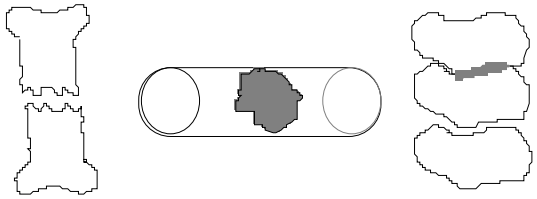

Figure 5
Some examples of connection alteration (left to right): a *fracture* within an object, an *obstruction* within a hole, a *fusion* between objects.

Maybe, the only generalization on this typology is that morphologies are relatively *visible* respect to other aspects of an organism (except plain anatomy). Possibly, their joint classification is due to such functional feature.

The status of morphologies is complicated by the fact that some morphology names are polysemous:

- Both a condition and the function that caused the condition ("inflammation", "ulcer", "fracture", "wound", "hyperplasia"),
- Both an object and the function that produced the object ("neoplasm", "hemorrhage"),

- Both an object O and the condition created in another object O' by O ("obstruction").

Such cases are metonymically ambiguous and are relevant to the ontology-lexicon interface (they are 'alternations', §3.4). For example: "the fracture has been caused by a fall" vs. "the fracture is transverse"; "the obstruction occurred in the jejunum" vs. "the obstruction has been removed".
Conceptual analysis puts into evidence other issues concerning morphologies. The most important is the dependence between a morphological condition, a function, and the related organ.
For example, an "ulcer" (as a condition) of a stomach implies that the stomach *embodies* an *ulceration function* (an ulcer as a function).
Another example is the mereological import of morphologies: some are featured by an organ, some only by a part of an organ. For instance, an "ectopic heart" is wholly ectopic, but an "ulcerated stomach" is only partly ulcerated.
The case about morphologies should have shown that a good definition of domain concepts does often require generic theories. In this case, a set including at least dependence, mereological, topological, and actor relations (§6.).
A further issue with morphologies concerns the representation primitives that should be used to model morphological properties, such as colors, shapes, configurations, etc. For example, assuming the Loom description logic used in the ONIONS methodology, one has several choices (|R| denotes relations, |C| denotes concepts, |i| denotes instances, and |P| denotes properties):

- Morphological properties are instances (of a class), which can fill a dedicated slot, e.g.:
  `(filled-by |R|Has-Color |i|Yellow).`
  This solution does not allow a morphology subsumption hierarchy with kinds of yellow.
- Morphological properties are types (classes) that restrict a dedicated slot, e.g.:
  `(some |R|Has-Color |C|Yellow), where (defconcept Yellow :is-primitive |C|Color).`
  This is good at maintaining subsumption, but morphologies are taken as abstract objects, which is a problematic ontological choice.
- Morphological properties are Loom "properties" (unary relations), e.g.:
  `(|P|Yellow), where (defproperty Yellow :is-primitive |P|Color).`
  This is good ontologically, but creates the formal problem of talking functionally of an implicit morphology as a property, not as (an instance of) a class. There is also the problem that Loom does not maintain a separate hierarchy for properties, which are mixed with concepts.
- Morphological properties are binary relations with a 'boolean' range, e.g.:
  `(exactly 1 |R|Yellow 'T), where (defrelation Yellow :is-primitive |R|Color).`
  This is a little tricky, but allows to maintain a separate hierarchy, and the constraint to be added to the axiom list. Unable to functionally express an implicit morphology.
- Morphological properties are properties or relations, but they also have a *reified* counterpart, e.g.:
  `(defconcept |C|Yellow :reifies |R|Yellow)`

The drawback of this solution is that two kinds of entities must be maintained for one notion. It also sounds tricky from an ontological viewpoint.

The experimented solutions in our research are the last three. Currently, the binary relation solution seems the best balance of pros and cons.

6. A Brief Description of the Generic Theories in the ON9.2 Library

6.1 Top-level concepts

The ON9.2 ontology library is an evolving collection of modules that specify generic, intermediate, and domain ontologies (Fig. 7); they are partly available on-line at our WWW site [25].
The basic notions are defined in a small top-level, and in the generic ontologies described below in the next sections. Alternative theories are allowed.
The module: "top-level" contains the most general distinctions (the upper part in Fig. 6) between the entities that are assumed to have identity criteria in the domains covered by ON9.2 theories.
The top concept is "entity", which subsumes the classic distinction between "occurrents" - "processes", "situations" and "temporal intervals" – and "continuants": "objects" and "regions".
There is a third concept directly subsumed by entity: "sign", used to account for the symbolic use of any entity (see §6.5).
Processes are distinguished into voluntary "acts" and "material-functions" carried out by unintentional objects. Objects are primarily differentiated by the layer of reality to which they pertain: material, biological, etc. (§6.3).
Further distinctions are mainly motivated by their relevance in the biomedical domain. It should be remarked that almost all distinctions in our library are related to some necessity raised by the task of integrating biomedical terminologies.

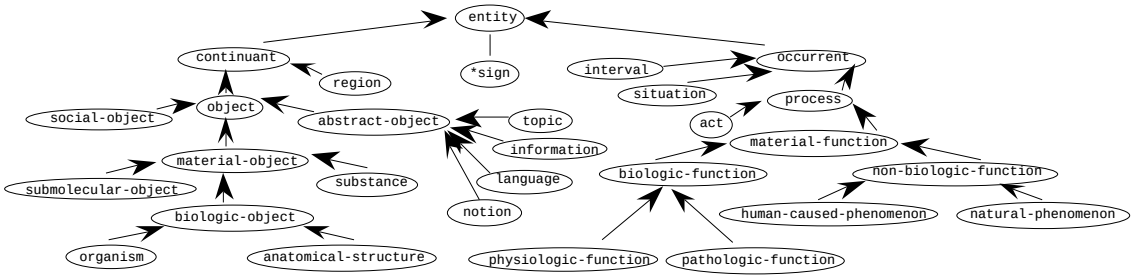

Figure 6
The top-level concepts in the ON9.2 ontology library. They are disjoint classes, except "*sign" (a 'role').

Properties (unary relations), binary relations, and n-ary relations (a small set) have independent top-levels, defined in the theory: "structuring-concepts".

6.2 "Formal ontology" theories

"Formal ontology" theories are the 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 whatever conceptualization. Thus, it has a meaning different from that more or less accepted in AI, where it means "formalized, or semantically explicit, theory".

Formal ontology theories are essential in the axiomatization of other generic theories: *localization*, *morphology*, *actors*, *time*, etc.

Theory: *dependence* is an introductory set of dependence relations, as defined in philosophical work of Simons [51], Varzi [60], etc. A (formal-) ontologically relevant dependence may be causal, physiological, psychological, functional, and proper. The proper dependence is such when something cannot exist without something else.

Theory: *mereology* presents a version of classical extensional mereology that is compliant with both Leonard-Goodman calculus of individuals [34] and Tarski's axioms [56].

Theory: *meronymy* specializes mereology by defining special notions of "whole" and "part" widely used in domain ontologies: societies, collections, systems, etc. Some relations here are not transitive (while "part" is transitive by default); most relations range over specialized domains. The definitions come from the work of Gerstl [16] and others.

Theory: *topology* is a small fragment of classic topology. It defines the basic "connected" relation, various kinds of weak contact, and several of properties of wholes (distributed, self-connected, closed, etc.). Most axioms are a reinterpretation of axiomatizations given by Varzi [60] and by Asher and Vieu [2].

Theory: *topo-morphology* specializes topology by defining special connexity relations used in common knowledge and in some domains, like "attached", "connects", "sequence", "branching", and the relations and properties to talk about various kinds of holes ("cavity", "channel", etc.).

In theory: *equality* we provide some relations involving *identity*. Identity is a much-discussed philosophical matter. In the current ON9.2 library, no particular grounding theory is provided. We only found that some relations are needed to the task of ontological engineering. In particular, we distinguish between:

- equality and difference applied to numbers,
- equality and difference applied to function values ranging on a non-numerical domain,

- equality and difference as partial identity with explicit neutralization of some property (space, time, morphology, etc.), e.g. two situations can be equal but time, two objects can be equal but localization, and
- mereological sameness defined as reciprocal parthood.

6.3 "Stratificational" theories

The "stratificational" theories in ON9.2 are the theory of *layers*, and of *granularity*. They help organizing entities of a domain according to the 'life form' they are about (cf. the Wittgenstein's notion of meaning as basically dependent on the form of life that is producing it [62]). For example, the same object (say, a spleen) has different identity criteria when it is considered at a molecular biological level, or from the macroscopic viewpoint.
Theory: *layers* defines the so-called "strata" [22]: Material, Biological, Psychological, Social, Abstract; this theory also specializes strata according to some scientific granularities [4]: Atomic, Molecular, etc. The basic intuition is that reality is 'layered', and the layers have a complex inter-dependence.
Theory: *granularity* implements Sowa's adaptation [52] of Searle's ontology of intentionality [50], which makes a fundamental distinction between "epistemic" and "actual" identity. For example, a "surgical knife" has an "actual" identity described by its form, material, color, etc., and an "epistemic" identity given by the building and measuring systems that forged and tested its cutting edge. The theory also recognizes an "intentional" identity that pertains to the way the world is considered by the human (or another organism's) form of life. For example, functional aspects of objects pertain to the intentional level: the identity of a "surgical knife" intentionally relies on its particular cutting functionality.

6.4 "Individuation" theories

"Individuation" theories are *localization*, *time*, *and morphology*.
Theory: *localization* axiomatizes regions and some special relations: Exactly-Located, Generically-Located, Partly-Located, Wholly-Located [7]. The main assumption is that every material object is located at some region, and a region is the only entity that can be located at itself. One consequence is that all localizations of an object O that refer to another object P actually refer to the region of P (see theory: positions). This consequence involves that many uses of localization relations – such as "the needle is in the box", or "the artery near the femur" ("femoral artery") are metonymic; to support such use, the composed relation "has-reference-location" has been defined:

```
(defrelation has-reference-location
   :annotations ((DOCUMENTATION "The metonymic use of location: anything can
    be (wholly, partly, or generically, but not exactly) located at some
    entity's region."))
   :is (:and locative-relation
        (:composition located unique-location)))
```

Theory: *position* is a domain application of localization theory: related positions and coordinated positions. It is inspired by the common sense use of linguistic prepositions and some cognitive semantics models [6][35].

Theory: *morphology* contains some basic and anatomical morphology notions: substance composition, morphological properties, etc. This is a difficult field, since few references are available, and morphological notions are strictly related to functional and physical ones. For example, an "inflammation" is a pathological process, but it causes a physical modification involving the change of a morphological property: an organ becomes "inflamed".

There are three different ontologies of time in our library:

- *Temporal-mereology* (originally formulated by Allen [1], see also an adaptation in [52]) uses mereological concepts in its definitions; its relations apply directly to intervals.
- *Unrestricted-time* aims at representing the common sense metonymy by which we use temporal relations ranging over processes and situations rather than intervals. It also defines Kamp's parallel time lines (platforms) [30].
- *Simple-time*, reused from the Ontolingua Server, follows the temporal mereology approach, and also defines some notions for dealing with "absolute" time expressions.

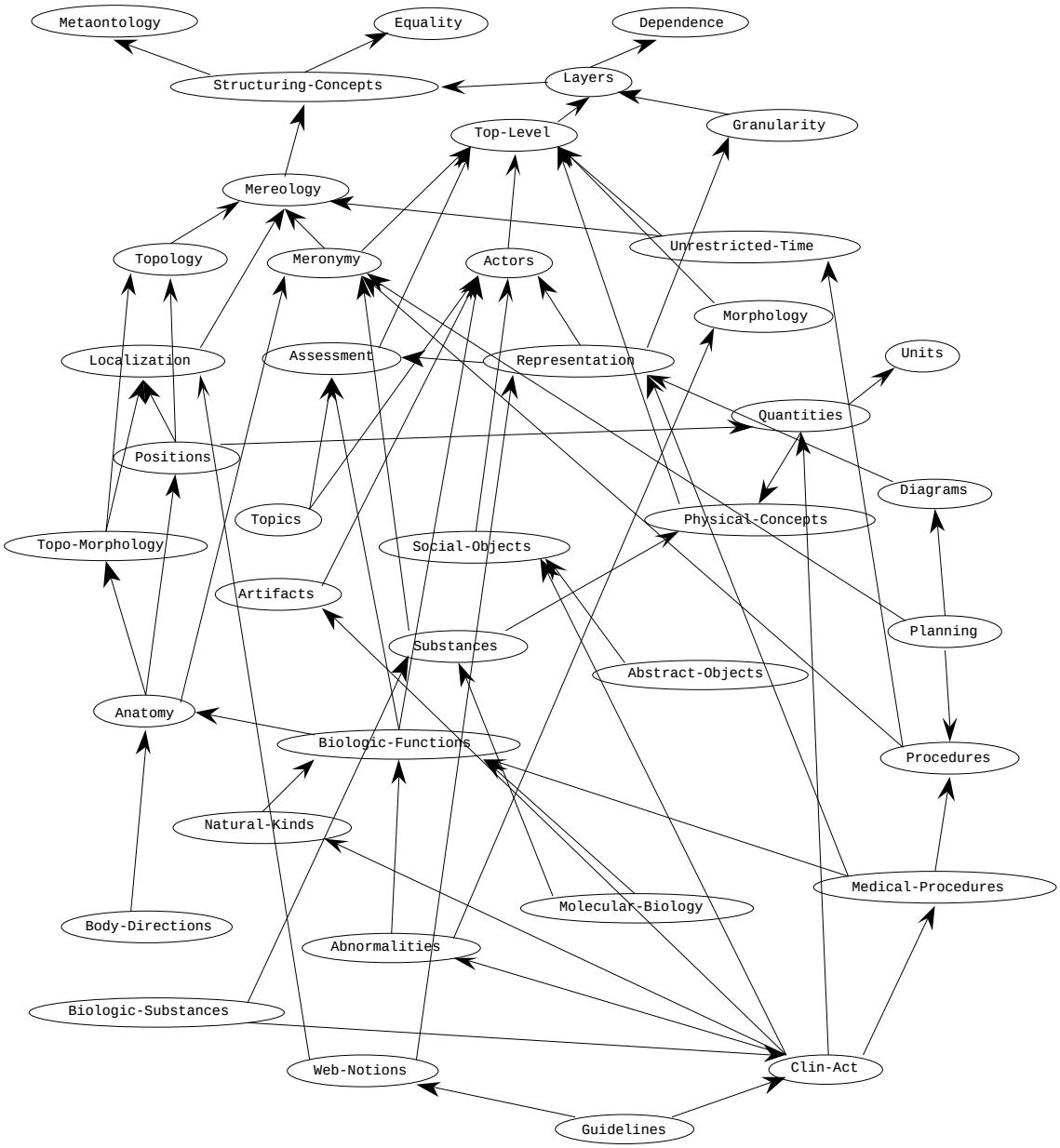

Figure 7
The inclusion lattice of the ON9.2 ontology library. Arrows mean 'includes'.

6.5 "Actors" theory

The theory of *actors* is a summary of various ontologies concerning event structure, taken from linguistics [12], narratology [45], and AI literature. The original suggestion came from Sowa [52], who stresses the relevance of Aristotle's "aitiai" for modeling processes, actors, scenes, situations, scripts, agents, etc. *Actors* has been the most used generic theory for ontology integration in medicine and has been customized to this purpose.

An "actor" is a relation with a range restricted to "process". It allows focalizing on the roles intervening in the development of a process, sometimes called "participants". Four main kinds of actor relations are defined:

- A "from-actor" (cf. Aristotle's "arché") relates a process with an entity involved in the starting part of the process. Such entities are usually active participants in the process. The intuition stands on the cognitive "path-origin" schema [29]. E.g. "performs", "effects", "cause-of".
- A "to-actor" (cf. Aristotle's "tèlos") relates a process with an entity involved only in the conclusive part of the process, or only undergoing the process. Such entities are usually passive participants in the process. The intuition stands on the cognitive "path-destination" schema [29]. E.g. "experiencer", "recipient", "goal", "product".
- An "in-actor" (cf. Aristotle's "ousìa") relates a process with something that 'hosts' the process. The intuition stands on the cognitive "container" schema [29]. E.g. "bearer", "embodies".
- A "by-actor" (cf. Aristotle's "hyle") relates a process with an entity that 'accompanies' the development of a process. The intuition stands on the cognitive "force" schema [29]. E.g. "instrument", "resource", "method".

Moreover, by exploiting relation composition, we have defined several "pseudo-actors": complex relations involving some elementary actor relation; e.g. "affects" is a relation composed by "from-actor" and "has-bearer". See also §6.7.

6.6 "Epistemological" theories

Current epistemological theories are *representation* and *assessment.*

Theory: *representation* includes some relations and concepts related to intentionality, interpretation, symbols, etc. The basic relations are "aware" and "interpretant". These notion derive from semiotics (e.g. [42]): some entity *e* is an interpretant of some other entity *f* when an (aware) agent in a context uses *e* as a defining element of *f*.

Such entity *e* is defined as a 'role' (see §6.7), since anything can be an interpretant of something else. "Interpretant" is then used to introduce relations such as "interprets", "copy", "represents", "judgment", and concepts like "sign" and "information".

Theory: *assessment* includes various relations pertaining to the 'epistemic' aspects of ontology: notions of belief, relevance, conventionality, typicality, and various specialized assessments. 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.

6.7 Metaontology

"Metaontology" is a 'representation ontology' (see §2.). It axiomatizes some meta-level categories on the basis of the work of Guarino [20] 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", etc. The distinctions between unary predicates are especially important. The proper semantic characterization is given in [38].
Intuitively, a "type" is a predicate that can be hardly dismissed by their instances, such as "person", "dog", "hepatitis". A "role" is a predicate that necessarily depends on another predicate: this is the case of entities focused according to an accessory, temporary, or functional aspect, e.g. "hypochondriac", "patient", "hormone". A "property" is a predicate that excludes countability of its instances and also necessarily depends on another predicate: it represents a feature of entities, independently from the actual entities, e.g. "red", "abnormal", "thick". To stress the difference between roles and properties: an "abnormal structure" is a role, not a property.
Other literature used to define generic ontologies includes cognitive semantics "schemas", linguistics notions, and some mathematical and engineering theories (measure units, geometry, algebra, etc.).

6.8 Domain theories

The domain part of the ON9.2 library is still evolving, since it is supposed to include the modules deriving from the ontologization of the UMLS Metathesaurus (§4.). General biomedical concepts and some specialized relations are included in the modules: "natural-kinds", "anatomy", "body-directions", "biologic-functions", "clinical-activities", "medical-procedures", "clinical-guidelines", molecular-biology", "biologic-substances", etc.

7. Conclusions

In this paper, we outlined the research based on the ONIONS methodology: its principles, tools, results, and some case studies. Our research has a twofold purpose. On one hand, it aims at building an explicit, reusable, easily maintainable ontology library for the clinical and biological domains, without focalizing on a specific application. On the other hand, it is immediately exploited in applications such as intelligent retrieval of clinical information (e.g. clinical guidelines over the WWW), and integration of clinical data within hospital departments. The task is indeed a huge one and the ambitious goal of completing a detailed, axiomatized, and modular integration of large terminologies with half-million concepts is still to come.
Nevertheless, our research aims at showing that large-scale integration of terminologically intensive data can take advantage from the framework of formal ontology and lexical semantics. This requires an effort to understand the cognitive basis of the lexicon, the interface between lexicon and conceptual structures,

and the somewhat intricate investigations of philosophy. Then again, as Ludwig von Boltzmann put it: "there is nothing more practical than a good theory".

Acknowledgements

We wish to thank Nicola Guarino for his precious suggestions and useful hints and the anonymous referees whose remarks gave us the opportunity of improving our paper.

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] Baker PG, Brass A, Bechhofer S, Goble C, Paton N, Stevens R, "TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview", in *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, ISMB98,* Montreal, 1998
[4] Blois M, *Information and Medicine*, Berkeley, University of California Press, 1980.
[5] Borgida A, "Description Logic in Data Management", *IEEE Transactions on Knowledge and Data Engineering* 7(5): 671-682, 1995.
[6] Cardona GR, *I sei lati del mondo,* Bari, Laterza, 1985.
[7] Casati R, Varzi A, "The Structure of Spatial Localization", *Philosophical Studies*, 82, 1996.
[8] Ceusters W, Buekens F, De Moor G, Waagmeester A, "The Distinction between Linguistic and Conceptual Semantics in Medical Terminology and its Implications for NLP-Based Knowledge Acquisition", in C Chute (ed): *Proceedings of IMIA WG6 Conference on Natural Language and Medical Concept Representation* (IMIA WG6, Jacksonville, 1997) 71-80.
[9] Coté RA, Rothwell DJ, Brochu L, eds. *SNOMED International* (3rd ed.), Northfield, Ill, College of American Pathologists, 1994.
[10] Duschka OM, and Genesereth MR, "Infomaster - An Information Integration Tool," in *Proceedings of the International Workshop "Intelligent Information Integration"* during the 21st German Annual Conference on Artificial Intelligence, KI-97. Freiburg, Germany, 1997.
[11] Farquhar A, Fikes R, Rice J, “The Ontolingua Server: a Tool for Collaborative Ontology Construction”, *Proceedings of Knowledge Acquisition Workshop,* Banff, participants edition, 1996.
[12] Fillmore CJ, "Frames and the Semantics of Understanding" *Quaderni di Semantica*, 6, 2, 1985.
[13] Gabrieli E, “A New Electronic Medical Nomenclature”, *J. Medical Systems,* 3, 1989.
[14] GALEN Project, documentation available at the URL: http://www.cs.man.ac.uk/mig/galen
[15] Gangemi A, Pisanelli DM, Steve G, "Ontology Alignment: Experiences with Medical Terminologies", *FOIS 98*, Guarino N (ed), Amsterdam, IOS-Press, 1998.
[16] Gerstl P, Pribbenow S, "Midwinters, Endgames, and Body Parts: A Classification of Part-Whole Relations". *International Journal of Human-Computer Studies*, 43, 1996.

[17] Goñi A, Mena E, Illaramendi A, "Querying Heterogeneous and Distributed Data Repositories Using Ontologies", in Charrel JP et al. (eds.) Information Modeling and Knowledge Bases IX, Amsterdam, IOS Press, 1998, 19-34.

[18] Guarino N (ed.), *Formal Ontology in Information Systems*, Amsterdam, IOS-Press, 1998.

[19] Guarino N, "Formal Ontology and Information Systems" in N Guarino, *Formal Ontology in Information Systems*, Amsterdam, IOS Press, 1998.

[20] 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.

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

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

[23] http://igm.nlm.nih.gov

[24] http://salmon.cs.tu-berlin.de:8888/leser/research/bib.html

[25] http://saussure.irmkant.rm.cnr.it

[26] http://saussure.irmkant.rm.cnr.it/onto/ontoint.html

[27] http://www.mwsearch.com

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

[29] Johnson M, *The Body in the Mind*, Chicago UP, 1989.

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

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

[32] Lehman F, Foxvog D, "Putting Flesh on the Bones: Issues that Arise in Creating Anatomical Knowledge Bases with Rich Relational Structures", AAAI-98 Workshop on Sharing Information in Bioinformatics and Medical Knowledge Bases, 1998.

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

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

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

[36] MacGregor RM, “A Description Classifier for the Predicate Calculus” *Proc. of AAAI 94*, 1994.

[37] Mallery JC, “A Common LISP Hypermedia Server”, *Proceedings WWW 94*, 1994.

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

[39] National Library of Medicine, *UMLS Knowledge Sources*, 1998 edition, available from the NLM, Bethesda, Maryland.

[40] Neches R et al., “Enabling Technology for Knowledge Sharing”, *AI Magazine*; 3, 1991.

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

[42] Peirce CS, "On Signs and the Categories", in *I fondamenti della semiotica cognitiva*. Torino, Einaudi (1980).

[43] Pisanelli DM, Gangemi A, Steve G, "An Ontological Analysis of the UMLS Metathesaurus", *Journal of American Medical Informatics Association,* vol. 5 (symposium supplement), 1998.

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

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

[46] Pusteovsky J, *The Generative Lexicon,* Cambridge, MA, MIT Press, 1995.

[47] 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 Europe MIE94,* 1994.

[48] Rector A, Solomon WD, Nowlan WA, Rush T, "A Terminology Server for Medical Language and Medical Information Systems", *Methods of Information in Medicine,* 34, 1995.

[49] Schulze-Kremer S, "Adding Semantics to Genome Databases: Towards an Ontology for Molecular Biology", 5th Int. Conf. on Intelligent Systems for Molecular Biology, Halkidiki, Greece, AIII Press, Menlo Park, 1997.

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

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

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

[53] Spackman KA, Campbell KE, Coté RA, "SNOMED RT: A reference terminology for health care", *Proceedings of AMIA 97 Conference,* 1997.

[54] 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.

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

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

[57] Tuttle MS, Chute MD, Safran C, Abelson DJ, Campbell KE, Panel: Enterprise Experience with a Reusable Vocabulary Component, *Journal of American Medical Informatics Association,* vol. 5 (symposium supplement), 1998.

[58] Ungerer F, Schmid H-J, "An Introduction to Cognitive Linguistics", London, Longman, 1996.

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

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

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

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

[63] Zeng Q, Cimino JJ, "Automated Knowledge Extraction from the UMLS", *Journal of American Medical Informatics Association,* vol. 5 (symposium supplement), 1998.