Invited chapter. In: Chen H, Fuller S, Hersh WR, Friedman C, editors. Medical informatics:
Advances in knowledge management and data mining in biomedicine: Springer-Verlag; 2005 (in press).

Chapter #2.5

BIOMEDICAL ONTOLOGIES

Olivier Bodenreider[1] and Anita Burgun[2]
[1]U.S. National Library of Medicine, 8600 Rockville Pike, MS 43, Bethesda, Maryland 20894, USA; [2]EA 3888 Laboratoire d'Informatique Médicale, Université de Rennes I, Avenue du Pr Léon Bernard, 35043 Rennes Cedex, France

Abstract: Ontology design is an important aspect of medical informatics, and reusability is a key issue that is determined by the level of compatibility among ontology concepts and among the theories of the biomedical domain they convey. In this article, we examine OpenGALEN, the UMLS Semantic Network, SNOMED CT, the Foundational Model of Anatomy, and the MENELAS ontology as well as descriptions of the biomedical domain in two general ontologies, OpenCyc and WordNet. Using the representation of *Blood* in each system, we examine issues in compatibility among these ontologies. The presence of additional knowledge is also illustrated and some issues in creating and aligning biomedical ontologies are discussed.

Key words: Biomedical ontology, biomedical knowledge representation, GALEN, UMLS, SNOMED CT, Foundational Model of Anatomy.

1. INTRODUCTION

The purpose of biomedical ontology is to study classes of entities (i.e., substances, qualities and processes) in reality which are of biomedical significance. Examples of such classes include substances such as the mitral valve and glucose, qualities such as the diameter of the left ventricle and the catalytic function of enzymes, and processes such as blood circulation and secreting hormones. Unlike biomedical *terminology*, whose purpose is to collect the names of entities employed in the biomedical domain, biomedical *ontology* is concerned with the principled definition of biological classes and the relations among them. In practice, as they are more than lists of terms but

do not necessarily meet the requirements of formal organization, the many products developed by biomedical terminologists and ontologists often fall between terminologies and ontologies and constitute an “ontology gradient”.

Ontologies may be categorized according to the domain they represent or the level of detail they provide (Figure 1). *General ontologies* represent knowledge at an intermediate level of detail independently of a specific task. In such ontologies, upper levels reflect theories of time and space, for example, and provide notions to which all concepts in existing ontologies are necessarily related. *Domain ontologies* represent knowledge about a particular part of the world, such as medicine, and should reflect the underlying reality through a theory of the domain represented. Finally, ontologies designed for specific tasks are called *application ontologies.* Conversely, *reference ontologies* are developed independently of any particular purpose and serve as modules sharable across domains.

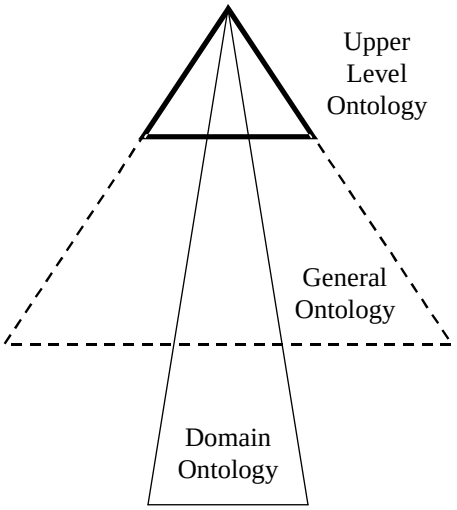

Figure #2.5-1. Kinds of ontologies

Core categories should be sharable across ontologies. Lower levels of upper level ontologies as well as general categories should be compatible with the equivalent semantic areas in the corresponding domain ontologies. For example, *Disease* in a general ontology should be compatible with that concept in a biomedical ontology. In addition, generic theories and meta-

level categories should be shared by every type in every ontology. For example, a representation of anatomy should re-use a generic theory of spatial objects. In turn, as anatomy is central to biomedicine and essentially stable, an ontology of anatomy can serve as a reference for ontologies relying on a representation of the human body, e.g., for an ontology of Diseases. In practice, however, these ideals are not always achieved. More generally, constructing biomedical ontologies that accommodate knowledge sharing by both humans and computer systems is challenging.

Ontologies play a fundamental role in medical informatics research (Musen 2002), contributing, for example, to natural language processing (e.g., Hahn et al. 1999), interoperability among systems (e.g., Degoulet et al. 1998), and access to heterogeneous sources of information, including the Semantic Web (e.g., Pisanelli et al. 2004). Increasingly, ontologies act as enabling resources in a variety of biomedical applications.

The objective of this paper is not to examine how applications benefit from using ontologies, but rather to present the characteristics of some major biomedical ontologies. In particular, we investigate how existing ontologies give differing views of the biomedical domain. First, we examine the representation of biomedicine in general systems such as OpenCyc and WordNet. We then describe three systems in the biomedical domain, GALEN, the UMLS, and SNOMED CT. A reference ontology, the Foundational Model of Anatomy, is also explored. Finally, as an example of an application ontology, we examine the MENELAS project. After a brief presentation of the characteristics of these ontologies, we look at the concept *Blood* in each system to illustrate common features and differences. Issues in building a single, sharable framework for representing biomedical knowledge are discussed.

This study was conducted at the U.S. National Library of Medicine as part of the Medical Ontology Research project (Bodenreider 2001), which focuses on developing methods for acquiring biomedical ontologies from existing resources and for validating them against other knowledge sources. References for the ontologies presented in this paper are listed in the appendix (Table 3) along with a summary of their main characteristics (Table 4). It is beyond the scope of this paper to present the techniques (e.g., description logics and frames) and tools (e.g., Protégé) used for representing ontologies. The interested reader is referred to references such as (Sowa 2000; Brachman and Levesque 2003).

2. REPRESENTATION OF THE BIOMEDICAL DOMAIN IN GENERAL ONTOLOGIES

2.1 OpenCyc

Cyc,® a general ontology developed by Cycorp, Inc., is built around a core of more than 1,000,000 hand-coded assertions (expressed in the formal language CycL) that capture "common sense" knowledge and enable a variety of knowledge-intensive applications. "Microtheories" are groups of assertions sharing a common set of assumptions focused according to a particular parameter, such as domain, level of detail, or time interval. OpenCyc,™ the upper level, publicly available part of the ontology, contains 6,000 concepts and 60,000 assertions about those concepts.

In OpenCyc as illustrated in Figure 2, *Thing*, the universal set, is the collection of everything. *Thing* is partitioned into *Set or collection* vs. *Individual* on the one hand and *Intangible* vs. *Partially tangible* on the other. Entities in OpenCyc are both represented as instances of sets, e.g., *Cancer* is an instance of the type *Disease Type* (`#$isa #$Cancer #$DiseaseType`) and organized in class/subclass hierarchies (`#$genls #$Cancer #$AilmentCondition`). Further specification may be provided by functions. *CancerFn*, for example, expresses that body parts can be the location of cancers. This function has domain animal body parts and range specific cancers: e.g., (`#$CancerFn #$Throat`).

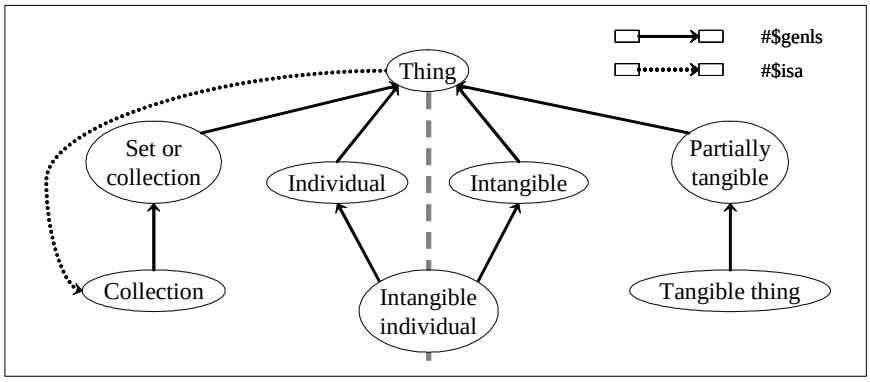

Figure #2.5-2. Top level in OpenCyc (partial representation)

Microtheories such as Biology or Ailment are relevant in the biomedical domain and have two primary benefits: (1) some assertions have microtheories as arguments: Everything true in Vertebrate Physiology is also true in Ailment and (2) some entities have distinct representations under

distinct microtheories: in Animal Physiology, subordinates of *Sensor* include *Nose*, *Skin*, and *Ear*, while in Naïve Physics they include *Tactile sensor* and *Electromagnetic radiation sensor*.

2.2 WordNet

WordNet® is an electronic lexical database developed at Princeton University (Fellbaum 1999) that serves as a resource for applications in natural language processing and information retrieval. The core structure in WordNet is a set of synonyms (synset) that represents one underlying concept. Synset formation is based on synonymy (one meaning expressed by several words) and polysemy (one word having several distinct meanings). There are separate structures for each linguistic category covered: English nouns, verbs, adjectives, and adverbs. For example, the adjective "renal" and the noun "kidney," although similar in meaning, belong to two distinct structures, and a specific relationship, "pertainymy," relates the two forms.

The current version of WordNet (2.0) contains over 114,000 noun synsets categorized into nine hierarchies, each starting with a "unique beginner" (see Figure 3). Each synset in the noun hierarchy belongs to at least one *is-a* tree (hyponymy) and may additionally belong to several *part-of*-like trees (meronymy). Hyponymy relations are established between synsets according to the following definition: A concept represented by the synset {x,x',…} is said to be a hyponym of the concept represented by the synset {y,y',…} if native speakers of English accept sentences constructed from frames such as "An x is a kind of y" (Fellbaum 1999). WordNet has been influenced by cognitive psychology as well as linguistics, and its hierarchies are not based on formal ontology theory. Gangemi et al. (2001) provide an ontological analysis of WordNet's top level and propose a revised, principled taxonomy.

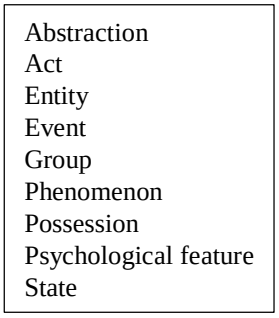

Figure #2.5-3. Top level in WordNet ("unique beginners")

Many concepts that represent health disorders in medical terminologies, when present in WordNet, are categorized appropriately; for example, *Leukemia* is a hyponym of *Cancer* (Burgun and Bodenreider 2001a; Burgun and Bodenreider 2001b). However, in some instances a medical sign or symptom appears only as a hyponym of a non-medical concept: the hypernym of *Vasoconstriction* (decrease in the diameter of blood vessels) is *Constriction*. This view emphasizes physical mechanism rather than pathology, and as a consequence, there is no formal relationship between *Vasoconstriction* and the biomedical domain in WordNet.

3. EXAMPLES OF MEDICAL ONTOLOGIES

3.1 GALEN

GALEN (Generalised Architecture for Languages, Encyclopaedias, and Nomenclatures in medicine) is a European Union project (1992-1999) that seeks to provide re-usable terminology resources for clinical systems. An ontology, the Common Reference Model, is formulated in a specialized description logic, the GALEN Representation and Integration Language (GRAIL), and is a core feature of GALEN (Rector et al. 1997). This ontology aims to represent "all and only sensible medical concepts," independently of any application. OpenGALEN provides a point of access to the GALEN Common Reference Model and to descriptions and specifications of the GALEN technology.

A key feature of GALEN is that it was constructed by defining the representation formalism and top level knowledge before populating the ontology. In addition, unlike traditional terminological resources whose terms are pre-coordinated, GALEN essentially provides the building blocks required for describing terminologies, as well as a mechanism for combining simple concepts. For example, the concepts *Adenocyte* and *Thyroid gland* are present in GALEN. However, instead of providing an explicit representation for *Adenocyte of thyroid gland*, GALEN indicates that it can be described by a combination of concepts: (*Adenocyte* which < *is structural component of Thyroid gland* >). The current version of OpenGALEN (December 2002) contains about 25,000 concepts. The GALEN ontology has been used for representing complex structures such as descriptions of medical procedures (Trombert-Paviot et al. 2000).

The major division in top level categories (Figure 4) is between *Phenomenon*, which subsumes structures, processes and substances, and *Modifier Concept*. The latter notion is used to distinguish concepts that represent things with independent existence (physical objects, for example)

from dependent concepts such as modifiers (*Mild severity*), states (*Pathological state*) or roles (*Infective role*). In addition to a hierarchy of categories, GALEN provides a rich hierarchy of associative relationships used to define complex structures. Its representation of partitive relations is particularly developed (Rogers and Rector 2000), including *has surface division* (`Hand has-surface-division Palm`), *has solid division* (`Heart has-solid-division Cardiac Septum`), *has layer* (`Heart has-layer Myocardium`), *has blind pouch division* (`Caecum has-blind-pouch-division Appendix Vermiformis`), *has linear division* (`Intestine has-linear-division Jejunum`), *has specific structural component* (`Knee Joint has-specific-structural-component Patella`), and *is specifically made of* (`Blood Clot is-specifically-made-of Coagulated Blood`).

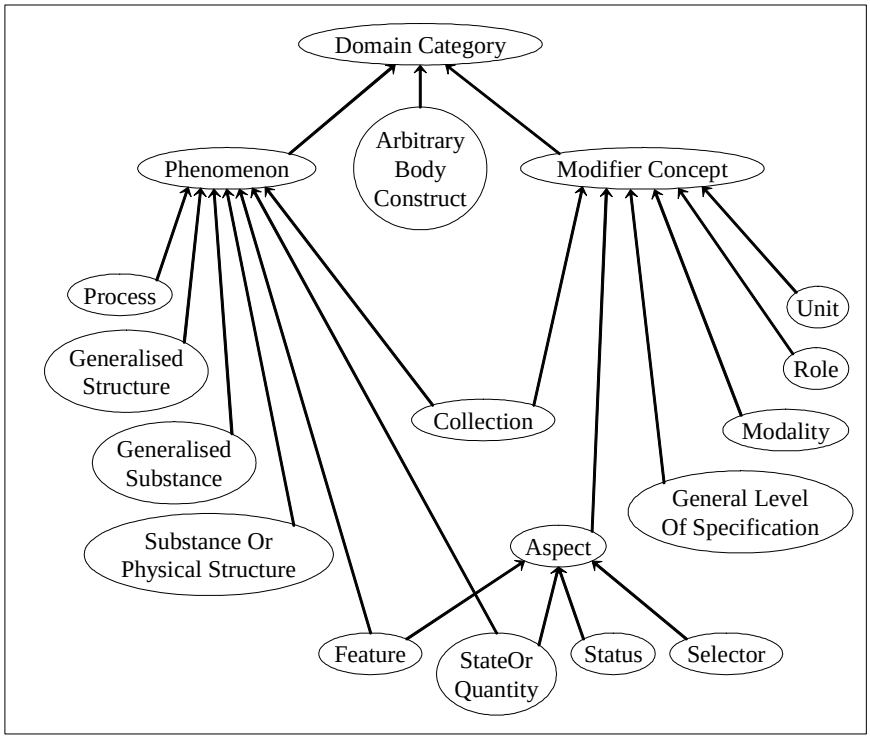

Figure #2.5-4. Top level in OpenGALEN

3.2 Unified Medical Language System

The Unified Medical Language System® (UMLS)® was developed by the National Library of Medicine to help health care professionals and

researchers access biomedical information from a variety of sources (Lindberg et al. 1993). The Metathesaurus,® a large repository of concepts, and the Semantic Network, a limited network of 135 semantic types, integrate over one million concepts from more than a hundred vocabularies and terminologies (2004AB version). While the structure of each source is preserved in building the Metathesaurus, equivalent terms are clustered into a semantically unique concept. Interconcept relationships are either inherited from underlying vocabularies or specifically generated. Since the Metathesaurus imposes no restrictions on sources, it cannot provide the kind of organization expected from an ontology. In contrast, the Semantic Network is developed independently of the vocabularies integrated in the Metathesaurus and serves as a basic, high-level ontology for the biomedical domain (McCray 2003). As illustrated in Figure 5, semantic types from the Semantic Network are used to categorize all UMLS concepts (McCray and Nelson 1995).

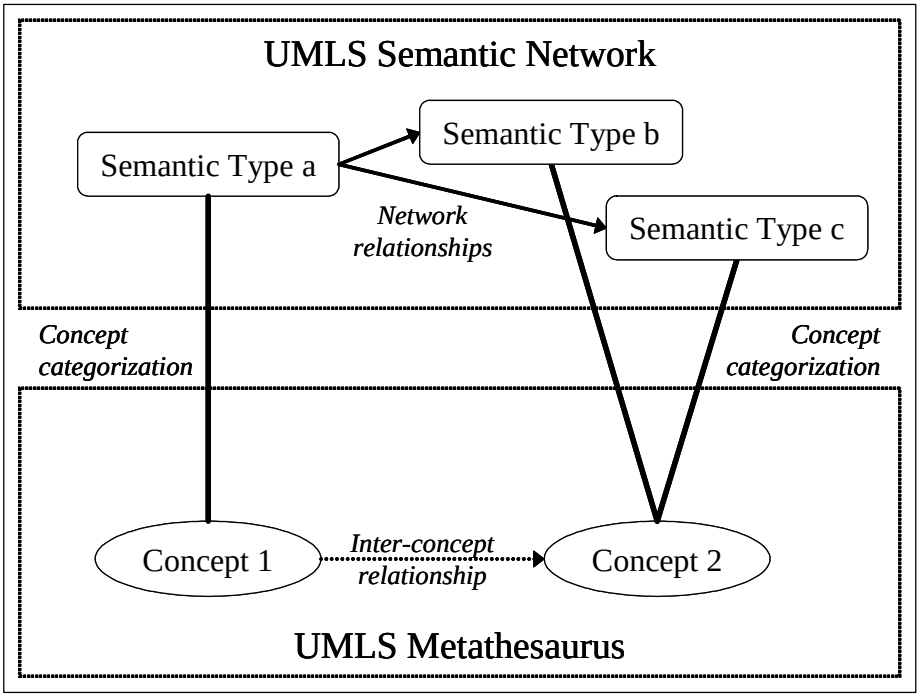

Figure #2.5-5. The two-level structure in the UMLS

At the highest level, the Semantic Network is organized around the opposition of entities and events, and two single-inheritance hierarchies reflect this distinction. The immediate children of *Entity* are *Physical Object*

and *Conceptual Entity*, while *Event* has *Activity* and *Phenomenon or Process* as direct descendants (Figure 6). Each semantic type in the network has a textual definition and appears in one of these hierarchies. In addition to the taxonomy, associative relationships in five subcategories are defined between semantic types: physical (e.g., *part_of*, *branch_of*, *ingredient_of*), spatial (e.g., *location_of*, *adjacent_to*), functional (e.g., *treats*, *complicates*, *causes*), temporal (e.g., *co-occurs_with*, *precedes*), and conceptual (e.g., *evaluation_of*, *diagnoses*). Since each Metathesaurus concept is assigned at least one semantic type, relationships between semantic types also define the allowable semantics for relationships between concepts (McCray and Bodenreider 2002).

The categorization of concepts by semantic type is subject to the economy principle (similar to the notion of parsimony developed in (Gruber 1995; Swartout et al. 1996)) and has three key features: (1) Since the most specific semantic type in the taxonomy is assigned to a concept, level of granularity varies across the UMLS (McCray and Hole 1990). (2) Due to single-inheritance tree structure rather than a lattice allowing multiple inheritance, a Metathesaurus concept cross-categorized by two semantic types is assigned to both types. (3) Rather than proliferating semantic types, concepts that cannot be categorized by existing sibling types are assigned their common supertype (McCray and Nelson 1995). The consequences of the economy principle for representing knowledge in the UMLS are discussed elsewhere (Burgun and Bodenreider 2001c).

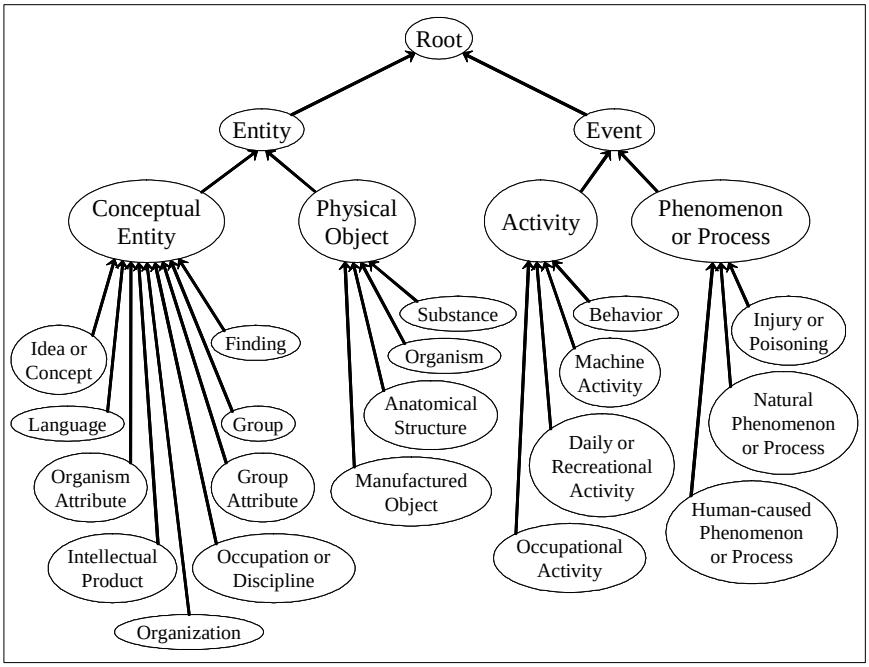

Figure #2.5-6. Top level in the UMLS Semantic Network

3.3 The Systematized Nomenclature of Medicine

The Systematized Nomenclature of Medicine (SNOMED®) Clinical Terms® (SNOMED CT), developed by the College of American Pathologists, was formed by the convergence of SNOMED RT and Clinical Terms Version 3 (formerly known as the Read Codes). SNOMED CT is the most comprehensive biomedical terminology recently developed in native description logic formalism. The version described here (January 31, 2004) contains 269,864 classes[1], named by 407,510 names[2]. SNOMED CT is now available as part of the UMLS[3] at no charge for UMLS licensees in the U.S. It is therefore likely to become widely used in medical information systems.

Each SNOMED CT concept is described by a variable number of elements. For example, the class *Viral meningitis* has a unique identifier (58170007), two parents (*Infective meningitis* and *Viral infections of the central nervous system*), several names (*Viral meningitis*, *Abacterial*

[1] SNOMED CT has a total of 357,135 classes of which 269,864 are "current"

[2] Among the 957,349 names in SNOMED CT, 407,510 correspond to the 269,864 "current" classes, excluding fully specified names and keeping only names whose status is "current"

[3] http://umlsinfo.nlm.nih.gov/

meningitis, and *Aseptic meningitis, viral*). The roles (or semantic relations) present in the definition of this concept are listed in Table 1.

Table #2.5-1. Roles present in the definition of *Viral meningitis*

Role	Value
Causative agent	*Virus*
Associated morphology	*Inflammation*
Finding site	*Meninges structure*
Onset	*Sudden onset;* *Gradual onset*
Severitiy	*Severities*
Episodicity	*Episodicities*
Course	*Courses*

SNOMED CT consists of eighteen independent hierarchies reflecting, in part, the organization of previous versions of SNOMED into "axes" such as *Diseases*, *Drugs*, *Living organisms*, *Procedures* and *Topography*. The first level concepts are listed in Table 2 with their frequency distribution.

Table #2.5-2. The eighteen top-level concepts in SNOMED CT and their frequency distribution

Top-level concepts	Frequency
Attribute	991
Body structure	30,652
Clinical finding	95,605
Context-dependent categories	3,649
Environments and geographical locations	1,620
Events	87
Observable entity	7,274
Organism	25,026
Pharmaceutical / biologic product	16,867
Physical force	199
Physical object	4,201
Procedure	46,066
Qualifier value	8,134
Social context	4,896
Special concept	178
Specimen	1,053
Staging and scales	1,098
Substance	22,267

3.4 Foundational Model of Anatomy

Development of the Foundational Model of Anatomy (FMA) at the University of Washington grew out of earlier work to enhance the anatomical content of the UMLS. By focusing exclusively on the representation of structure, the FMA expects to serve as a reference ontology, i.e., to allow other ontologies of which anatomy is a component to be aligned with it (Rosse and Mejino 2003). Specifically, the goal of the FMA is to provide a conceptualization of the material objects and spaces that constitute the human body. It integrates an Anatomical Ontology with two much smaller structures: the Physical State Ontology and the Spatial Ontology. The latter represents geometric objects and three-dimensional shape classes, and also distinguishes between bona fide (real) and fiat (virtual) boundaries of volumes, surfaces, and lines. The Anatomical Ontology contains nearly 70,000 concepts originally limited to gross anatomy and is now being extended to cellular and sub-cellular phenomena. FMA is implemented in Protégé[4], a frame-based ontology editing environment developed at Stanford University.

Definitions of physical anatomical entities in the Foundational Model of Anatomy are formulated by specifying constraints (Michael et al. 2001) based on spatial dimension, mass, and inherent three-dimensional shape, as well as the structural units that make up the body. Relationships, however, are constrained to the structural organization of physical anatomical entities. The top level of the taxonomy is *Anatomical entity*, which is divided into *Physical anatomical entity* and *Non-physical anatomical entity* (Figure 7). Physical entities have spatial dimension, while non-physical entities, such as *Developmental stage*, do not. Further distinction is made between physical entities that have mass, such as anatomical structures and body substances (*Material physical anatomical entity*), and those that do not, including anatomical spaces, surfaces, lines, and points (*Non-material physical anatomical entity*). The attribute of inherent three-dimensional shape contrasts anatomical structures, which are objects, with body substances.

[4] http://protege.stanford.edu/

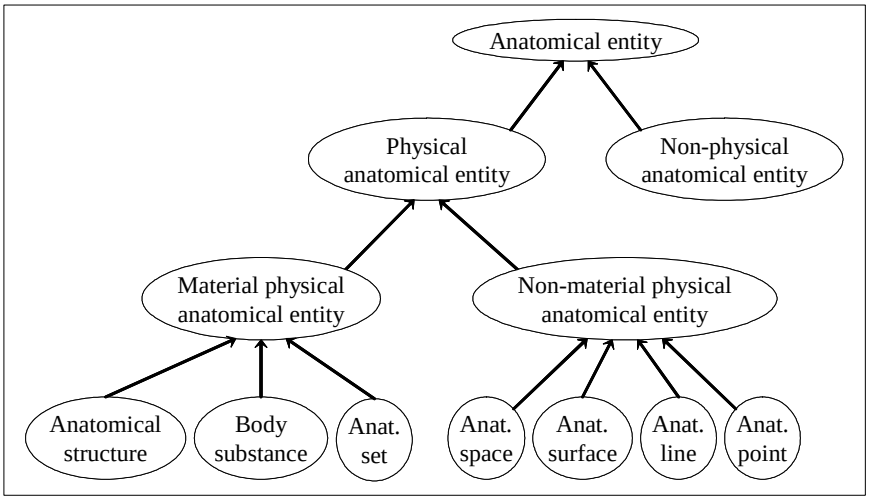

Figure #2.5-7. Top level in the Foundational Model of Anatomy (Anatomical taxonomy)

In addition to the anatomical taxonomy, hierarchies have been formulated using the transitive *part-of* relation as well as two anatomical relations, *branch-of* and *tributary-of*, which represent relationships among tree-like structures such as nerves, arteries, veins, and lymphatic vessels. Moreover, the FMA extends these relationships to boundary, orientation, connectivity, and location; the latter is specified using containment, adjacency, and anatomical coordinates (Mejino et al. 2001).

3.5 MENELAS ontology

MENELAS, a European Union project for accessing medical records in several European languages (Zweigenbaum 1994), takes a knowledge-based approach to natural language understanding. A pilot application covering coronary artery disease has been developed, and resources (represented as conceptual graphs) include domain-specific syntactic and semantic lexicons as well as an ontology of coronary artery diseases enhanced with structured encyclopedic knowledge for each concept.

The MENELAS ontology (see Figure 8 for the top level) has 1,800 concepts and 300 relationship types acquired from several sources, including interviews with physicians, reuse of existing terminological resources, and corpus analysis. It was initially developed as a lattice (Bouaud et al. 1994); however, to avoid ambiguities due to multiple inheritance, the principles of opposition of siblings and unique semantic axis were later adopted, leading to a tree structure (Zweigenbaum et al. 1995). Concept labels in the ontology are simply mnemonic; the actual meaning of a concept comes from its

position in the hierarchy. For example, *Physical object* is a child of *Abstract object*, which in turn is a child of *Substratum*. The latter concept is defined as having instances in the world and is opposed to *Ideal object*: *Apple* is an *Abstract object*, whereas *Two* is an *Ideal object*.

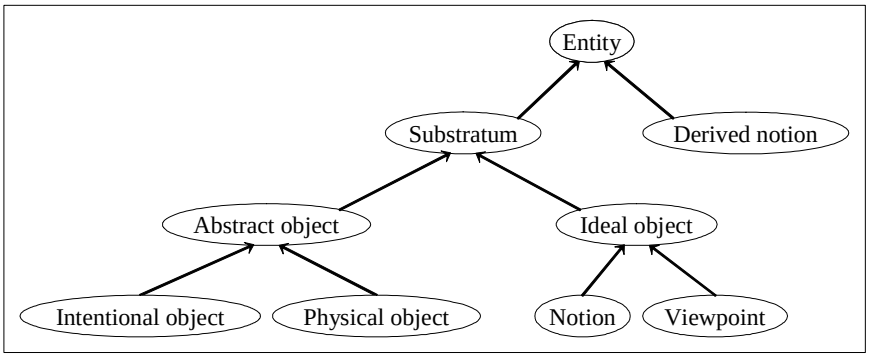

Figure #2.5-8. Top level in MENELAS

Relations are categorized according to the kinds of concepts they link. Relations between physical objects, for example, link mass objects and countable objects (*contains*, *has for dosage*, and *constituted of*) or real objects and pseudo-objects (*component of*). The *part of* relation links any kind of physical object and has children *part fragment* and *part segment*. There is also a relation, *functional part*, to represent functional viewpoints. Models and schemas provide additional knowledge, which may be limited to the domain-specific and task-oriented context of the MENELAS application. For example, the model for organ component includes the notion of duct in order to accommodate the coronary arteries.

4. REPRESENTATIONS OF THE CONCEPT *BLOOD*

Having discussed the general characteristics and top level organization of several ontologies, we now examine the representation of blood in these systems and analyze the differences among representations. We also show how most ontologies provide a rich representation compared to mere taxonomies by including additional knowledge.

4.1 *Blood* in biomedical ontologies

What makes the representation of *Blood* interesting is the dual nature of blood as both tissue and fluid, a dichotomy reflected in medical dictionary definitions: (1) "the fluid that circulates through the heart, arteries, capillaries, and veins, carrying nutriment and oxygen to the body cells" (Dorland's); and (2) "the 'circulating tissue' of the body; the fluid and its suspended formed elements that are circulating through the heart, arteries, capillaries, and veins" (Steadman's). In the following discussion, comparison of the ontologies is based on textual and formal definitions of *Blood* as well as ontological properties of that concept.

Although not represented as a type in **OpenCyc**, *Blood* is a specialization of *Mixture*, along with *Mud*, *Air*, and *Carbonated beverage*. (*Blood* referring to lineage is represented separately.) *Mixture* is a subclass of *Partially tangible* and represents a homogeneous, partially tangible thing composed of two or more different constituents which have been mixed. Because its constituents do not form chemical bonds, a mixture may be resolved by a separation event. As a mixture, *Blood* is an element of the collection *Existing stuff type* (`#$isa #$Mixture #$ExistingStuffType`), which implies that division in time or space does not destroy its stuff-like quality. In OpenCyc, *Blood* is represented differently from *Sweat* and *Semen*, which are subordinates of *Bodily secretion*. In addition, *Sweat*, considered as a waste, is also a descendant of *Excretion substance*.

Blood is defined in **WordNet** as "the fluid (red in vertebrates) that is pumped by the heart. Blood carries oxygen and nutrients to the tissues and carries waste products away; the ancients believed that blood was the seat of the emotions." There are five other meanings of "blood," including one referring to temperament or disposition. The direct hypernym of *Blood* is *Liquid body substance*. (The complete hierarchy for *Blood* in WordNet is given in Figure 9a.) *Blood*, *Sweat*, and *Semen*, are categorized as *Liquid body substance*. Unlike *Blood*, *Sweat* is linked to *Liquid body substance* through the synset *Secretion*.

In **OpenGALEN**, *Blood* is a subordinate of *Soft tissue* as well as *Lymphoid tissue*, *Integument*, and *Erectile tissue*, among others. The hierarchy for *Blood* in GALEN appears in Figure 9b. This structure is actually a lattice, since *Substance* is the common subtype of *Generalised substance* and *Substance or physical structure*, both being subtypes of *Phenomenon*. In GALEN, *Blood* is represented differently from *Sweat* and *Semen*, which are subordinates of *Body fluid*.

Blood has the semantic type *Tissue* in the **UMLS Metathesaurus**, which is defined as "An aggregation of similarly specialized cells and the associated intercellular substance. Tissues are relatively non-localized in

comparison to body parts, organs or organ components." *Tissue* is a subordinate of *Fully-formed anatomical structure* in the Semantic Network (Figure 9c has the entire *is-a* hierarchy for *Blood*). In the UMLS, *Blood* is not assigned the same semantic type as *Sweat* and *Semen*, which are categorized as *Body substance*. Moreover, in the Metathesaurus, ancestors of *Blood* include *Body fluid*, *Body substance*, *Soft tissue* and *Connective tissue*.

In **SNOMED CT**, *Blood* is found in the concept category *Substance* as a subordinate of *Blood material*, as well as *Blood component*. (The hierarchical environment for *Blood* in SNOMED CT is given in Figure 9d.) Multiple inheritance allows *Body fluid*, an ancestor of *Blood*, to inherit from both *Body substance* and *Liquid substance*. These two concepts are descendants of the top level category *Substance*. Subordinates of *Body fluid* also include *Sweat* and *Semen*, as well as *Lymph* and *Pus*.

The **Foundational Model of Anatomy** (FMA) represents *Blood* as a subordinate of *Body substance*, which is defined as "a material physical anatomical entity in a gaseous, liquid, semisolid or solid state, with or without the admixture of cells and biological macromolecules; produced by anatomical structures or derived from inhaled and ingested substances that have been modified by anatomical structures as they pass through the body." In addition to *Blood*, this definition covers other cellular fluids, such as *Semen*, as well as secretions (e.g., *Saliva* and *Sweat*), transudates (e.g., *Lymph*, and *Cerebrospinal fluid*), excretions (e.g., *Feces* and *Urine*), along with *Respiratory air* and *Aqueous humor of eyeball*. *Blood* is not considered to be a tissue in the FMA. The complete *is-a* hierarchy for *Blood* is represented in Figure 9e, and this lineage is distinct from that of *Tissue*, largely because substances, as defined in the FMA, do not have inherent three-dimensional shape. *Tissue* inherits properties from its ancestor *Anatomical structure*, which is a sister of *Body substance* and is differentiated from it by the feature inherent 3D shape.

In **MENELAS**, *Blood* (along with *Lymph*) is a subordinate of *Body fluid*. The ancestors of *Blood* can be found in Figure 9f. One of these, *Mass object*, has three subtypes: *Agglomerate* (divided into *Inorganic agglomerate* and *Organic agglomerate*), *Substance* (*Biochemical substance* and *Chemical substance*), and *Tissue* (*Body fluid* and *Connective tissue*). *Blood* as a child of *Body fluid* belongs to a different branch from the one dominated by *Substance*. Furthermore, *Tissue*, defined as a set of cells, is differentiated from *Substance*, defined as a set of molecules. A "model" (which provides additional knowledge) is associated with the concept *Body fluid* and emphasizes one property of fluids, namely viscosity, a feature pertinent to natural language understanding in the MENELAS application. The representation of *Body fluid* as tissue in MENELAS is noncanonical, given that other ontologies separate fluids and substances from tissue. *Semen* is

outside the scope of this application ontology for interpreting coronary angiography reports, while *Sweat* is categorized as *Cutaneous sign* (sweating), rather than *Substance*.

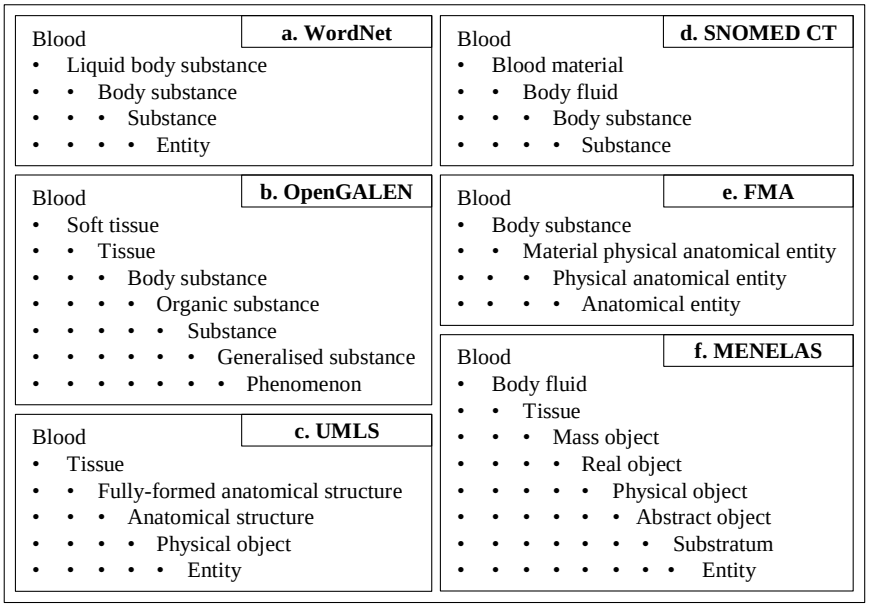

Figure #2.5-9. Representation of *Blood* in several biomedical ontologies

4.2 Differing representations

The differing representations of *Blood* in several systems raise issues about compatibility among ontologies. Obviously, the representation of most concepts is simpler than that of *Blood*, and the ontologies studied often provide compatible views on the biomedical domain. What makes the representation of *Blood* more complex is that two different superordinates are found: *Tissue* and *Body substance*. GALEN and the UMLS Semantic Network categorize *Blood* as *Tissue* while the Foundational Model of Anatomy categorizes it as *Body substance*. In between, WordNet, SNOMED CT and MENELAS categorize *Blood* as *Body fluid*, itself categorized as *Body substance* in WordNet and SNOMED CT, but as *Tissue* in MENELAS. Finally, in GALEN, *Tissue* is a subtype of *Body substance*. A composite representation of *Blood* is shown in Figure 10.

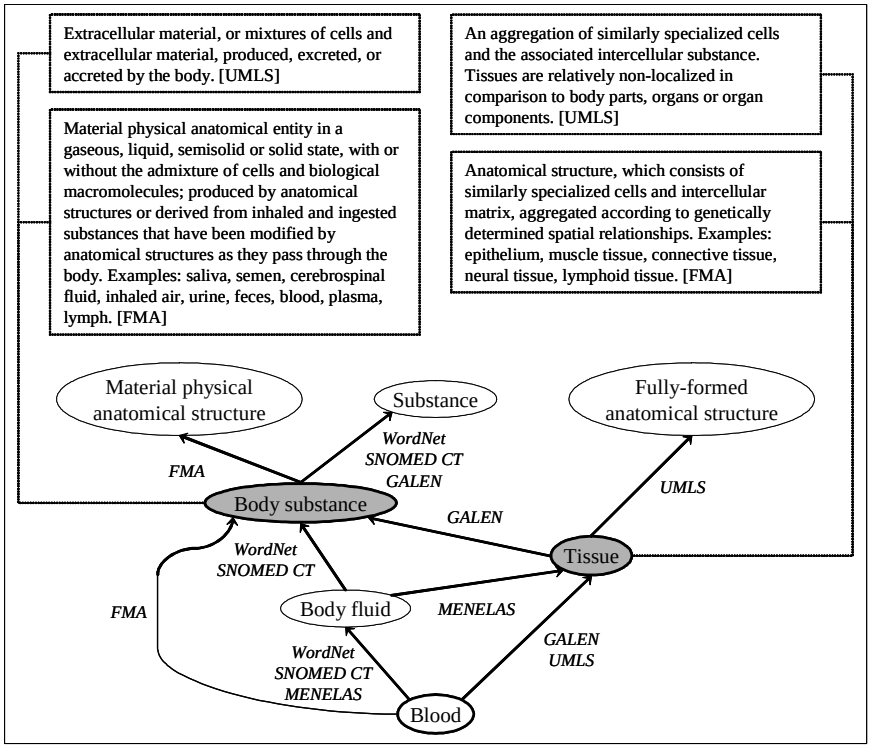

Figure #2.5-10. Composite representation of *Blood*

Superficially, this dual representation of *Blood*, as both *Tissue* and *Body substance*, does not reveal any major incompatibility, such as circular hierarchical relationships. However, a unified representation in which *Blood* is a common subtype of *Tissue* and *Body substance* would violate the constraint of opposition of siblings. Analyzed more carefully, the definitions of *Tissue* in the Foundational Model of Anatomy (FMA) and the Semantic Network are closely related but not equivalent (the complete definitions are shown in Figure 10). In both systems, *Tissue* is a kind of anatomical structure consisting of "similarly specialized cells and intercellular substance/matrix". The difference between the two systems lies in the precision – found only in the FMA – that this aggregation must follow "genetically determined spatial relationships". Blood cells in suspension in plasma or aggregated after sedimentation are indeed similarly specialized and correspond to the definition of *Tissue* in the UMLS Semantic Network. However, their spatial organization differs from that of an epithelium, muscle tissue and neural tissue in that it is not genetically determined but rather depend on the characteristics of blood circulation. This additional criterion is particularly relevant for disambiguating the classification of

Blood in the FMA. Moreover, the categorization of *Blood* as *Body substance* rather than *Tissue* in the FMA is consistent with the distinction introduced between *Anatomical structure* (of which *Tissue* is a subtype) and *Body substance* through the property has inherent 3D shape, which is present in *Tissue* and absent in *Body substance*.

The representation of *Blood* illustrates other differences across ontologies. While most ontologies represent the prototypical form of blood (i.e., the fluid circulating in the cardiovascular system), GALEN distinguishes between liquid and coagulated blood. The issue here is that the properties inherent to fluids are inherited by *Blood* in WordNet, SNOMED CT, FMA and MENELAS. As a consequence, if GALEN were integrated with these representations as shown in Figure 10, *Coagulated blood*, a descendant of *Blood*, would wrongly inherit such properties. Analogously, *Body substance* is likely to represent different entities in FMA and in GALEN. As mentioned earlier, *Body substance* in FMA is a *Material physical anatomical entity* with no inherent three-dimensional shape. In GALEN, *Body substance* is more general, encompassing both *Tissue* and *Body fluid* and defined as an *Organic substance* playing a role in physiology.

4.3 Additional knowledge

Taxonomy, i.e., the arrangement of concepts in *is-a* hierarchies, plays a central role in ontologies, of which such hierarchies constitute the backbone. In addition to the relative position of *Blood* in their hierarchies, most ontologies provide additional knowledge about *Blood* through properties attached to this concept and through the associative relations of *Blood* to other concepts. OpenCyc categorizes *Blood* as a *Mixture*, indicating that it can be subject to events such as *Separation mixture*. Erythrocyte sedimentation, resulting from the reversible separation of blood components, is an example of such events. In SNOMED CT, *Blood* is involved in the definition of other concepts through specific roles which provide additional knowledge about it. *Blood* can be analyzed (e.g., `Blood specimen` *`has specimen substance`* `Blood`), can be the object of medical procedures (e.g., `Transfusion of whole blood` *`has direct substance`* `Blood` and `Finger-prick sampling` *`has direct substance`* `Blood`) and can enter in the composition of clinical drugs (e.g., `Antithrombin III preparation` *`has active ingredient`* `Blood`). As a *Body fluid* in MENELAS, *Blood* acquires the viscosity property. *Blood* is also a subtype of *Mass object* and inherits the general knowledge represented for this type through relations (e.g., *Mass object* may be a component of *Countable object*) and properties (e.g., quantity, expressed with quantitative values and units). GALEN identifies two distinct

physical states for *Blood*: *Liquid blood* and *Coagulated blood*, (both represented as descendants of *Blood*). In addition, *Blood* inherits the properties of *Body substance* (e.g., `Body substance` *`plays physiological role`* `Organic role`). Additionally, GALEN extends the representation of *Blood* through roles such as `Blood` *`has countability`* `Infinitely divisible`. This role, inherited from *Substance*, expresses that *Blood* is not a discrete object. By categorizing *Blood* as *Tissue* in the UMLS, potential relationships with other kinds of entities can be inferred from the Semantic Network. Relationships of *Tissue* to other Semantic Types, result in predicates including `Tissue` *`produces`* `Biologically active substance`, `Tissue` *`is a location of`* `Pathologic function`, `Embryonic structure` *`is a developmental form of`* `Tissue`, and `Tissue` *`surrounds`* `Tissue`. In the Foundational Model of Anatomy, *Blood* inherits from *Body substance* the value *False* for the property has inherent 3D shape. The anatomical structures containing *Blood* including *Cavity of cardiac chamber* and *Lumen of cardiovascular system* are represented through relations such as `Blood` *`contained in`* `Cavity of cardiac chamber`.

5. ISSUES IN ALIGNING AND CREATING BIOMEDICAL ONTOLOGIES

As more biomedical ontologies are created, users might be tempted to integrate these sets of concepts and relations into a single system. However, the analysis of the differences in representation of *Blood* illustrated the limitations of a naïve approach to merging ontologies, even when representations occur within a single theory of the domain (i.e., Western medicine). While difference in granularity is usually not a problem, differing naming conventions, the lack of reliable textual definitions and the lack of explicit and consistently applied classificatory principles may result in merging difficulties. Additional difficulty is encountered when attempting to merge ontologies that convey different theories of the domain (e.g., Western and Oriental medicine or modern medical knowledge and pre-scientific representations of the human body). In this case, the target system must be able to clearly identify the underlying theories and to represent them separately. Tools have been developed to assist the ontology developer in merging existing ontologies (Noy and Musen 1999).

Ontology design can benefit from two complementary approaches. First, some methodologies such as the Protégé software engineering methodology, aim at providing a clear division between domain ontologies and domain-independent problem-solvers that, when mapped to domain ontologies, can

solve application tasks (Musen 1998). Second, ontologies can be improved by drawing on the results of recent research in philosophy called formal ontology. For example, Guarino et al. (2000) have developed methods built around the fundamental philosophical theories of identity, unity, rigidity and dependence, that can be used to reduce inconsistencies in *is-a* hierarchies. Mereotopology, the theory of parts and boundaries, addresses issues in *part-of* hierarchies. Exploiting these theories helps design principled ontologies. Applied to the biomedical domain, formal ontology addresses, for example, distinctions between a person and its body, or between being a person and being a patient. More generally, formal ontology helps create consistent upper-level ontologies to which domain ontologies can be hooked. For example, the principles of mereotopology have been applied to the representation of anatomical structures and subdivisions of the human body.

6. CONCLUSION

Although general ontologies and limited application ontologies may be useful, biomedical applications (e.g., clinical decision support systems, medical language processing and information retrieval) would benefit from large, principled domain ontologies. We examined some of the biomedical ontologies currently available and found that none of them fully meets the requirements of formal organization. Not surprisingly, we observed a certain lack of compatibility among their representations. Several factors contribute to this situation. First, there is no agreement on an upper level ontology to which a biomedical ontology could hook its concepts. Second, there is no unique theory of the domain, and some characteristics of biomedicine make it particularly difficult to represent (e.g., large number of concepts and vagueness of some concepts). Finally, pragmatic aspects rather than formal principles often prevail in the design of biomedical ontologies. The contribution of formal ontology has been acknowledged and will undoubtedly benefit medical ontology. Meanwhile, we believe that identifying and clarifying the core concepts and relationships of the domain will contribute to improve the sharability of existing ontologies as well as the interoperability of the applications that rely on them.

7. ACKNOWLEDGMENTS

The authors would like to thank Jeremy Rogers, Cornelius Rosse, Jacques Bouaud and Pierre Zweigenbaum for providing valuable insights about the ontologies to which they contributed. Special thanks to Tom

Rindflesch for his encouragement, useful comments and invaluable editorial assistance on this manuscript.

8. REFERENCES

Bodenreider, O. 2001. *Medical Ontology Research (Report to the Board of Scientific Counselors)*, Lister Hill National Center for Biomedical Communications, Bethesda, Maryland.

Bouaud, J., Bachimont, B., Charlet, J., and Zweigenbaum, P. 1994. Acquisition and structuring of an ontology within conceptual graphs, in: *ICCS'94 Workshop on Knowledge Acquisition using Conceptual Graph Theory*, University of Maryland, College Park, MD, pp. 1-25.

Brachman, R.J. and Levesque, H.J. 2003. *Knowledge representation and reasoning*, Morgan Kaufmann, Amsterdam; Boston.

Burgun, A. and Bodenreider, O. 2001a. Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System, *Proc NAACL Workshop, "WordNet and Other Lexical Resources: Applications, Extensions and Customizations"*:77-82.

Burgun, A. and Bodenreider, O. 2001b. Mapping the UMLS Semantic Network into general ontologies, *Proc AMIA Symp*:81-85.

Burgun, A. and Bodenreider, O. 2001c. Aspects of the taxonomic relation in the biomedical domain, in: *Collected papers from the Second International Conference "Formal Ontology in Information Systems"* (ed. C. Welty and B. Smith), ACM Press, pp. 222-233.

Degoulet, P., Sauquet, D., Jaulent, M.C., Zapletal, E., and Lavril, M. 1998. Rationale and design considerations for a semantic mediator in health information systems, *Methods Inf Med* **37**(4-5):518-526.

Fellbaum, C., ed. 1999. *WordNet: An electronic lexical database*, MIT Press, Cambridge, Massachusets.

Gangemi, A. and Oltramari, A. 2001. A formal ontology approach to refine lexical taxonomies: the case of WordNet top level, in: *Collected papers from the Second International Conference "Formal Ontology in Information Systems"* (ed. C. Welty and B. Smith), ACM Press, pp. 285-296.

Gruber, T.R. 1995. Toward principles for the design of ontologies used for knowledge sharing, *International Journal of Human-Computer Studies* **43**(5-6):907-928.

Guarino, N. and Welty, C. 2000. A Formal Ontology of Properties, in: *EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management*, (ed. R. Dieng and O. Corby), Springer-Verlag, pp. 97-112.

Hahn, U., Romacker, M., and Schulz, S. 1999. How knowledge drives understanding--matching medical ontologies with the needs of medical language processing, *Artif Intell Med* **15**(1):25-51.

Lindberg, D.A., Humphreys, B.L., and McCray, A.T. 1993. The Unified Medical Language System, *Methods Inf Med* **32**(4):281-291.

McCray, A.T. 2003. An upper-level ontology for the biomedical domain, *Comparative And Functional Genomics* **4**(1):80-84.

McCray, A.T. and Bodenreider, O. 2002. A conceptual framework for the biomedical domain, in: *The semantics of relationships: an interdisciplinary perspective* (ed. R. Green, C.A. Bean, and S.H. Myaeng), Kluwer Academic Publishers, Boston, pp. 181-198.

McCray, A.T. and Hole, W.T. 1990. The scope and structure of the first version of the UMLS Semantic Network, *Proc Annu Symp Comput Appl Med Care*:126-130.

McCray, A.T. and Nelson, S.J. 1995. The representation of meaning in the UMLS, *Methods Inf Med* **34**(1-2):193-201.

Mejino, J.L., Jr., Noy, N.F., Musen, M.A., Brinkley, J.F., and Rosse, C. 2001. Representation of structural relationships in the Foundational Model of Anatomy, *Proc AMIA Symp*:973.

Michael, J., Mejino, J.L., Jr., and Rosse, C. 2001. The role of definitions in biomedical concept representation, *Proc AMIA Symp*:463-468.

Musen, M.A. 1998. Domain ontologies in software engineering: use of Protege with the EON architecture, *Methods Inf Med* **37**(4-5):540-550.

Musen, M.A. 2002. Medical informatics: searching for underlying components, *Methods Inf Med* **41**(1):12-19.

Noy, N.F. and Musen, M.A. 1999. An algorithm for merging and aligning ontologies: automation and tool support, in: *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99) Workshop on Ontology Management*, AAAI Press, Orlando, Florida.

Pisanelli, D.M., Gangemi, A., Battaglia, M., and Catenacci, C. 2004. Coping with medical polysemy in the semantic web: the role of ontologies, *Medinfo* **2004**:416-419.

Rector, A.L., Bechhofer, S., Goble, C.A., Horrocks, I., Nowlan, W.A., and Solomon, W.D. 1997. The GRAIL concept modelling language for medical terminology, *Artif Intell Med* **9**(2):139-171.

Rogers, J. and Rector, A. 2000. GALEN's model of parts and wholes: experience and comparisons, *Proc AMIA Symp*:714-718.

Rosse, C. and Mejino, J.L., Jr. 2003. A reference ontology for biomedical informatics: the Foundational Model of Anatomy, *J Biomed Inform* **36**(6):478-500.

Sowa, J.F. 2000. *Knowledge representation: logical, philosophical, and computational foundations*, Brooks/Cole, Pacific Grove, Ca.

Swartout, B., Patil, R., Knight, K., and Russ, T. 1996. Toward Distributed Use of Large-Scale Ontologies, in: *Proceedings of the 10th Workshop on Knowledge Acquisition, Modeling and Management*, (ed. B. Gaines and M.A. Musen), Banff, Canada.

Trombert-Paviot, B., Rodrigues, J.M., Rogers, J.E., Baud, R., van der Haring, E., Rassinoux, A.M., Abrial, V., Clavel, L., and Idir, H. 2000. GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures, *Int J Med Inf* **58-59**:71-85.

Zweigenbaum, P. 1994. Menelas - an Access System for Medical Records Using Natural-Language, *Computer Methods and Programs in Biomedicine* **45**(1-2):117-120.

Zweigenbaum, P., Bachimont, B., Bouaud, J., Charlet, J., and Boisvieux, J.F. 1995. Issues in the Structuring and Acquisition of an Ontology for Medical Language Understanding, *Methods of Information in Medicine* **34**(1-2):15-24.

9. APPENDIX

Table #2.5-3. References for the ontologies mentioned in this chapter

Ontology	**URL**
Foundational Model of Anatomy	http://fma.biostr.washington.edu/
MENELAS	http://www.biomath.jussieu.fr/~pz/Menelas/
OpenCyc™	http://www.opencyc.com/
OpenGALEN	http://www.opengalen.org/
SNOMED CT®	http://www.snomed.org/
Unified Medical Language System®	http://umlsks.nlm.nih.gov/ (free UMLS registration required)
WordNet®	http://www.cogsci.princeton.edu/~wn/

Table #2.5-4. Some characteristics of the ontologies mentioned in this chapter

Name	**Version**	**Date**	**Scope**	**Objective**	**Formalism**	**Number of concepts**	**Number of relation-ship types**	**Number of assertions (explicitly represented)**
OpenCyc™	0.7	Dec. 2002	General	To support commonsense reasoning	CycL	6,000	n/a	60,000
WordNet®	2.0	Aug. 2003	General	Lexical reference	Graph of synsets	152,000	7	344,000
OpenGALEN	6	Dec. 2002	Clinical medicine	To support terminology services	Description logic (GRAIL)	25,000	594	216,000
UMLS® Semantic Network	2004 (AC)	Nov. 2004	Bio-medicine	To provide a consistent categorization of all concepts represented in the UMLS Metathesaurus	Semantic network	135	54	6864
SNOMED CT®		Jan. 31, 2004	Clinical medicine	Capturing, sharing and aggregating health data	Description logic	270,000	50	1.5 M
Foundational Model of Anatomy		Dec. 2003	Anatomy	Reference ontology	Frame-based	70,000	170	1.5 M
MENELAS	Final	March 1995	Coronary artery diseases	To support natural language processing	Conceptual graphs	1,800	300	n/a