KR-MED 2004 Proceedings

Udo Hahn, Chair

Sponsored by the AMIA Formal Biomedical Knowledge Representation Special Interest Group

Published by the
American Medical Informatics Association
4195 St. Elmo Avenue, #401
Bethesda, MD 20814
Phone: 301.657.1291
Fax: 301.657.1296
www.amia.org

Foreword

These are the proceedings of the First InternationalWorkshop on *Formal Biomedical Knowledge Representation* (KR-MED 2004), held in Whistler (British Columbia, Canada) on the 1st of June 2004. It is the first of this kind, organized by the recently founded Special Interest Group *Formal (Bio-)Medical Knowledge Representation* of the *American Medical Informatics Association* (AMIA). This workshop is collocated with KR 2004, the Ninth International Conference on the *Principles of Knowledge Representation and Reasoning*.

The engineering of large-scale domain knowledge, mostly in form of controlled vocabularies, taxonomies, and classification systems constitutes an important branch of activities in the field of Medical Informatics. The recent growth of interest in genomics and molecular biology has set another focus on the organization of the fast growing terminological knowledge in this domain. Despite recent advances in using formal languages for biomedical concept representation, many fundamental issues (ontological basis, expressivity, scalability) remain unresolved. Hence, it seemed to us as a natural move to discuss the challenges and requirements we have to offer directly with the KR community.

As the chairman of the programme committee I had the pleasure to collaborate with the following members of the PC:

- Olivier Bodenreider, NLM, USA
- James Cimino, Columbia University, USA
- Peter Elkin, Mayo Clinic, USA
- John Gennari, University of Washington, USA
- Ian Horrocks, University of Manchester, UK
- Mark Musen, Stanford University, USA
- Domenico Pisanelli, CNR, Italy
- Alan Rector, University of Manchester, UK
- Cornelius Rosse, University of Washington, USA
- Barry Smith, Leipzig University, Germany
- Chris Welty, IBM Research, USA

They have done a great job in reviewing the submitted papers, five on the average, in due time. Thank you all! For this workshop, we had 28 submissions out of which we selected 12 papers for presentation at the workshop. While 43% acceptance rate may be rather low for a a kick-off workshop, this may also guarantee a high level of quality of the papers that made it. Hence, we plan to publish a selection of the best papers to appear in a special issue of *Artificial Intelligence in Medicine*. So, stay tuned.

I also want to extend my thanks to the members of the Organizing Committee, *viz.* Stefan Schulz, Freiburg University Hospital, Germany, and Ronald Cornet, Amsterdam Academic Medical Center, The Netherlands. Their work was mainly behind the scene, but so important for the success of the whole enterprise.

Let us enjoy our workshop!
Udo Hahn, Freiburg University, Germany,
Chair of the KR-MED 2004 Programme Committee

Table of Contents

Axioms for parthood and containment relations in bio-ontologies

Thomas Bittner
Institute for Formal Ontology and Medical Information Science
University of Leipzig
thomas.bittner@ifomis.uni-leipzig.de

Abstract

To fix the semantics of different kinds of parthood relations we require axioms which go beyond those characterizing partial orderings. I formulate such axioms and show their implications for bio-ontologies. Specifically, I discuss parthood relations among masses, for example among body substances such as blood and portions thereof, and among components of complexes, for example between your stomach and your gastro-intestinal system. I contrast these with the relation of being contained in (as your lungs are contained in your thorax).

The axioms considered are rooted in mereology, the formal theory of parts and wholes. By making explicit the differences between the different kinds of relations they support different kinds of data integration in bioinformatics.

Introduction

The growth of bioinformatics has led to an increasing number of evolving ontologies which must be correlated with the existing terminology systems developed for clinical medicine. A critical requirement for such correlations is the alignment of the fundamental ontological relations used in such systems, and especially of the relation of part-of [16, 26].

However, there is one problem that stands in the way of achieving such integration: existing terminology systems and ontologies are marked by an inadequate degree of semantic consistency at their foundations [27]. The ambiguities and inconsistencies which result from the lack of a standard unified framework for understanding the basic ontological relationships that structure these domains are an obstacle to ontology alignment and data integration, and thus also to the sort of automatic processing of biomedical data which is the presupposition of advances in this field.

Part-whole relations play a critical role in medical concept representation. As Rogers and Rector [20] point out, this is most obvious in the modeling of anatomy; but it also true of the representation of surgical procedures, as well as of many physiological and disease processes, as also of the chemical pathways which lie beneath all of these.

Part-whole relations have long been the subject of extensive study in philosophy [2, 24], linguistics [31], knowledge representation [10, 9], and more recently in bio-informatics [11, 22, 20, 17]. In particular, it has long been recognized that several different subtypes of the part-of relation may be identified [19, 31, 9, 13]. This recognition underlies the modeling of the part-of relation in GALEN [20] and in the Foundational Model of Anatomy (FMA) [21, 16]. All such relations are, when taken singly, treated formally as partial orderings. However there does not exist a formal treatment of what *distinguishes* such relations one from another.

In this paper I give axiomatic theories for three sorts of partial ordering relations: (i) the component-of relation between components and the complexes they form (my mouth, my oropharynx, and my gastro-intestinal system are components of my alimentary system); (ii) part-of relations among masses such as body-substances in the sense of FMA (the blood in your left ventricle is part of the blood in your body); and (iii) containment relations (my brain is contained in my skull, my lungs are contained in my thorax).

The formal characterization will be purely mereological and will exploit the classification of formal theories given for example by Simons [23] or Varzi [29]. Thus no resources from topology or geometry are required. Moreover, in all that follows I consider entities at a single moment in time. The full formal characterization of all the part-whole relations contained in a system like the FMA or GALEN will need to go further than what is presented here. Distinctions of the type here discussed will however be indispensable to further progress in this field.

Partial ordering structures

In this paper formal theories of different kinds of partial order relations are discussed. Each of the theories is presented in a single-sorted first-order predicate logic with identity. I use the letters x, y, and z for variables. Predicates always begin with a capital letter. The logical connectors $\neg, =, \wedge, \vee, \rightarrow, \leftrightarrow$ have their usual meanings: not, identical-to, and, or, if ...then, if and only if (iff). I write (x) to symbolize universal quantification and $(\exists x)$ to symbolize existential quantification. Leading universal quantifiers are assumed to be understood and are omitted.

Properties of partial orderings

I introduce the binary primitive $x < y$ interpreted as the generic relation of proper partial ordering, i.e., x stands to y in the relation of proper partial ordering.
In terms of $<$, I define the relations of (improper) partial order and overlap: x and y are in the relation of improper partial order iff either $x < y$ or x and y are identical ($D_{\leq}$); x and y overlap iff they share a common entity in the partial ordering hierarchy (D_O):

$$\begin{array}{ll} D_{\leq} & x \leq y \equiv x < y \vee x = y \\ D_O & O\,xy \equiv (\exists z)(z \leq x \wedge z \leq y) \end{array}$$

I now add axioms to the effect that the relation of proper partial ordering, $<$, is asymmetric and transitive (APO1-APO2).

$$\begin{array}{ll} APO1 & x < y \rightarrow \neg y < x \\ APO2 & (x < y \wedge y < z) \rightarrow x < z \end{array}$$

It then follows that proper partial ordering is irreflexive (TPO1) and that (improper) partial ordering $\leq$ is reflexive, antisymmetric, and transitive (TPO2-4)[1]:

$$\begin{array}{ll} TPO1 & \neg x < x \\ TPO2 & x \leq x \\ TPO3 & (x \leq y \wedge y \leq x) \rightarrow x = y \\ TPO4 & (x \leq y \wedge y \leq z) \rightarrow x \leq z \end{array}$$

Examples of partial ordering structures

I now discuss three examples of partial order relations: the component-of relation, the containment relation, and the part-of relation as it holds between masses.

The complement-of relation. Consider the component-of relation between components and complexes of my alimentary system. Figure 1 shows the component-of structure of my alimentary system according to the FMA [21]. My mouth, my oropharynx, and my gastrointestinal system are components of my alimentary system. In general, the nodes c and d in the graph structure are connected by an arrow iff entity c is a component of the complex d.

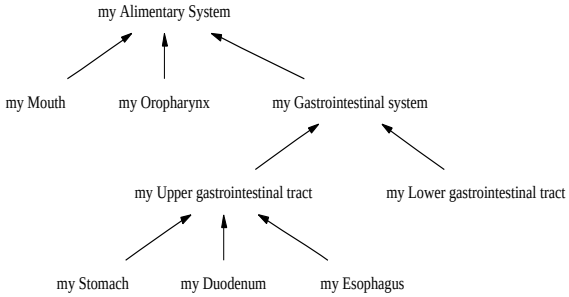


Figure 1: Component-of relations between the components of my Alimentary system

To see that the component-of relation satisfies the axioms of proper partial orderings (APO1-2) consider that components are distinct from the complexes they form. Since my stomach is a component of my alimentary system, the alimentary system is not a component of my stomach. Also the alimentary system is identical to itself but not a component of itself. Moreover, the component-of relation is transitive. My stomach is a component of my upper gastro-intestinal tract. My upper gastro-intestinal tract is a component of my gastro-intestinal system. And also my stomach is a component of my gastro-intestinal system.
As an example for overlap of complexes consider the alimentary system and the respiratory system according to the FMA. Both have the oropharynx as a component and hence overlap in the sense of definition D_O.

Containment is the second example of a proper partial ordering relation. For a non-medical example consider the relation between your backpack and the books therein, or the relation between your wallet and the coins therein, or the relation between the coins and the backpack in the case where the wallet with the coins is in the backpack.
For a medical example of containment consider the relation which holds between my pericardial sac and my thorax in the sense that my thorax forms a container for my pericardial sac, which in turn is contained in my thorax (Figure 2). The same relation of containment holds between my heart and my pericardial sac in the sense that my pericardial sac is a container for my heart. Clearly, containment understood in this sense is asymmetric and transitive. For example. The pericardial sac is a container for my heart, but the latter is not a container for the former. Since my heart is con-

[1] The formal proofs are omitted here but can be obtained from the author.

tained in my pericardial sac and my pericardial sac is contained in my thorax, and it also holds that my heart is contained in my thorax.

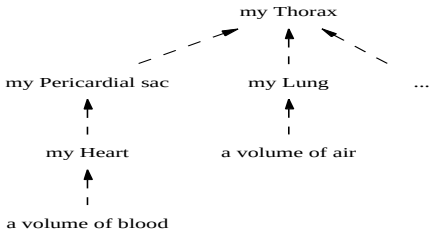


Figure 2: Containment relations

Notice that the interpretation of the containment relation employed here is different from those in the FMA [7] and GALEN [8]. Both interpret containment as a relation between an entity and (a part of) a space that is enclosed by a container. For example for GALEN [8] the heart is contained in the mediastinum, which is a part of the thoracic space.

Here, in contrast, the relation of containment always holds between entities – the contained entity (e.g., a volume of blood) and the container (e.g., my heart). Containers can themselves be contained in other containers (e.g., my heart is contained in my pericardial sac, which in turn is contained in my thorax).

Containers have properties, like having-a-cavity, which distinguish them from non-containers. The characterization of those properties, however, is beyond the realm of mereology. This requires at least the resources of topology and a theory of location [3, 6].

The advantage of the interpretation applied here is fourfold. Firstly, we focus on what *containment* means and not on what a container is. The former question can be answered within a mereological framework the latter cannot. Secondly we need only a single category in order to characterize containment – entities. In the interpretation of containment applied in the FMA and in GALEN one needs (at least) two categories: contained entities like the heart; and regions, like the thoracic space, which are enclosed by container-like entities. Thirdly, representing containment as relation of partial order between entities allows us to characterize the similarities and differences between parthood and containment in a very explicit manner.

Fourthly, representing containment as relation between entities allows us to distinguish it from the relation of *location*, which holds between entities and regions [4]. Often both relations are used in combination, for example, in order to say that the heart is *contained* in the thorax and within the thorax it is *located* in a region to which we refer to as the middle mediastinum, and which is a *part of* the region which is *enclosed by* the thorax. In general for specifying the semantics of relations in complex systems like the FMA or GALEN it is important to characterize relations in separation first by employing the simplest possible theory. Complex relations then can be described by combining the theories characterizing the components of the complex relation.

The parthood relation among masses is the third example of a partial ordering relation. Examples of masses are body-substances like saliva, semen, cerebrospinal fluid, inhaled air, urine, feces, blood, plasma, etc. The relation I have in mind here is the relation which holds between the blood in my body and the blood in my left ventricle. Notice that we do not have a relation of containment here. Rather names of containers like 'my body' or 'my left ventricle' are used here only in order to refer to certain quantities or portions of the blood in my body at a certain moment in time.

One can now verify that the parthood relation among masses is a proper partial ordering relation: the blood in my heart is a proper part of the blood in my body (but not *vice versa*), the blood in my right ventricle is a proper part of the blood in my heart, and the blood in my right ventricle is a proper part of the blood in my body.

From these examples we can see that all three relations share the property that they form partial ordering structures. Yet they are quite different in nature. It will our task in the remainder of this paper to characterize these distinctions formally.

Complexes

The characteristic property of complexes is that we can represent their partonomic structure using trees as indicated in Figure 1.

The formal theory of the relation component-of employs a binary primitive $x <_{cp} y$ which is interpreted as 'the entity x is a component(-part) of the entity y'. We then add the axioms for asymmetry and transitivity for $<_{cp}$ (ACP1-2)

$$\begin{array}{ll} ACP1 & x <_{cp} y \rightarrow \neg y <_{cp} x \\ ACP2 & (x <_{cp} y \wedge y <_{cp} z) \rightarrow x <_{cp} z \end{array}$$

together with definitions for the improper component-of relation (which includes identity) and for component-overlap ($D_{\leq_{cp}}$ and $D_{O_{cp}}$)

$$\begin{array}{ll} D_{\leq_{cp}} & x \leq_{cp} y \equiv x <_{cp} y \vee x = y \\ D_{O_{cp}} & O_{cp}\, xy \equiv (\exists z)(z \leq_{cp} x \wedge z \leq_{cp} y). \end{array}$$

One can see that these axioms and definitions are exactly analogous to what was presented in the section

on properties of partial orderings. As shown above it then follows that the component-of relation, $<_{cp}$, is irreflexive and that $\leq_{cp}$ is a partial ordering.

Axioms for the tree structure

We now characterize the specific character of the component-of relation beyond the fact that it has the structure of a (proper) partial ordering. We do so by adding axioms which constrain the partial order in such a way that the resulting component-of hierarchy is a finite tree structure.

For this purpose we introduce two additional predicates, one which holds for the root of the tree structure ($D_{root_{cp}}$) and another which holds for atomic components, i.e., entities without a component ($D_{At_{cp}}$).

$$\begin{array}{ll} D_{root_{cp}} & root_{cp}\ x \equiv (y)(y \leq_{cp} x) \\ D_{At_{cp}} & At_{cp}\ x \equiv \neg(\exists y)(y <_{cp} x) \end{array}$$

The component-of relation $\leq_{cp}$ is now governed by further axioms in addition to ACP1-2 (the $<_{cp}$-counterparts of APO1-2). These additional axioms fall into two groups, axioms which enforce the tree structure and the finiteness of this structure respectively. We start by discussing the first group:

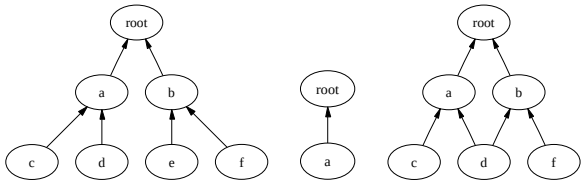
$$\begin{array}{ll} ACP3 & (\exists x) root_{cp}\ x \\ ACP4 & O_{cp}\ xy \rightarrow (x \leq_{cp} y \vee y <_{cp} x) \\ ACP5 & x <_{cp} y \rightarrow (\exists z)(z <_{cp} y \wedge \neg\ O_{cp}\ xz) \end{array}$$

ACP3 demands that every component-tree has a root. Using the antisymmetry of $\leq_{cp}$ we can then prove that there exists exactly one root. This rules out the structure in Figure 3(d) from being a component-of tree.

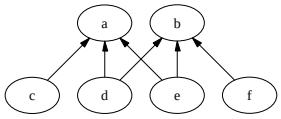
ACP4 is a version of what I shall call the *no-partial-overlap principle* (NPO). It rules out the possibility of partial overlap of components by demanding that if the complexes x and y share a common component then either x is a component of y, or x and y are identical, or y is a component of x. From this it follows that cycles like the one shown in Figure 3(c) cannot occur in component-of-trees.

Notice that the no-partial-overlap principle (NPO) also rules out the possibility that two different body systems which overlap (like the respiratory system and the alimentary system which share the component oropharynx) can exist within in the same component-of tree. This is because the two systems belong to distinct partitions of the human body (in the sense of the theory of granular partitions [1]), which is to say to different anatomical views or perspectives.

For example, the respiratory system has as components everything that is involved in the respiration process, and the alimentary system has as components everything that is is involved in the process of nutrition intake, digestion, and excretion. Clearly, there are parts of the body which have multiple functions, and therefore are components of different bodily systems. Each system has its own component-of tree with the particular system as a whole as the root. This corresponds to the view defended by Rector et al. [18] who argue that it is an important aspect of the design of ontologies to represent different views by means of separate tree structures.



(a) Component tree (b) Chain (c) Partial overlap (cycles)



(d) Multiple roots

Figure 3: Component trees and non-trees.

ACP5 demands that if x is a component of y then there exists a component z of y such that x and z do not overlap. This rules out cases where a complex has only a single proper component. In particular, it rules out graphs like the one shown in Figure 3(b) from being representations of component-of trees. ACP5 is a version of what, following Simons [23], I call the *weak supplementation principle* (WSP).

The second group of axioms that characterizes the component-of relation beyond the properties of being a partial ordering are axioms which enforce the finiteness of the component tree. ACP7 ensures that every complex has at least one atom as component. This ensures that no branch in the tree structure is infinitely long [30, 15]. Finally ACL8 is an axiom schema which enforces that every complex is either an atom or has only finitely many components. This ensures that com-

ponent trees cannot be arbitrary broad.

$$\begin{array}{ll} ACP6 & (\exists y)(At_{cp}\ y \wedge y \leq_{cp} x) \\ ACP7 & \neg At_{cp}\ y \rightarrow (\exists x_1, \ldots, x_n)((\bigwedge_{1\leq i\leq n} x_i <_{cp} y) \wedge \\ & \quad (z)(z <_{cp} y \rightarrow \bigvee_{1\leq i\leq n} z = x_i)) \end{array}$$

Here $(\bigwedge_{1\leq i\leq n} x_i <_{cp} y)$ is an abbreviation for $x_1 <_{cp} y \wedge \ldots \wedge x_n <_{cp} y$ and $\bigvee_{1\leq i\leq n} z = x_i$ for $x_1 = z \vee \ldots \vee x_n = z$.

Extensionality

Extensionality is a property of the component-of relation which tells us that two complexes are identical if and only if they have the same components. For example if the complex c_1 has the components a and b and the complex c_2 has the components a and b then c_1 and c_2 are the same complex. This kind of reasoning might seem trivial from a human perspective, but it may be very useful to enable a computer to identify and to distinguish complexes by means of their components. Moreover, when specifying the semantics of the component-of relation it is important that the property of extensionality is covered by the formalism.

In this context it is important to stress once more that we here assume an atemporal framework in which we consider reality only as it exists *at a single moment in time*. This means that we do not take into account the fact that a complex can have different components at different times. For example, I might lose one of my fingers but still my hand before and after the accident are the same complex. How things preserve their identity while undergoing changes in this way is a difficult and controversial subject. For discussions see for example [28, 12, 14].

Given the above axioms for the component-of relations, we can in fact prove that it has the property of extensionality. This is because, using ACP1, ACP2, ACP4, and ACP5, we can prove that two complexes are identical if and only if they have the same components (TCP1). Moreover using ACP6 we can prove in addition that two complexes are identical iff they have the same atomic components (TCP2).

$$\begin{array}{ll} TCP1 & (\exists z)(z <_{cp} x) \rightarrow \\ & \quad (x = y \leftrightarrow (z)(z <_{cp} x \leftrightarrow z <_{cp} y)) \\ TCP2 & x = y \leftrightarrow (z)(At_{cp}\ z \rightarrow (z \leq_{cp} x \leftrightarrow z \leq_{cp} y)) \end{array}$$

Notice that neither TCP1 nor TCP2 is derivable from the axioms for a partial ordering alone. Both are consequences of the partial ordering axioms *in conjunction* with the specific axioms which we added in order to characterize the component-of relation. Consequently, relations which are only characterized to be a partial ordering may or may not be extensional in the sense described above. Therefore omitting axioms ACP3-7 means leaving important properties of the relation in question unspecified.

To be sure, the principles discussed here are built implicitly into systems like the FMA or GALEN. The important point, however, is that in order to explicate relations like component-of it is critical to make such axioms explicit.

Theorem TCP2 is also interesting from a computational perspective. Clearly, when comparing complexes it is much easier to check the identity only of atomic components rather than of all components.

Parthood among masses

An important aspect of entities classified as *masses* is that they do not have any compositional structure. This means that parts can be carved out from the original mass in an arbitrary fashion. Consider, for example, body-substances like blood, plasma, urine, etc. They can be separated arbitrarily into quantities, for example, by pouring them into containers or – abstractly – by applying fiat boundaries [25]. According to the FMA [7] we can distinguish, for example, the blood in containers like my right ventricle, my artery, my coronary artery, and so on; we can apply fiat boundaries and distinguish the blood in the left part and the right part of my body or the blood in the upper and lower parts of my body. All these operations carve out parts or quantities of the original mass. (See also [9].)

We start the formal treatment of the parthood relation among masses by introducing the binary primitive $x <_M y$ which is interpreted as 'the mass x is a proper part of the mass y'. We then add the axioms for asymmetry and transitivity (referred to as AM1-2) together with definitions for the improper parthood relation and for overlap (referred to as D_{O_M} and $D_{\leq_M}$, respectively) along the lines discussed in the opening paragraphs of the section on complexes. In this section we omit the statement of those axioms and definitions. As discussed above we then can prove that $\leq_M$ is a partial ordering relation.

In contrast to the component-of relation, the parthood relation among masses does not form a tree structure. This is because partial overlap can occur between masses. Consider, for example, the relation of overlap between the blood in the left part of my body and the blood in the upper part of my body. They partly overlap since they share a common quantity of blood, namely the blood in the upper left part of my body, but neither is part of the other. Consequently, we cannot have the no-proper-overlap principle (NPO) as an axiom or theorem in our theory of $<_M$.

On the other hand we clearly need the weak supplementation principle (WSP) to be an axiom or theorem of such a theory, since WSP ensures that there cannot

be a mass that has a single proper part. Adding WSP as an axiom to this theory, however, is insufficient if we want to be able to identify and to distinguish masses in terms of their proper parts by means of a principle of extensionality similar to the one for complexes discussed above. (For details on why this is the case see [23].)

In order to characterize $\leq_M$ beyond its structure as a partial ordering we add an axiom to the effect that if x is not a part of y then there exists a z such that z is part of x and z does not overlap y (AM3).

$$AM3 \quad \neg x \leq_M y \rightarrow (\exists z)(z \leq_M x \wedge \neg\, O_M\, zy)$$

To see that AM3 is a sensible axiom consider the blood in my heart and the blood in my left ventricle. Clearly, the former is not a part of the latter. Moreover, the blood in my heart has parts, for example the blood in my right ventricle, which do not overlap with the blood in the left ventricle.

Using AM3 we can then prove the $<_M$ -counterpart of the weak supplementation principle (WSP) as a theorem (TM1), which then ensures that there cannot be a mass that has a single proper part. Using AM4 we can also prove that two masses are identical if and only if they have the same proper parts (TM2).

$$\begin{array}{ll} TM1 & x <_M y \rightarrow (\exists z)(z <_M y \wedge \neg\, O_M\, zx) \\ TM2 & (\exists z)(z <_M x) \rightarrow \\ & \quad (x = y \leftrightarrow (z)(z <_M x \leftrightarrow z <_M y)) \end{array}$$

Consequently, the property of extensionality holds for $<_M$.

The theory governing the compositional structure of masses, formed by AM1-3 together with the definitions for $\leq_M$ and O_M, is known in the literature as extensional mereology [23].

Containment

Consider Figure 2. Here we have a sequence of nested containers: my heart, containing a certain quantity of blood; my pericardial sac containing my heart; my thorax containing, among other anatomical entities, my pericardial sac. As pointed out above, containment understood in this sense is irreflexive, asymmetric, and transitive.

In our theory of containment we now introduce a binary primitive $x <_{ct} y$, which is interpreted as 'the entity x is contained in the entity y' together, with the axioms of asymmetry and transitivity (referred to as ACT1-2). We also add the usual definitions for overlap O_{ct} and for improper containment which includes identity $\leq_{ct}$, exactly analogous to those in the opening paragraphs of the section on complexes. We then can prove that $\leq_{ct}$ is a partial ordering relation.

Notice that, in contrast to the case of masses and complexes, we cannot here have the weak supplementation principle (WSP) either as an axiom or as a theorem in a theory of containment. This is because there are examples of containers with only one contained entity: my brain is contained in my skull; my sister is carrying a single baby in her uterus; my pericadial sac contains my heart as the only entity, etc. Those examples would be ruled out by a theory which contained WSP.

On the other hand, our theory of containment should permit us to identify or distinguish containers – at a given point in time – by means of the entities they contain. We therefore add an axiom to the effect that if (i) x has at least one contained entity, and (ii) every entity contained in x is also contained in y, then x is contained in y (ACT3).

$$\begin{array}{ll} ACT3 & ((\exists z) z <_{ct} x \wedge \\ & \quad (z)(z <_{ct} x \rightarrow z <_{ct} y)) \rightarrow x \leq_{ct} y \end{array}$$

The idea of modeling containment using the axioms ACT1-3 is due to Brock Decker. For details see [5].

Using the definition of $\leq_{ct}$ and the axioms ACT1-3 we can now prove that two containers x and y are identical iff they are non-empty and they contain the same entities (TCT1):

$$\begin{array}{ll} TCT1 & ((\exists z) z <_{ct} x \wedge \\ & \quad (z)(z <_{ct} x \leftrightarrow z <_{ct} y)) \leftrightarrow x = y \end{array}$$

Consider Figure 2. Like complexes, containers form tree-like structures in the sense that (1) there is a maximal container and (2) containers do not partially overlap. The structure is tree-*like* since there can be containers with only a single contained entity and hence nodes with a single child node in the corresponding tree representation (as the one shown in Figure 3(b)).

Formally we define predicates for the root, $root_{ct}$, and for atoms, At_{ct}, in terms of $\leq_{ct}$ exactly analogous to the definitions $D_{root_{cp}}$ and $D_{At_{cp}}$ in the section on complexes.

$$\begin{array}{ll} D_{root_{ct}} & root_{ct}\, x \equiv (y)(y \leq_{ct} x) \\ D_{At_{ct}} & At_{ct}\, x \equiv \neg(\exists y)(y <_{ct} x) \end{array}$$

The root here is understood as the maximal container and atoms are understood as entities which themselves do not contain any other entities.

We then add axioms ACT4 and ACT5 in terms of $root_{ct}$, $<_{ct}$, $\leq_{ct}$ and O_{ct} exactly analogous to ACP3 and ACP4.

$$\begin{array}{ll} ACT4 & (\exists x) root_{ct}\, x \\ ACT5 & O_{ct}\, xy \rightarrow (x \leq_{ct} y \vee y <_{ct} x) \end{array}$$

Here ACT4 enforces the existence of a root container. ACT5 is an instance of the no-partial-overlap principle (NPO) and rules out the partial overlap of containers.

Finally we add axioms ACT6 and ACT7 enforcing the condition that the resulting tree-like containment structures are finite. ACT6+7 are the $<_{ct}$-counterparts of ACP6+7.

$$\begin{array}{ll} ACT6 & (\exists y)(At_{ct}\, y \wedge y \leq_{ct} x) \\ ACT7 & \neg At_{ct}\, y \rightarrow (\exists x_1, \ldots, x_n)((\bigwedge_{1 \leq i \leq n} x_i <_{ct} y) \wedge \\ & \quad (z)(z <_{ct} y \rightarrow \bigvee_{1 \leq i \leq n} z = x_i)) \end{array}$$

Conclusions

The theories of the component-of, mass-part-of, and contained-in relations presented in this paper share the fact that they all are partial orderings and satisfy the principle of extensionality. (In Table 1 this is indicated by the + symbols.) That the principle of extensionality is satisfied means that at a given moment in time we can identify and distinguish masses in terms of their proper parts, complexes in terms of their components, and containers in terms of the entities they contain. The fact that this kind of reasoning is permitted, however is not implied by the underlying partial ordering structure. Other principles needed to be added in order to support this kind of reasoning. I showed that the same principles allow us to distinguish these relations formally.

relation	part. order	WSP	NPO	EXT
component-of	+	+	+	+
mass-part-of	+	+	–	+
contained-in	+	–	+	+

Table 1: Theories of partial ordering relations and their underlying principles.

The principles which allow us to distinguish the three relations are the weak supplementation principle (WSP) and the no-proper-overlap (NPO) principle. The former holds in the theories of the component-of and the mass-part-of relations but not in the theory of the contained-in relation (indicated by the $-$ symbol). The weak supplementation principle in the theories of component-of and mass-part-of tells us that a mass or a complex cannot have a single proper part or component. The no-partial-overlap principle in the theories of component-of and contained-in tells us that there cannot be partial overlap among components of complexes and among containers.

Acknowledgments

My thanks go to Barry Smith and Anand Kumar for helpful comments. Support from the Wolfgang Paul Program of the Alexander von Humboldt Foundation is gratefully acknowledged.

References

[1] T. Bittner and B. Smith. A theory of granular partitions. In M. Duckham, M. F. Goodchild, and M. F. Worboys, editors, *Foundations of Geographic Information Science*, pages 117–151. London: Taylor & Francis, 2003.

[2] H. Burkhardt and C.A. Dufour. Part/whole i: History. In H. Burkhardt and B. Smith, editors, *Handbook of Metaphysics and Ontology*, pages 663 – 673. Muenchen, Philosophia, 1991.

[3] R. Casati and A.C. Varzi. *Holes and Other Superficialities*. MIT Press, Cambridge, Mass., 1994.

[4] R. Casati and A.C. Varzi. The structure of spatial localization. *Philosophical Studies*, 82(2):205–239, 1995.

[5] B. Decker. Some axioms for beer glasses and backpacks. Technical report, Department of Philosophy, University at Buffalo, 2003.

[6] M. Donnelly. On parts and holes: The spatial structure of the human body. In *Proceedings of MedInfo 2004*, 2004.

[7] FMA. *The Foundational Model of Anatonomy*, Dec. 2003.

[8] GALEN. *OpenGALEN anatonomy, version 1.7, Build 930*, Feb. 2003.

[9] P. Gerstl and S. Pribbenow. Midwinters, end games, and body parts: a classification of part-whole relations. *Int. J. Human-Computer Studies*, 43:865–889, 1995.

[10] N. Guarino, S. Pribbenow, and L. Vieu. Modeling parts and wholes. *Data & Knowledge Engineering*, 20(3):257–258, 1996.

[11] U. Hahn, S. Schulz, and M. Romacker. Partonomic reasoning as taxonomic reasoning in medicine. In *Proceedings of the 16th National Conference on Artificial Intelligence and 11th Innovative Applications of Artificial Intelligence Conference*, pages 271–276, 1998.

[12] K. Hawley. *How things persist*. Oxford : Clarendon Press, 2001.

[13] I. Johansson. On the transitivity of the parthood relations. In H. Hochberg and K. Mulligan, editors, *On Relations and Predicates*. Ontos-verlag, 2004.

[14] E. J. Lowe. *A survey of Metaphysics*. Oxford University Press, 2002.

[15] C. Masolo and L. Vieu. Atomicity vs. infinite divisibility of space. In C. Freksa and D. Mark, editors, *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science. International Conference*

COSIT'99, volume 1661 of *Lecture Notes in Computer Science*. Springer-Verlag, 1999.

[16] J. Mejino, N. Noy, M. Musen, and C. Rosse. Representation of structural relationships in the foundational model of anatomy. 2003.

[17] J. L. V. Mejino, A. V. Agoncillo, K. L. Rickard, and C. Rosse. Representing complexity in part-whole relationships within the foundational model of anatomy. In *Proceedings of the American Medical Informatics Association Fall Symposium*, 2003.

[18] A. Rector, J. Rogers, A. Roberts, and C. Wroe. Scale and context: Issues in ontologies to link health- and bio-informatics. In *Proceedings of the AMIA 2002 Anual Symposium*, pages 642–646, 2002.

[19] N. Rescher. Axioms for the part relation. *Philosophical Studies*, 6:8–11, 1955.

[20] J. Rogers and A. Rector. Galen's model of parts and wholes: experience and comparisons. In *Proceedings of the AMIA Symp 2000*, pages 714–8, 2000.

[21] C. Rosse and J. L. V. Mejino. A reference ontology for bioinformatics: The foundational model of anatomy. *Journal of Biomedical Informatics*, in press, 2003.

[22] S. Schulz and U. Hahn. Mereotopological reasoning about parts and (w)holes in bio-ontologies. In C. Welty and B. Smith, editors, *Formal Ontology in Information Systems. Collected Papers from the 2nd International Conference*, pages 210 – 221, 2001.

[23] P. Simons. *Parts, A Study in Ontology*. Clarendon Press, Oxford, 1987.

[24] P. Simons. Part/whole ii: Mereology since 1900. In H. Burkhardt and B. Smith, editors, *Handbook of Metaphysics and Ontology*, pages 673 – 675. Muenchen, Philosophia, 1991.

[25] B. Smith. Fiat objects. *Topoi*, 20(2):131–48, 2001.

[26] B. Smith and C. Rosse. The role of foundational relations in the alignment of biomedical ontologies. 2003.

[27] B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the gene ontology. In *Proc. Annual Symposium of the American Medical Informatics Association*, pages 609– 613, 2003.

[28] J. J. Thomson. Parthood and identity across time. *Journal of Philosophy*, 80:201–220, 1983.

[29] A. Varzi. Parts, wholes, and part-whole relations: The prospects of mereotopology. *Data and Knowledge Engineering*, 20(3):259–86, 1996.

[30] A. Varzi. Mereology. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Stanford: CSLI (internet publication), 2003.

[31] M.E. Winston, R. Chaffin, and D. Herrmann. A taxonomy of Part-Whole relations. *Cognitive Science*, 11:417–444, 1987.

Investigating Subsumption in DL-Based Terminologies: A Case Study in SNOMED CT

Olivier Bodenreider [1], Barry Smith [2,3], Anand Kumar [2], Anita Burgun [4]

[1] US National Library of Medicine, Bethesda, Maryland, USA
[2] Institute for Formal Ontology and Medical Information Science, Univ. Leipzig, Germany
[3] Department of Philosophy, University at Buffalo, New York, USA
[4] Laboratoire d'Informatique Médicale, Université de Rennes I, France

Formalisms such as description logics (DL) are sometimes expected to help terminologies ensure compliance with sound ontological principles. The objective of this paper is to study the degree to which one DL-based biomedical terminology (SNOMED CT) complies with such principles. We defined seven ontological principles (for example: each class must have at least one parent, each class must differ from its parent) and examined the properties of SNOMED CT classes with respect to these principles. Our major results are: 31% of the classes have a single child; 27% have multiple parents; 51% do not exhibit any differentiae between the description of the parent and that of the child. The applications of this study to quality assurance for ontologies are discussed and suggestions are made for dealing with multiple inheritance.

INTRODUCTION

Biomedical terminologies and ontologies are increasingly taking advantage of Description Logics (DL) in representing knowledge. GALEN[1] and SNOMED Clinical Terms® (in what follows SNCT)[2] were both developed in a native DL formalism. Several other groups have worked at converting existing terminologies into terminologies with a DL formalism (UMLS® Metathesaurus® [1-3], UMLS Semantic Network [4], Gene Ontology™ [5], National Cancer Institute Thesaurus [6]). Protégé-2000's OWL plug-in now also allows developers of frame-based resources to export their ontologies into DL formalism.

The validation of an ontology by a DL-based classifier allows compliance with certain rules of classification (e.g., absence of terminological cycles) and it brings also other benefits in terms of coherence checking and query optimization [7, 8]. However, neither a DL formalism nor the use of a classifier can ensure compliance with all principles of a sound ontology [9].

The objective of this paper is to study the degree to which one DL-based biomedical terminology complies with such ontological principles. We selected SNCT as target for this evaluation because it is the most comprehensive biomedical terminology recently developed in native DL formalism. Another reason for our choice is that SNCT will soon be available as part of the UMLS[3] (at no charge for UMLS licensees in the U.S.) and is therefore likely to become widely used in medical information systems.

This paper is organized as follows. We first define a limited number of basic ontological principles with which biomedical ontologies are expected to be compliant. (These are in effect principles of good classification.) We then give a brief description of SNCT, we present the methods used to test the compliance of SNCT with these principles, and we summarize our results. Finally, we discuss the application of this method to quality assurance in ontologies and terminologies, laying special emphasis on the role of creating partitions in ontologies, and we also outline other implications of our results.

BACKGROUND

Terms, classes, and instances. We shall refer to the nodes in SNCT not as concepts but rather on the one hand as *terms* (where we are interested in the hierarchy itself, as a syntactic structure), and on the other hand as *classes* (where we are interested in the biological entities to which these terms refer). It is classes, not concepts, which stand in *IS A*, *PART OF* and similar relations in biomedical ontologies. Classes have *instances*. In the biomedical domain, instances are generally represented in health information systems (e.g., electronic patient records) or in biomedical experiments (e.g., in the form of microar-

[1] http://www.opengalen.org/
[2] http://www.snomed.org/snomedct_txt.html
[3] http://umlsinfo.nlm.nih.gov/

ray experiments), while biomedical terminologies and ontologies are focused on classes and their relations.

Relations among classes. The possible relations of class *A* to class *B* are defined in Table 1. *A* is the root of a given taxonomy if and only if every class in the taxonomy is a child of *A*; conversely, *A* is a leaf of a given taxonomy if and only if *A* has no children.

Relation	Definition
$A = B$	*A* and *B* are the same entity (i.e., they have the same definition, and thus also the same family of instances at any given time)
A *IS A* B	1. *A* and *B* are classes and 2. all instances of *A* are instances of *B*
A is a child of *B*	1. A *IS A* B, 2. $A \neq B$, and 3. if A *IS A* C and C *IS A* B then $A = C$ or $C = B$
A and *B* are siblings	1. there is some *C* of which *A* and *B* are both children and 2. $A \neq B$
A is a parent of *B*	*B* is a child of *A*
C is a differentia of *A* with respect to *B*	1. A *IS A* B, 2. $A \neq B$, and 3. instances of *A* are marked out within the wider class *B* by the fact that they exemplify *C*

Table 1 – Definition of the relations between classes A *and* B

Principles of classification. Scientific classification has evolved from Aristotle to Linnaeus to large and varied classifications of modern times. Along the way, classification principles were elaborated. One such principle, resulting from the use of a unique *fundamentum divisionis* or single classificatory principle in differentiating the species of each successive genus, is that subclasses be mutually exclusive and jointly exhaustive [10]. Some other highly general organization and classification principles – which we believe rest on a wide consensus among those working on biomedical terminologies [11, 12] – are:

- Each hierarchy must have a single root
- Each class (except for the root) must have at least one parent
- Non-leaf classes must have at least two children
- Each class must differ from each other class in its definition. In particular: each child must differ from its parent and siblings must differ from one another

Principles of subsumption. More interestingly, principles can also be derived from the study of the way subsumption is in fact treated in biomedical terminologies and ontologies. As noted by Bernauer [13], two major types of difference can be observed between a parent and its child: the introduction in the child of a new "criterion" (introduction of a *role* in DL parlance), and the *refinement* of an already existing criterion (corresponding to DL's *refinement of a role value*[4]). For example, the introduction of the role *CAUSATIVE AGENT* with value *Infectious agent* explains the subsumption relation of *Meningitis* to *Infective meningitis*. Similarly, the subsumption relation of *Infective meningitis* to *Viral meningitis* is explained by the refinement of the role value for *CAUSATIVE AGENT* since *Infectious agent* subsumes *Virus*. Such refinement can be a matter of specialization as in the previous example, where the role value for the parent is more generic than that for the child. Less frequently, partitive refinement can occur. For example, *Neuropathy* subsumes *Peripheral motor neuropathy* because the value in the parent of the role *FINDING SITE* (*Nerve structure*) includes as part the corresponding value in the child (*Peripheral motor neuron*).

The following *inheritance principle* is standardly taken for granted in work on ontologies and terminologies: if *A* is a child of *B* then all properties of *B* are also properties of *A*. As a corollary, no cycles are allowed in an *IS A* hierarchy. Additionally, one inheritance principle based on our approach to subsumption can be expressed as follows: All roles of a parent class must either be inherited by each child or refined in the child. From the perspective of the child, differentia from child to parent should uniquely result in every case either from refinement of the value of a common role or introduction of a new role

Single vs. multiple inheritance. Some of the principles presented above are the object of a large consensus (e.g., *that each class must have at least one parent* is needed if a terminology is to have a proper hierarchical structure). Others, however, still spur debate among terminology developers. This is the case in regard to the issue of single vs. multiple inheritance, i.e., of whether classes should be allowed to have more than one parent. As noted by Cimino: "There is some disagreement, however, as to whether concepts should be classified according to a single taxonomy (strict hierarchy) or if multiple classifications (polyhierarchy) can be allowed." While it is beyond the scope of this paper to argue for or against multiple inheritance, we will make some suggestions for dealing with this issue in the discussion.

[4] Also called role filler in DL parlance.

MATERIALS

SNOMED CT was formed by the convergence of SNOMED RT and Clinical Terms Version 3 (formerly known as the Read Codes). The version used in this study (January 31, 2004) contains 269,864 classes. The first level is subdivided into eighteen classes listed in Table 2 with their frequency distribution.

Class	Freqency
Attribute	990
Body structure	30,651
Clinical finding	95,604
Context-dependent categories	3,648
Environments and geographical locations	1,619
Events	86
Observable entity	7,273
Organism	25,025
Pharmaceutical / biologic product	16,866
Physical force	198
Physical object	4,200
Procedure	46,065
Qualifier value	8,133
Social context	4,895
Special concept	177
Specimen	1,052
Staging and scales	1,097
Substance	22,266

Table 2 – The 18 first-level classes in SNOMED CT and their frequency distribution

Role	Value
CAUSATIVE AGENT	*Virus*
ONSET	*Sudden onset; Gradual onset*
SEVERITY	*Severities*
EPISODICITY	*Episodicities*
COURSE	*Courses*
ASSOCIATED MORPHOLOGY	*Inflammation*
FINDING SITE	*Meninges structure*

Table 3 – Roles present in the description of Viral meningitis

Each SNCT class has a description[5] consisting of 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 meningitis*, and *Aseptic meningitis, viral*). The roles present in the description of this class are listed in Table 3.

In addition to a unique identifier, each class is assigned a unique, fully specified name consisting of a regular name suffixed (in parentheses) with a reference to what SNCT calls the "primary hierarchy" of the class, the latter corresponding roughly to one of the top-level classes in the hierarchy. For example, the fully specified name for *Viral meningitis* is *Viral meningitis (disorder)*[6]. This assignment to a primary hierarchy is not explicitly recognized as a property of the class in the SNCT representation. However, because the corresponding high-level category can be easily extracted from the fully specified name of the class, we found it useful it to use it for purposes of categorizing SNCT classes. Thus for example we will use *disorder* as the category for *Viral meningitis*. The list and frequency distribution of such categories in SNCT is presented in Table 4.

administrative concept	54	*navigational concept*	165
assessment scale	870	*observable entity*	7,274
attribute	991	*occupation*	4,153
body structure	25,395	*organism*	25,026
cell	603	*person*	302
cell structure	501	*physical force*	199
context-dependent category	3,649	*physical object*	4,201
disorder	62,301	*procedure*	42,782
environment	1,007	*product*	16,867
environment / location	1	*qualifier value*	8,080
ethnic group	254	*regime/therapy*	3,284
event	87	*religion/philosophy*	145
finding	33,304	*social concept*	21
geographic location	612	*special concept*	1
inactive concept	7	*specimen*	1,053
life style	21	*staging scale*	15
morphologic abnormality	4,153	*substance*	22,267
namespace concept	5	*tumor staging*	213

Table 4 – The list of high-level categories ("primary hierarchies") in SNOMED CT and their frequency distribution

Inheritance in SNCT is indicated by the presence of *IS A* relationships among classes. For example, the class *Fracture of calcaneus* subsumes two classes (*Closed fracture of calcaneus* and *Open fracture of calcaneus*). The difference between the descriptions of the classes *Fracture of calcaneus* and *Closed fracture of calcaneus* lies in the presence of a specialized value for the role *ASSOCIATED MORPHOLOGY* in the child (*Fracture, open*[7]) compared to that of the parent (*Fracture*). Also of note, the class *Fracture* subsumes *Fracture, open*. The refinement of the value of the

[5] Throughout this paper, we use 'description' with the common meaning that is also standard in the DL-context, i.e., to refer to the list of properties of a given class (more precisely: of its instances), expressed by roles. In SNOMED CT parlance, however, a description corresponds to a name for a class.

[6] The primary hierarchy for *Viral meningitis* is *Clinical finding*, while the category mentioned in parentheses in the fully specified name is *disorder*.

[7] Despite similarities in their names, *Fracture, open (morphologic abnormality)* and *Open fracture (disorder)* are distinct classes in SNOMED CT.

role *ASSOCIATED MORPHOLOGY* between the two classes constitutes the differentia, while the other roles are all inherited from the parent class.

METHODS

The methods presented below were developed for testing the compliance of SNCT with the seven principles listed in Table 5.

P1	Each class must have at least one parent
P2	Non-leaf classes must have at least two children
P3	Children should have exactly one parent
P4	Each hierarchy must have a single root
P5	Each child's description must differ from its parent's description
P6	All roles of a parent class must either be inherited by each child or refined in the child
P7	Differentia from child to parent should uniquely result in every case either from refinement of the value of a common role or introduction of a new role

Table 5 – Ontological principles studied in SNCT

Quantitative analysis: Number of parents, children, and roots

By simply counting the number of parents and children for each class, we verify the degree of compliance with **P1**, **P2**, and **P3**. Additionally, the existence of a path between each class and the eighteen top-level classes is tested by traversing the graph of all classes in SNCT from each class upwards. We use this method for verifying **P4**.

Qualitative analysis of differentiae

In order to verify SNCT's compliance with **P5**, we analyze the differentiae in pairs of parent-child classes by comparing the roles and role values for each class in the pair. First, we verify that at least one role or one role value is present in the description of the child but not in that of the parent.

The second step consists in examining the roles shared by the two classes and those specific to each class. All roles of the parent are searched for in the description of the child in order to verify compliance with **P6**.

The relationship between the values of a role shared by the parent and child classes is examined and is expected to be either specialization (*IS A*) or partitive refinement (*PART OF*). The presence of roles specific to the child is also examined. The number of differentiae (i.e., the number of role values refined and of roles introduced in the child) is recorded. This step is used to verify **P7**.

RESULTS

Quantitative analysis: Number of parents, children, and roots

Number of children

The number of children per class ranges from 0 to 2532. The frequency distribution of the number of children is presented in Figure 1. 196,237 classes (73%) have no children. These classes are leaf nodes in the SNCT hierarchy. Examples of such classes include the substance *Tartrate dehydratase*, the finding *Anuria*, the organism *Trypanosoma evansi*, and the body structure *Upper left third premolar tooth*.

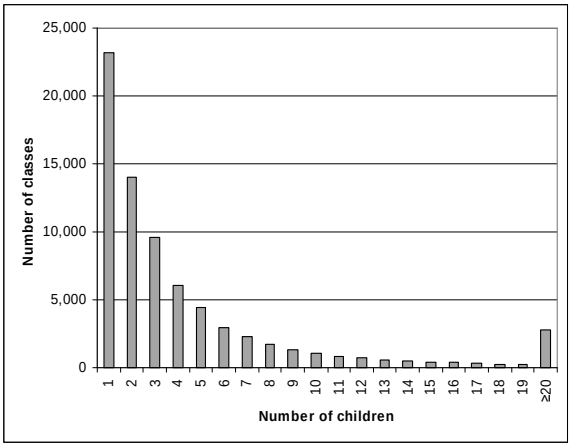


Figure 1 – Distribution of the number of children

Out of 73,627 classes with children, 23,174 classes (31.5%) have a single child. This proportion is relatively constant across SNCT categories. Examples of classes with a single child include {*Cervical secretion sample*, child: *Cervical mucus specimen*} (*specimen*), {*Deferoxamine*, child: *Deferoxamine mesylate*} (*substance*), {*Multiple polyps*, child: *Multiple adenomatous polyps*} (*morphologic abnormality*), and {*Referral to general medical service*, child: *General medical self-referral*} (*procedure*).

8,034 classes (11%) have ten children or more and 150 have more than 99 children. The median number of children is 2. Example of classes with a large number of children include *Infectious gastroenteritis* (10 children), *Operation on heart valve* (25 children), *Sodium compound* (51 children), and *Disorder of eye proper* (100 children).

Some classes have an unusually large number of children, including *Veterinary proprietary drug AND/OR biological* (2532 children), *Biochemical test* (996 children), the substance *Oxidoreductase* (580 children), the organism *Bos taurus* (551 children),

and *Congenital malformation* (505 children). Although these classes often correspond to large collections of drugs, tests, or disorders, the large number of children in these classes may point to issues such as a lack of organization or incomplete descriptions.

Number of parents

Except for the root, every class of SNCT has at least one parent. The number of parents per class ranges from 1 to 13.[8] The frequency distribution of the number of children is presented in Figure 2. 195,053 classes (72.3%) have a single parent, 53,517 classes (19.8%) have two parents, 13,969 classes (5.2%) have three, 4,692 classes (1.7%) have four, and 2,632 classes (1.0%) have five or more.

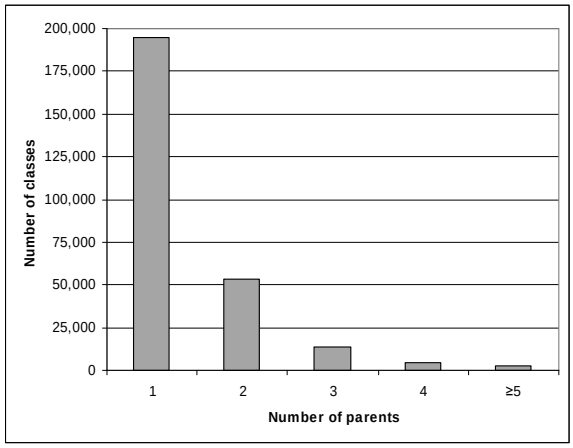


Figure 2 – Distribution of the number of parents

Overall, the proportion of classes having multiple parents, i.e., exhibiting multiple inheritance, is 27.7%. This proportion tends to be higher in some categories (e.g., around 45% for *body structure*, *disorder*, and *procedure*) and lower in others (e.g., around 5-15% for *cell*, *organism*, and *substance*).

Number of roots

Except for the root and for the eighteen top-level classes of SNCT excluded from this test, each class can be linked hierarchically to exactly one top-level class. This means that SNCT consists of eighteen independent hierarchies.

Qualitative analysis of differentiae

Existence of a differentia between parent and child

Out of the 377,681 parent-child relations examined, 193,957 (51%) do not exhibit any differentiae between the description of the parent and that of the child. However, the presence or absence of differentiae in children varies considerably across categories. In most categories – including *geographical location*, *organism*, and *substance* – no differentiae are ever mentioned. In the other categories, the proportion of children exhibiting differentiae in their description ranges from 29% (*cell*) to 86% (*specimen*).

Number and nature of differentiae

When there does exist a differentia between a child and its parent, i.e., when their descriptions are not identical, the difference in the descriptions can affect one role or multiple roles, and one or more values within each role.

Single differentia. Out of the 183,724 parent-child relations where there is at least one differentia between the child and its parent, 102,426 (56%) exhibit exactly one differentia. For example, the classes *Fracture of calcaneus* and *Open fracture of calcaneus* presented earlier differ only by the value of their common role *ASSOCIATED MORPHOLOGY*. In 60% of the cases, the differentia comes from the refinement of the value for a given role; in 40% of the cases, it comes from the introduction of a new role in the child. The example above (*Fracture of calcaneus*) illustrates the refinement (from *Fracture* to *Fracture, open*) of the role *ASSOCIATED MORPHOLOGY*. Conversely, the introduction of the role *FINDING SITE* (with value *Ear structure*) differentiates the class *Otitis* from its parent *Inflammatory disorder*.

Multiple differentiae. In case of multiple differentiae, the differentiae involved reflect the introduction of several roles (34%), the refinement of several values (20%), or the combination of introducing at least one role and refining at least one value (46%). For example, *Endoscopy of jejunum* differs from *Procedure on jejunum* by 1) the introduction of two roles (*METHOD*, with value *Inspection – action*, and *ACCESS INSTRUMENT*, with value *Endoscope, device*) and 2) the refinement of the role *ACCESS* (from *Surgical access values* to *Endoscopic approach – access*). Figure 3 illustrates the roles introduced and inherited for the class *Endoscopy of jejunum*. Not surprisingly, multiple differentiae are often associated with multiple inheritance. In the example above, the role *METHOD* is actually inherited (and refined from *Evaluation – action* to *Inspection - action*) from *Gastrointestinal investigation*, the second parent of *Endoscopy of jejunum*. The role *ACCESS INSTRUMENT*, however, is truly specific to *Endoscopy of jejunum* (i.e., not present in any of its parents).

Our analysis of differentiae reveals a number of **other potentially problematic issues**. In 7,226 cases, some

[8] The three classes with 13 parents are *Anoscopy with coagulation for control of hemorrhage of mucosal lesion*, *Mandibuloacral dysostosis*, and *Entire sternocleidomastoid muscle*.

role or value present in the parent is not inherited or refined in the child. For example, the role *ONSET* has two possible values in the class *Subjective visual disturbance* (*Sudden onset* and *Gradual onset*), of which *Gradual onset* is not inherited by its child class *Sudden visual loss*. The role *ONSET* is involved in roughly half of the cases where some role is specific to a parent class but eleven other roles are also involved in this phenomenon.

In 21,799 cases, although the parent and child classes share a role, the values of this role are neither identical (inherited by the child from the parent) nor such as to stand in any taxonomic relation (with the specialized value in the child) or meronomic relation (with the part in the child). For example, the class *Diabetic retinopathy* and its child *Diabetic retinal microaneurysm* share the role *FINDING SITE*, but their values for this role (*Retinal structure* and *Visual pathway structure*) do not stand in a hierarchical relation. Typically, this problem is associated with multiple inheritance. The role value which does not stand in hierarchical relation with corresponding role values in one parent most often does in one of its other parents. In the example above, *Retinal structure* is actually inherited from *Retinal microaneurysm*, the other parent of *Diabetic retinal microaneurysm*.

DISCUSSION

The work described in this paper is in the tradition of studies auditing large medical terminologies such as [14]. However, we are interested here not just in the consistency of the terminological structure but also in compliance with general classification principles. We found SNCT to be fully compliant with principles such as *each class must have at least one parent* and *each hierarchy must have a single root*. In contrast, we observed non-compliance with many other principles, the consequence of which will be presented next. We will then revisit the problem of single vs. multiple inheritance and outline a possible solution to it.

Application to quality assurance for ontologies

Non-leaf classes with a single child

The recognition by biologists of the phylum *Chordata* rests on the distinction of several subphyla: *Vertebrata* (or *Vertebrates*), *Cephalochordata*, and *Urochordata*. Compared to *Vertebrates*, the latter two might be of lesser relevance to clinical medicine. However *Vertebrates* is defined in opposition to the two other subphyla and all three should therefore be represented in a well-formed ontology of organisms. Moreover, in a world in which *Vertebrates* had only one child, the distinction between parent and child would not be made by biologists. Therefore, the presence of such cases is reason to suspect the presence of error.

The review of a limited number of classes having a single child suggests the following possible issues. One is the incompleteness of the hierarchy (e.g., *Subphylum Vertebrata* is the only subphylum recorded in SNCT for *Phylum Chordata*). Another issue is the presence of a hybrid class, resulting from the intersection of two parent classes, as the single child of at least one of the two parent classes (e.g., *Closure of abdominothoracic fistula*, hybrid child of *Closure of fistula of thorax* and *Abdomen closure*) and single child of *Closure of fistula of thorax*). Finally, the presence of redundant classes, where a parent and a child class bear no differences, can also be at the origin of single child classes. This issue is discussed in detail in the next section.

Among the 23,174 single child classes, 12,928 (56%) have a single parent and therefore do not correspond to hybrid classes. Examples of such classes can be found in virtually every category and include the procedure *Arthroscopy of toe* (single child of *Arthroscopy of foot*), the disorder *Congenital absence of lobe of liver* (single child of *Congenital absence of liver*), and the substance *Urine* (single child of *Urinary tract fluid*).

Absence of difference in the description between children and parents

Beyond hierarchy, one of the major reasons for interest in DL-based systems is that they promise to make available for formal reasoning tools detailed descriptions for each class, representing through roles the defining characteristics of these classes. However, DL systems can also accommodate classes with minimal descriptions (i.e., restricted to bare subsumption links). We reviewed a small number of classes (in the domain of disorders) for which no difference was provided between the parent and the child in terms of roles or role values. The major issue brought to light by this limited analysis seems to be the incompleteness of many descriptions. For example, while no difference is provided between the descriptions of *Bullous lichen planus* and *Lichen planus*, such a difference is provided for *Bullous dermatosis* (*ASSOCIATED MORPHOLOGY* with value *Blister*) and *Skin lesion*. In other cases, the representation of some characteristics seems to have been purposely omitted (e.g., *COURSE* for acute and subacute variants of diseases, although *Acute* exists as a class). Generally, morphologic distinctions seem better represented than physiological ones. Also of note, some classes represent what are in fact mere collections (e.g., *Extrapyramidal disease*). These classes are defined in extension (i.e., via a list of their subclasses) rather

than in intension (i.e., via a list of characteristics). Extensional definitions are less desirable since they imply the need for more radical revisions in light of the discovery of new types of cases.

Finally, in some cases, there is actually no difference to be represented between the parent and the child class (e.g., *Closed fracture of skull without intracranial injury* vs. *Closed fracture of skull*). The issue, in this case, is the presence of two classes for representing one biomedical entity. The distinction between the two classes lies not in the biomedical entity they represent (i.e., the skull is fractured, but not open), but merely in the knowledge of the physician that intracranial injuries might be associated with such fractures. In other words, this distinction is epistemological in nature and, arguably, should not be represented in an ontology. It would be a valuable extension of the current DL in SNCT if ways could be found to do justice to operators, such as 'with' and 'without,' which play an important role in the organization of SNCT's term hierarchy. As things stand, the information conveyed by such operators is not accessible in ways which would support reasoning with terminological knowledge in medicine. This means more generally that the information conveyed by the compositional structure of SNCT's terms is at the moment not available for automatic retrieval.

Presence of roles specific to the parent class

In most of the cases we examined, the presence in a parent's description of roles not inherited by its children has to do with the representation of specialization in DL-based structures. As noted earlier, *Subjective visual disturbance* is described as having possibly a *Sudden onset* or a *Gradual onset*. However, the only valid onset for its child *Sudden visual loss* is *Sudden onset*. Therefore, *Sudden visual loss* can be seen as a specialization of *Subjective visual disturbance*. This could be represented in DL form by '∀(*HAS-ONSET Onsets*)' for *Subjective visual disturbance* and '∃(*HAS-ONSET Sudden onset*)' for *Sudden visual loss* [15].

Characterizing inheritance

The uncontrolled use of *IS A* to signify a variety of different sorts of relations (including *PART OF*, *IS AN INSTANCE OF*, and so on) results in what Guarino has called '*IS A* overload', which is often associated in turn with examples of incorrect subsumption [16]. Examples of this phenomenon in SNCT include *Both testes IS A Testis Structure*, *Deferoxamine mesylate IS A Deferoxamine*, and *Urine sediment IS A Urine*.

IS A overload, which is often associated with multiple inheritance, may be alleviated by making explicit which sort of subsumption link is involved in each specific type of case – for example by replacing *IS A* as it occurs between *Viral meningitis* and *Infective meningitis* with *IS* A_{AGENT} or as it occurs between *Viral meningitis* and *Viral infection of the central nervous system* with *IS* A_{SITE}.

The use of such explicit subsumption links also enables a large taxonomy such as SNCT to be divided into *partitions* within which taxonomic reasoning can be more reliably performed. Through a locative partition, for example, which we can think of as a window or view on reality with a specific type of focus, *Viral meningitis* would appear in its locative guise: as a *Viral infection of the central nervous system*, and inferences could be performed safely along the *IS* A_{SITE} relationship within this partition. Analogously, in a causative partition, *Viral meningitis* would be linked to *Infective meningitis* and subsumption could be performed safely along the *IS* A_{AGENT} relationship. The locative and causative partitions would then yield complementary views of different aspects of one and the same reality. This view is illustrated in Figure 4, and the underlying formal theory is presented in [17].

CONCLUSIONS

SNCT is the most comprehensive biomedical terminology recently developed in native DL formalism and is expected to play an important role in clinical information systems. Unlike thesauri built for information retrieval purposes, SNCT should enable reasoning about biomedical knowledge. We have listed some principles, mostly related to classification, and tested the degree to which SNCT complies with them. While we found SNCT to be more coherent than many other terminologies, we also found the description of many of its classes to be minimal or incomplete, with possible detrimental consequences on inheritance.

Description logics provide a formalism suitable for representing many features of a variety of different domains – including the biomedical domain – in a way that can support automatic reasoning and information retrieval. In and of themselves, however, DLs do not systematically ensure compliance with the principles of classification required if reasoning is to be performed accurately. More than the use of any formalism, we believe that compliance with sound ontological principles is what guarantees the accuracy of reasoning.

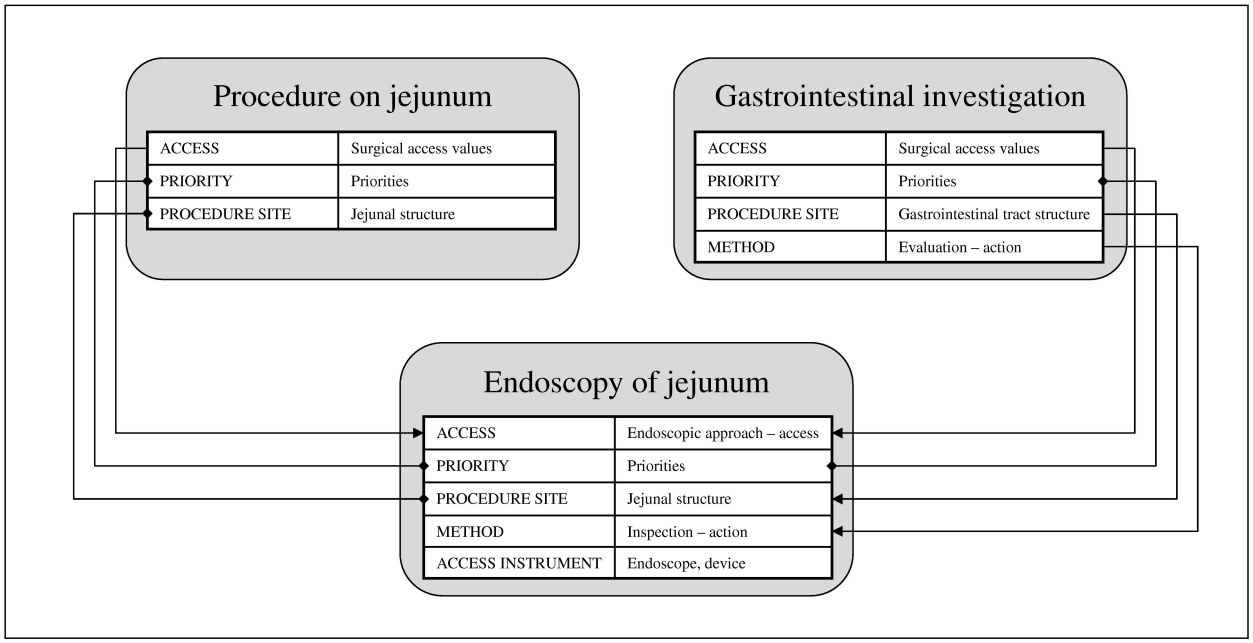


Figure 3 – Inheritance of role values for Endoscopy of jejunum.

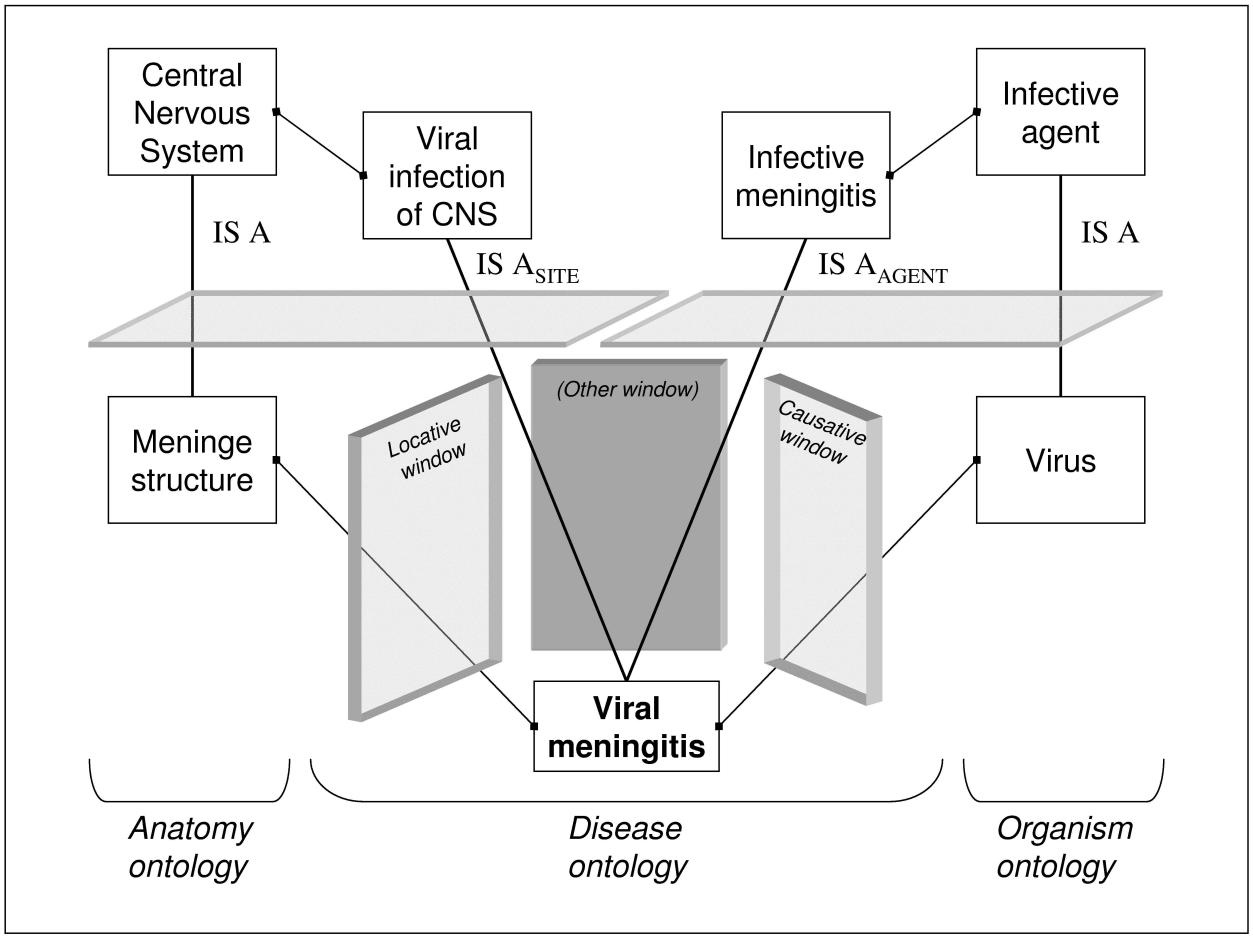


Figure 4 – Two views (locative and causative) on Viral meningitis.

Acknowledgements

Smith and Kumar are supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation.

References

1. Pisanelli DM, Gangemi A, Steve G. An ontological analysis of the UMLS Methathesaurus. *Proc AMIA Symp* 1998:810-4.
2. Cornet R, Abu-Hanna A. Usability of expressive description logics--a case study in UMLS. *Proc AMIA Symp* 2002:180-4.
3. Hahn U, Schulz S. Towards a broad-coverage biomedical ontology based on description logics. *Pac Symp Biocomput* 2003:577-88.
4. Kashyap V, Borgida A. Representing the UMLS Semantic Network using OWL: (Or "What's in a Semantic Web link?"). In: Fensel D, Sycara K, Mylopoulos J, editors. *The SemanticWeb - ISWC 2003*. Heidelberg: Springer-Verlag; 2003. p. 1-16.
5. Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput* 2003:624-35.
6. Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. *Journal of Web Semantics* 2003;1(1).
7. Horrocks I, Rector A, Goble C. A Description Logic based schema for the classification of medical data. In: Baader F, Buchheit M, Jeusfeld MA, Nutt W, editors. *Proceedings of the 3rd Workshop KRDB'96*; 1996. p. 24-28.
8. Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* 2000;16(2):184-5.
9. Ceusters W, Smith B, Flanagan J. Ontology and medical terminology: Why Description Logics are not enough. *Proceedings of TEPR 2003 - Towards an Electronic Patient Record. San Antonio, Texas, May 10-14, 2003* 2003:(CD-ROM publication).
10. Marradi A. Classification, Typology, Taxonomy. *Quality & Quantity* 1990;24(2):129-157.
11. Smith B. The Logic of Biological Classification and the Foundations of Biomedical Ontology. In: Westerståhl D, editor. *Invited Papers from the 10th International Conference in Logic Methodology and Philosophy of Science, Oviedo, Spain, 2003*: Elsevier-North-Holland; 2004. p. (to appear).
12. Michael J, Mejino JL, Jr., Rosse C. The role of definitions in biomedical concept representation. *Proc AMIA Symp* 2001:463-7.
13. Bernauer J. Subsumption principles underlying medical concept systems and their formal reconstruction. *Proc Annu Symp Comput Appl Med Care* 1994:140-4.
14. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. *J Am Med Inform Assoc* 1998;5(1):41-51.
15. Rector A. Defaults, context, and knowledge: Alternatives for OWL-indexed knowledge bases. *Pac Symp Biocomput* 2004:226-237.
16. Guarino N. Some ontological principles for designing upper level lexical resources. In: Rubio A, Gallardo N, Castro R, Tejada A, editors. *Proceedings of First International Conference on Language Resources and Evaluation. ELRA - European Language Resources Association, Granada, Spain*; 1998. p. 527-534.
17. Bittner T, Smith B. A theory of Granular Partitions. In: Duckham M, Goodchild MF, Worboys MF, editors. *Foundations of Geographic Information Science*. London: Taylor & Francis; 2003. p. 117-151.

STEEL: A Spatio-Temporal Extended Event Language for Tracking Epidemic Spread from Outbreak Reports

Hervé Chaudet
Faculté de Médecine de Marseille, Laboratoire d'informatique médicale (herve.chaudet@medecine.univ-mrs.fr)

Abstract

We propose a Spatio-Temporal Extended Event Language (STEEL) for representing and reasoning about events that are described in outbreak reports. This language is an extension of the Event Calculus based on mereotopological relationships and structured conglomeration of events, in which time is replaced with spatiotemporal location. It allows representing and building aggregates of events according to the spatiotemporal location of their occurrence. In a proof a concept study, we aimed at comparing the performances of an experimental implementation in Prolog of this language with 3 human experts during a question-answering task on a trial corpus of 35 outbreak reports. This experiment showed experts' agreement with the system's responses.

INTRODUCTION

The use of emailed reports for an early and wide dissemination of epidemiological information by the Internet shows an increasing success for monitoring epidemiological events since its introduction.[1] The descriptive possibilities of these texts and their ability to deal with unattended situations make them competitive, comparing to epidemic registries, for reporting emerging infectious disease outbreaks and unusual disease patterns, including biological threats. That is why we are developing a system for automatic processing of outbreak reports with a double objective of question answering and qualitative modeling.

However, if the use of emails make easier the notification of epidemiological events, the automatic analysis and use of the transmitted information is particularly challenging, especially in a question answering task, where the system must retrieve answers rather than documents in response to a question. This requires the ability of:

- Building a representation of narrative's content that identifies epidemiological information and keeps the links with the texts. This can be made by information extraction systems that produce fact templates[2,3] or by natural language processing systems that build syntactic-semantic representations of narratives in a canonical form.[4]
- Reasoning about these representations for building the epidemic history, that is the spatiotemporal evolution of outbreak characteristics.

The experience of BIOSTORM on medical databases shows that this reasoning is highly knowledge consuming, requires particularly an explicit modeling of epidemiologic knowledge and the use of temporal and spatial abstractions for epidemic history building.[5] Moreover, lessons learned from the information extraction systems quoted above show that information in outbreak reports is event centered and that recovering the structure of outbreak scenarios is particularly difficult because of complex event structuring, inclusion relationships between events, scattering of events in texts, and information uncertainty.[6]

This paper presents our attempt for representing and reasoning about events that are described in outbreak reports such as emailed by the ProMED global electronic reporting system (http://www.promedmail.org/). After outlining the difficulties in using information from outbreak reports that have guided us, we describe the Spatio-Temporal Extended Event Language (STEEL) that we have developed for knowledge representation. In this extension, we have replaced time with spatiotemporal location and added a notion of spatiotemporal event aggregate for representing complex constructs of events. The last section of this paper succinctly describes an evaluation of its adequacy for representing the outbreak report contents by comparing the performances of this system with human experts in an *ad hoc* query situation with a trial corpus of 35 outbreak reports issued from ProMED-Mail.

USING INFORMATION FROM OUTBREAK REPORTS

Emailed outbreak reports are short news stories, which aim at reporting series of connected events that describe the spread of epidemics of infectious diseases. The term of "event" is used here for all that refers to "actions, events, motions, accomplishments and processes".[7] Figure 1 shows an example of emailed report issued from ProMED-Mail, updating

information about an outbreak of dengue fever in Bangladesh.

Archive Number: 20021008.5493, Published Date: 08-OCT-2002

Bangladesh: Dengue Fever Continues to Spread
A dwindling supply of blood is exacerbating the dengue outbreak in Bangladesh. The total number of dengue-affected patients, according to the official accounts, stood at 4763 as of 16 Sep 2002. Of these, 45 have died so far. The Dengue Control Room sources said 509 persons affected with dengue virus were undergoing treatment at different hospitals across the country.
Meanwhile, it was reported that one person died from dengue hemorrhagic fever (DHF) in Magura district, while 4 others had been hospitalised in Jessore General Hospital in the last 24 hours. The Khulna City Corporation in the meantime has launched an anti-mosquito drive in the city. At least one person died in Chapai-Nawabganj district and 15 others have been hospitalised for DHF. All of them were admitted to the Rajshahi Medical College Hospital (RMCH) during the past 10 days. […]

Figure 1: Example of emailed report issued from ProMED-MAIL

The structure of outbreak reports is complex, intertwining and dispersing descriptive background information with story events throughout the narrative.[7] Three reasons may explain this:

- An emailed report may relate more than one epidemiological event, place or time.
- Its writing is highly influenced by a requirement of brevity.[8] For compactness, the story is crammed into a few complex sentences, complicating the structure of the narrative.
- It is often in a form of an update report relating the evolution of the epidemic characteristics since their last description.

The example in Figure 1 reports, in a single narrative, 10 events (8 related to patients), which concern a total of 6 spatial locations and 4 dates or time intervals, including the publication date. As illustrated, events "interlock and relate to each other in complex ways, forming inclusion relationships".[6] They frequently report aggregates of sub-events in a compact way, like in the sentence "Of them, 45 have died so far". An adequate representation of the events must capture these relationships, especially the sub-event composition.

A preliminary interview of experts in travel and tropical medicine, which are the main users of outbreak reports in our hospital, had reported that using information from outbreak reports requires reconstructing the relationships between events, their underlying temporal and spatial locations, and all descriptive background information that allows orientation in respect to person, place, time and epidemic situation.

THE KNOWLEDGE REPRESENTATION LANGUAGE

STEEL is a typed first-order logic language that is based on the Event Calculus (henceforth EC), which was introduced by Kowalski and Sergot[9] for representing and reasoning about the occurrence of events, the properties that events initiate and terminate, and the maximal validity intervals for which these properties hold uninterruptedly. Amongst the EC's extensions that have been developed in order to enhance its expressiveness, complex patterns of actions have been explored by Cervesato and Montanari,[10] showing the ability of process constructors for packaging up related events, and the problem of event's spatiotemporal location has been addressed by Galton[11] and Bennett.[12] STEEL carries on these works, introducing a joined spatiotemporal location of event, whose properties are used for ruling the building of event aggregates.

1. Language ontology. The basic ontology of STEEL comprises 4 basic types: events, fluents, time stamps, and spatial regions.

An Event corresponds to the performance or occurrence of an action over a specified time. If actions are time independent, defining "certain useful and relevant activities that may be conducted over some time by the agents to accomplish changes of state of the world",[13] events are time dependant. Discourse elements describing events can be identified on the basis of their "dynamic verbs".[7] Events are classified according to a scheme[14] that helps in the project of locating events in time:

- *Occurrence events* are the main event subclass and correspond to the events we want to place on the time axis (e.g. "one person died from dengue").
- *Reporting events* associate the source of information with an occurrence event (e.g. "it was reported that one person died").
- *Attitude events* are similar to reporting events, but they do not guarantee the reality of the information (e.g. "has died from a disease it was feared could be Ebola fever").
- *Aspectual events* that involve aspectual verbs like start, stop, begin etc.

Events may be instantaneous or may happen over a period of time,[15] defining the notion of event duration.

Fluents are valued expressions that describe the properties of system's objects (the value of a quality or a relation), and whose interpretations change from time to time (e.g. "dwindling supply of blood"). The fluent is valid when the object under consideration gets that specific property. Fluents' states are defined

according to events that can initiate or terminate them.

Time stamps. Time is a concept that cannot be easily represented,[16] and several suggestions have been proposed for natural language processing.[17,18,19] Our aim is to represent temporal entities in a convenient manner for inducing the times and ordering of events. In our ontology, time is an ordered set $(T,\prec)$ where elements of T are Shahar's time stamps,[20] which are issued from time expressions encountered in the narratives and can be placed on a time axis (e.g. "16 September 2002"). Formally, this choice allows using event name for time specification, as proposed in the New Event Calculus,[21] or as showed in a study about temporal preposition phrases.[22]

Spatial regions. Space is two-dimensional and corresponds to the set $S=\mathcal{R}x\mathcal{R}$, where $\mathcal{R}$ is the set of reals. A region is a subset of S, usually represented only by a name (e.g. "Bangladesh", "Magura district"). A point is a special kind of region (e.g. "Rajshahi Medical College Hospital"). This choice keeps a level of complexity in accordance with discourse objectives and the way spatial relations are expressed in natural language. As pointed by Asher and Vieu,[23] the mathematical conception of topological space is foreign to space as it is usually expressed in narrative texts, where the reader may use the spatial information contained in texts, even though this information does not contain any system of coordinates.

From the two last types, we define the spatiotemporal location of an event performance as a couple <t,l>, where t is a time expression and l a spatial region.

2. Language description. The basic types are used for defining 3 sets of language basic predicates.

Events and their influences on fluents. This set of modified EC predicates is described in Table 1. In a first approach, we have avoided to deal with locations simultaneously different in time and space.

The relation between t_1 and t_2 in the predicate *happens(e,<t_1,l>,<t_2,l>)* is formalized by the following axiom:

$$happens(e,<t_1,l>,<t_2,l>) \rightarrow t_1 \preceq t_2$$

Table 1: Description and meanings of basic language predicates related to event occurrences and their influences on fluent values.

Predicate	Meaning
$t_1 \prec t_2$	Time stamp t_1 is before time stamp t_2.
$t_1 = t_2$	Time stamp t_1 is equal to time stamp t_2.
happens(e,<t_1,l>,<t_2,l>)	Event e starts at spatiotemporal location <t_1,l> and ends at location <t_2,l> (note: happens(e,<t,l>) $\equiv_{def}$ happens(e, <t,l>,<t,l>)).
initiallyTrue(f,l)	Fluent f holds from time 0 at spatial location l.
initiallyFalse(f,l)	Fluent f does not hold from time 0 at spatial location l.
initiates(e,f,<t,l>)	Fluent f starts to hold after the occurrence of event e at spatiotemporal location <t,l>.
terminates(e,f,<t,l>)	Fluent f ceases to hold after the occurrence of event e at spatiotemporal location <t,l>.
releases(e,f,<t,l>)	Fluent f is no more subject to inertia after the occurrence of event e at spatiotemporal location <t,l> (its value becomes undetermined).

Spatiotemporal relations. Following Hazarika and Cohn's mereotopological theory of space-time,[24,25] spatiotemporal relations between objects can be represented with binary relations based on the notion of connection. Two entities are spatially connected (*sp-connected*) if they share at least a spatial point, though not necessarily simultaneously (e.g. Zaïre that has been renamed as Congo Démocratique). Temporal connection (*t-connected*) of two time intervals is defined by the existence of at least a common temporal point, though not necessarily at the same place. Finally two entities are spatiotemporally connected (*st-connected*) if the closures of these entities share at least a spatiotemporal point. This *α-connected(x,y)* primitive, where $\alpha \in \{st, sp, t\}$, allows defining a set of 10 others mereotopological relations that constitutes the basis of a qualitative representation language (Table 2).

Table 2: Definition of spatio-temporal mereological relations from the primitive *α-connected(x,y)* (where $\alpha \in \{st, sp, t\}$)

Relation	Predicate	Definition
x is disconnected from y	α-disconnected(x,y)	$\neg\alpha$-connected(x,y)
x is a part of y	α-partof(x,y)	$\forall z$. (α-connected(z,x) $\rightarrow$ α-connected(z,y))
x is a proper part of y	α-properpart(x,y)	α-partof(x,y) $\wedge$ $\neg\alpha$-partof(y,x)
x is identical with y	α-equal(x,y)	α-partof(x,y) $\wedge$ α-partof(y,x)
x overlaps y	α-overlap(x,y)	$\forall z$. (α-partof(z,x) $\wedge$ α-partof(z,y))
x is discrete from y	α-discrete(x,y)	$\neg\alpha$-overlap(x,y)
x partially overlaps y	α-partoverlap(x,y)	α-overlap(x,y) $\wedge$ $\neg\alpha$-partof(x,y) $\wedge$ $\neg\alpha$-partof(y,x)
x is externally connected to y	α-externconnected(x,y)	α-connected(x,y) $\wedge$ $\neg\alpha$-overlap(x,y)
x is a tangential proper part of y	α-tangproppart(x,y)	α-properpart(x,y) $\wedge$ $\exists z$. (α- externconnected (z,x) $\wedge$ α- externconnected (z,y))
x is a non tangential proper part of y	α-nontangproppart(x,y)	α-properpart(x,y) $\wedge$ $\neg\exists z$. (α- externconnected (z,x) $\wedge$ α- externconnected (z,y))

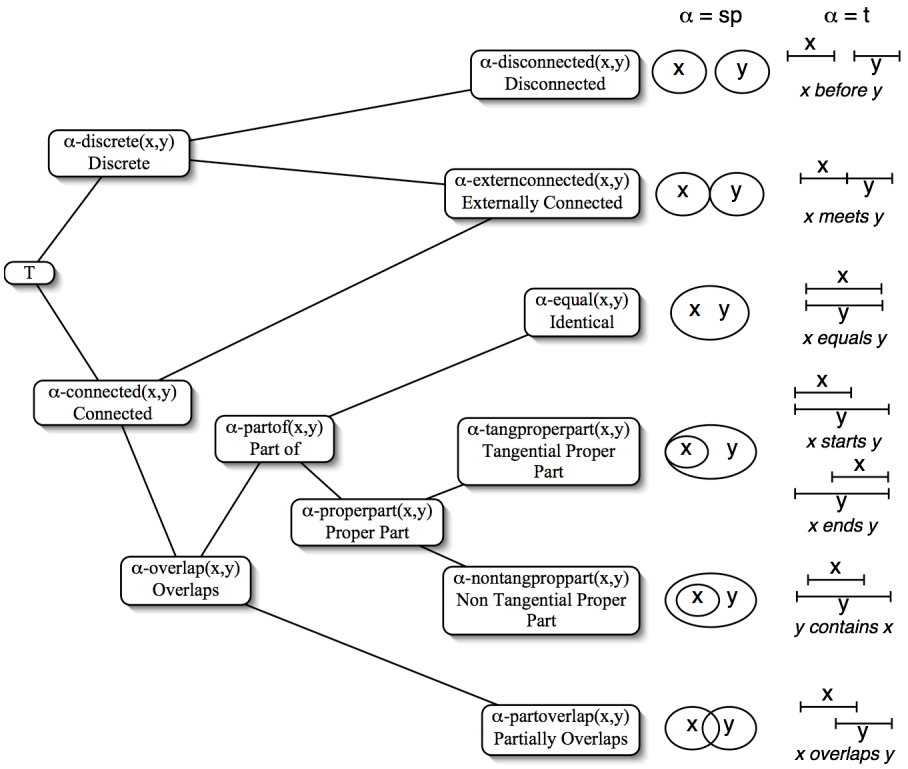


Figure 2: Subsumption lattice of basic mereotopological relations and graphical representation of their semantics for spatial and temporal domains

Figure 2 presents graphically these relations, their semantics for the (sp-) and temporal (t-) relations, the names given by Allen[17] in the last case, and shows their subsumption lattice. We have not figured the names of reciprocal relations. The six terminal relations in the lattice are provably Jointly Exhaustive and Pairwise Disjoint. Experimental results concerning the cognitive adequacy of the interval relations for spatial representation and reasoning showed that people use ordinal information similar to this calculus when representing and remembering spatial arrangements.[26]

The spatio-temporal connection of two entities can then be easily defined by the simultaneous existence of a member of the graph subtree of Figure 2:

st-connected(x,y) → t-connected(x,y) ∧ sp-connected(x,y)

These mereological relations allows introducing an axiom of spatial persistence, stating that an event happens in a region if it happens in a part of this region (e.g. if there is a case of Ebola fever in Gabon then there is a case of Ebola fever in Africa):

happens(e,<t_1,l_2>,<t_2,l_2>) ← happens(e,<t_1,l_1>, <t_2,l_1>) ∧ sp-partof(l_1,l_2)

Macro-events. Capturing the sub-event compositions that are reported in outbreak reports requires more than just representing the hierarchical relationships between events mentioned above in section 2. Representing complex pattern of events involves additional relations among events, such as sequentiality, simultaneity, iteration, or temporal delays between events. Cervesato and Montanari introduced macro-events, which are expressions defined by applying the following grammar,[10] where m is a macro-event, d and D are time expressions with d < D, and e is an event:

m ::= e (basic event)
| m_1 ;$_d^D$ m_2 (sequence with delay d to D)
| m_1 + m_2 (alternative)
| m_1 || m_2 (parallelism)
| m^n (n-time iteration)

We consider that a macro-event is an occurrence of a structured conglomeration of events, and is a direct subconcept of *Event*.

Each macro-event instance is defined by an instance of a macro-event structure (MES), where S is a formula obtained by applying recursively the grammar.

A resulting MES is a tree in which the leaves are event subconcepts. A MES can be used for defining subclasses of events, and if m is a macro-event and MES_m its structure, then $MEClass_m \doteq MacroEvent \sqcap MES_m$.

Table 3 presents how we have expressed these macro-events in predicate form. These rules give a

Table 3: Definition of macro-events in first order logic.

Event	Predicate	Definition
$m_1 ;_d^D m_2$	seqevent(m_1,m_2,d,D)	happens(m,<t_1,l>,<t_2,l>) ∧ meventdef(m, seqevent(m_1,m_2,d,D)) ↔ ∃ t_3, t_4 . (happens(m_1,<t_1,l>,<t_3,l>) ∧ happens(m_2,<t_4,l>,<t_2,l>) ∧ t_3+d ≤ t_4 ≤ t_3+D)
$m_1 + m_2$	altevent(m_1,m_2)	happens(m,<t_1,l>,<t_2,l>) ∧ meventdef(m, altevent(m_1,m_2)) ↔ happens(m_1,<t_1,l>,<t_2,l>) ∨ happens(m_2,<t_1,l>,<t_2,l>)
$m_1 \parallel m_2$	parevent(m_1,m_2)	happens(m, <t_1,l>,<t_2,l>) ∧ meventdef(m, parevent(m_1,m_2)) ↔ ∃ t_3, t_4, t_5, t_6 . (happens(m_1, <t_3,l>,<t_4,l>) ∧ happens(m_2, <t_5,l>,<t_6,l>) ∧ t_1=min(t_3,t_5) ∧ t_2=max(t_4,t_6))
m^n	iterevent(m,n)	happens(m, <t_1,l>,<t_2,l>) ∧ meventdef(m, iterevent(E,n)) ↔ ∃ t_3, t_4 . (happens(m_1, <t_1,l>,<t_3,l>) ∧ happens(m_2, <t_4,l>,<t_2,l>) ∧ meventdef(m_1, iterevent(E,n-1)) ∧ E(m_2) ∧ t_3 ≤ t_4)

logical framework for building the representation of complex events with coherent time boundaries The relation between a macro-event m and its MES is given by a predicate *meventdef(m, MES_m)*.
Macro-events can substitute plain events in the EC defined above, in particular in the predicates *initiates* and *terminates*, allowing properties to be started and ended by generic macro-events.

3. Continuity reasoning toolbox. History of epidemic spread must be reconstructed from the spatiotemporal connections of events and fluents that are described through partial observations reported in emailed texts. This task needs in particular to capture the notions of spatiotemporal continuity of fluents for determining the maximal validity intervals (henceforth MVI) and the structures of macro-events described over several texts (evolution of the number of deaths, of new cases…).

Table 4: Description and meanings of basic language predicates related to event influence on fluent value persistence.

Predicate	Meaning
clipped(<t_1,l>,f, <t_2,l>)	Fluent f is terminated between time t_1 and time t_2 at spatial location l.
declipped(<t_1,l>,f, <t_2,l>)	Fluent f is initiated between time t_1 and time t_2 at spatial location l.
holdsAt(f,<t,l>)	Fluent f holds at spatiotemporal location <t,l>.

Fluent persistence. Events' influences on the persistence of fluent values with respect to time and spatial location are expressed with a set of 6 axioms describing the semantics of 3 basic predicates that are presented in Table 4:

clipped(<t_1,l_1>,f,<t_4,l_1>) ↔ ∃ e,t_2,t_3,t_5,l_2 . happens(e,<t_2,l_2>,<t_3,l_2>) ∧ (terminates(e,f,<t_5,l_1>) ∨ releases(e,f,<t_5,l_1>)) ∧ $t_2 \prec t_5 \prec t_3$ ∧ $t_1 \prec t_5 \prec t_4$ ∧ sp-partof(l_1,l_2)

declipped(<t_1,l_1>,f,<t_4,l_1>) ↔ ∃ e,t_2,t_3,t_5,l_2 . happens(e,<t_2,l_2>,<t_3,l_2>) ∧ (initiates(e,f,<t_5,l_1>) ∨ releases(e,f,<t_5,l_1>)) ∧ $t_2 \prec t_5 \prec t_3$ ∧ $t_1 \prec t_5 \prec t_4$ ∧ sp-partof(l_1,l_2)

holdsAt(f,<t,l>) ← initiallyTrue(f,l) ∧ ¬clipped(<0,l>,f,<t,l>)

holdsAt(f,<t,l_1>) ← ∃ e,t_1,t_2,t_3,l_2 . happens(e,<t_1,l_2>,<t_2,l_2>) ∧ initiates(e,f,<t_3,l_1>) ∧ $t_1 \prec t_3 \prec t_2$ ∧ $t_3 \prec t$ ∧ ¬clipped(<t_3,l_1>,f,<t,l_1>) ∧ sp-partof(l_1,l_2)

¬holdsAt(f,<t,l>) ← initiallyFalse(f,l) ∧ ¬declipped(<0,l>,f,<t,l>)

¬holdsAt(f,<t,l_1>) ← ∃ e,t_1,t_2,t_3,l_2 . happens(e,<t_1,l_2>,<t_2,l_2>) ∧ terminates(e,f,<t_3, l_1>)∧ $t_1 \prec t_3 \prec t_2$ ∧ $t_3 \prec t$ ∧ ¬declipped(<t_3, l_1>,f,<t, l_1>) ∧ sp-partof(l_1,l_2)

Event aggregation and spatiotemporal continuity. For the purpose of constructing spatiotemporally located event aggregates, we introduced a constructor, written $\bigwedge_h$, which combines the events of two *happens* predicates for building the structure of the resulting macro-event, depending on their spatiotemporal and ontological relationships.
Let e_1 and e_2 be two instances respectively of $E_1 \sqsubseteq$ *Event* and $E_2 \sqsubseteq$ *Event*. *Happens(e_1,<t_1,l_1>,<t_1',l_1>)* and *happens(e_2,<t_2,l_2>,<t_2',l_2>)* are their representations in the knowledge base. The time interval during which each event occurs can be respectively defined as [t_1,t_1']=d_1 and [t_2,t_2']=d_2.
Two main cases of constructor's behavior must be considered depending on the ontological relationships between the two events.
If:

- E_1=E_2 with disjoint instances, that is:
  - The instances are differents: $e_1 \neq e_2$
  - Or, in the case of macro-events, the interpretations of events' classes are disjoint: $MEClass_{e1}^{\mathcal{I}} \cap MEClass_{e2}^{\mathcal{I}} = \emptyset$
- Or E_2=Macroevent ⊓ iterevent(E_1,*)

Then the instances are related to a same event and the macro-event constructor proceeds to its iteration.
In all other cases, the structure of the resulting event involves sequentiality or parallelism, depending on the st-relationships between the events. Table 5 summarizes the constructor's results.

Table 5: Results of the macro-event constructor $\wedge$h(happens(e1,<t1,l1>,<t1',l1>), happens(e2,<t2,l2>,<t2',l2>))

Case: E_1=E_2=E or E_2=Macroevent $\sqcap$ iterevent(E_1=E,*)	
sp relationships	*Results*
sp-partof(l',l) with (l'=$l_1 \wedge l=l_2$) or (l'=$l_2 \wedge l=l_1$)	happens(m,<t,l>,<t',l>) $\wedge$ meventdef(m,iterevent(E,*)) $\wedge$ t=min(t_1,t_2) $\wedge$ t'=max(t_1',t_2')
sp-partoverlap (l_1,l_2) $\vee$ sp-discrete(l_1,l_2)	happens(m,<t,l>,<t',l>) $\wedge$ meventdef(m,iterevent(E,*)) $\wedge$ t=min(t_1,t_2) $\wedge$ t'=max(t_1',t_2') $\wedge$ l=($l_1 \cup l_2$)
Otherwise	
sp relationship: case sp-equal(l_1,l_2)	
t relationships	*Results*
t-discrete (d_1,d_2)	happens(m,<t_1,l>,<t_2',l>) $\wedge$ meventdef(m,seqevent(e_1,e_2,d,d)) $\wedge$ d=t_2-t_1' $\wedge$ l_1=l_2=l
equals(d_1,d_2)	happens(m,<t,l>,<t',l>) $\wedge$ meventdef(m,parevent(e_1,e_2)) $\wedge$ t_1=t_2=t $\wedge$ t_1'=t_2'=t' $\wedge$ l_1=l_2=l
t-properpart(d_1,d_2)	happens(m,<t_2,l>,<t_2',l>) $\wedge$ meventdef(m, parevent(e_1,e_2)) $\wedge$ l_1=l_2=l
overlaps(d_1,d_2)	happens(m,<t,l>,<t',l>) $\wedge$ meventdef(m, parevent(e_1,e_2)) $\wedge$ t=min(t_1,t_2) $\wedge$ t'=max(t_1',t_2') $\wedge$ l_1=l_2=l
sp relationship: case sp-discrete(l_1,l_2) $\vee$ sp-partoverlap(l_1,l_2)	
t relationships	*Results*
t-discrete (d_1,d_2)	happens(m,<t_1,l>,<t_2',l>) $\wedge$ meventdef(m,seqevent(e_1,e_2,d,d)) $\wedge$ d=t_2-t_1' $\wedge$ l=($l_1 \cup l_2$)
equals(d_1,d_2)	happens(m,<t,l>,<t',l>) $\wedge$ meventdef(m,parevent(e_1,e_2)) $\wedge$ t_1=t_2=t $\wedge$ t_1'=t_2'=t' $\wedge$ l=($l_1 \cup l_2$)
t-properpart(d_1,d_2)	happens(m,<t_2,l>,<t_2',l>) $\wedge$ meventdef(m, parevent(e_1,e_2)) $\wedge$ l=($l_1 \cup l_2$)
overlaps(d_1,d_2)	happens(m,<t,l>,<t',l>) $\wedge$ meventdef(m, parevent(e_1,e_2)) $\wedge$ t=min(t_1,t_2) $\wedge$ t'=max(t_1',t_2') $\wedge$ l=($l_1 \cup l_2$)
sp relationship: case sp-properpart(l_1,l_2)	
t relationships	*Results*
t-discrete (d_1,d_2)	happens(m,<t_1,l>,<t_2',l>) $\wedge$ meventdef(m,seqevent(e_1,e_2,d,d)) $\wedge$ d=t_2-t_1'
equals(d_1,d_2)	happens(m,<t,l>,<t',l>) $\wedge$ meventdef(m,parevent(e_1,e_2)) $\wedge$ t_1=t_2=t $\wedge$ t_1'=t_2'=t'
t-properpart(d_1,d_2)	happens(m,<t_2,l>,<t_2',l>) $\wedge$ meventdef(m, parevent(e_1,e_2))
overlaps(d_1,d_2)	happens(m,<t,l>,<t',l>) $\wedge$ meventdef(m, parevent(e_1,e_2)) $\wedge$ t=min(t_1,t_2) $\wedge$ t'=max(t_1',t_2')

Macro-event occurrence and maximum validity intervals of properties. A macro-event m, which structure is MES_m, has occurred over a spatiotemporal interval [<t,l>,<t',l>], written *meo(m,t,t',l)*, iff: $\exists$ t_1, t_2, l_1 . meventdef(m,MES_m) $\wedge$ happens(m,<t_1,l_1>, <t_2,l_1>) $\wedge$ t$\preceq$ $t_1$$\preceq$$t_2$$\preceq$t' $\wedge$ sp-partof(l_1,l). The macro-event occurrence may be not explicitly present in the knowledge base, and determined recursively using the result definitions of the macro-event constructor $\bigwedge_h$.

The MVI of a property or a fluent p, written mvi(p,t,t',l), is the maximal spatiotemporal interval [<t,l>,<t',l>] over which p holds uninterruptedly. This can be written as:

mvi(p,t,t',l) $\leftrightarrow$ t$\prec$t' $\wedge$
(initiallyTrue(p,l) $\vee$ (happens(e_1,<t_1,l_1>,<t_2,l_1>) $\wedge$
initiates(e_1,p,<t,l>) $\wedge$ $t_1$$\prec$t$\prec$$t_2$ $\wedge$ sp-partof(l,l_1)) $\wedge$
happens(e_2,<t_3,l_2>,<t_4,l_2>) $\wedge$
(terminates(e_2,p,<t',l>) $\vee$ releases(e_2,p,<t',l>)) $\wedge$
$t_3$$\prec$t'$\prec$$t_4$ $\wedge$ sp-partof(l,l_2) $\wedge$
$\neg$clipped(<t,l>,p,<t',l>)

STEEL is able to determine the MVI, to check the truth of MVIs or macro-event occurrences, and to process Boolean combinations of MVI and macro-event occurrence verifications.

EVALUATION

As proof of concept, we have studied the adequacy of STEEL for representing outbreak report contents by comparing the performances of an experimental implementation of this language in Prolog with human experts in a query situation.

STEEL was implemented in SWI-Prolog (University of Amsterdam, http://www.swi-prolog.org/). The axioms and definitions from the previous section of this paper were transcribed into prolog rules. The whole language kernel is a module of about 200 rules. The trial knowledge base was built from a trial corpus of 35 emailed outbreak reports issued from the ProMED mail list and describing an outbreak of Ebola fever in Gabon from December 2001 until May 2002. The size of this trial corpus was 8105 words, 213 sentences. From this corpus, 224 events and 328 objects have been extracted with a simple annotation tool built ad hoc, which solicited the annotator for inferring the spatiotemporal locations of events when they were not directly specified in the text. Amongst these events, 148 were macro-events. The annotator responses were either a location or a relational (i.e. precedence, inclusion…) expression involving a location. Figure 3 shows an excerpt of the resulting knowledge base that is about 2000 prolog rules long.

Source

X-ProMED-Id: 20011205.2950 Date: 2001-12-05
Subject: PRO/EDR> Viral hemorrhagic fever, suspected - Gabon
Source: WHO Disease Outbreaks Report, Wed 5 Dec 2001 [edited]
On Tue 4 Dec 2001, WHO received reports of 7 deaths in an outbreak of suspected viral haemorrhagic fever in Ogooue Ivindo Province in the northeastern part of the country.

Representation

```
ist(happens(reports(proMED,system,proMedMail20012950),
[[2001,12,05],_]),system).
…
ist(happens(reports(_,WHO,[proMedMail20012950|1]),[[2001,
12,04],_]),proMedMail20012950).
ist(happens(event1,[time1,'Ogooué Ivindo Province'],
[time2,'Ogooué Ivindo Province']), [proMedMail20012950|1]).
ist(agent(event1,isPossibly('viral haemorrhagic fever',_)),
[proMedMail20012950|1]).
ist(sp-partof(northEastPart('Gabon'),'Ogooué Ivindo Province'),
[proMedMail20012950|1]).
…
instance(event1,macroevent).
…
meventdef(event1,itervent(death,7)).
```

Figure 3: Excerpt of the trial knowledge base.

For the experimentation we have built from the trial corpus a test set of 18 questions covering a spectrum of question types. Two examples of questions follow:

- What is the number of new cases of Ebola fever in Gabon between 2001-12-29 and 2002-1-6?
- What names of cities and villages located in Ogooué-Ivindo Province are mentioned in the outbreak reports?

These questions have been addressed in logic formalism to the system, which found a response in every case. The CPU time required for a response was 273±98 ms on a PowerPC G4 under Darwin/MacOSX at 1.2 Ghz with 512 Ko of L2 cache and a 167 Mhz bus.

Then we gave in a booklet the trial document collection to 3 experts in tropical and/or travel medicine that are usual users of ProMED-Mail. In a first test we have asked them to answer to the same questions. In a second test we have given them the system's responses, without indicating their origin, and asked them if these responses were satisfactory or not.

Each expert took between 75 and 90 minutes for completing the experiment. The expert-expert and expert-system answer accordance in the first test is summarized in Table 6. The number of different answers stands at 3 in all cases, except one case of total concordance between an expert and the system. Testing the homogeneity of accordance distribution shows that experts make no distinction between the system and another expert (Fisher's exact probability test on a 2x6 contingency table, p-value=0.4676). After merging the expert's responses, we found that a total of 4 questions have got at least one expert's response different from the system. We have considered this merging as the maximal discordance, and we have tested it with a binomial sign test on the number of agreements versus the number of discordances. The critical probability P_{H0}(number of agreements ≥ 14)=0.015442 confirms the correctness of the system's responses according to the experts. This conclusion is confirmed by the results of the second part of the experience, which reports that all system's responses are appropriate for all experts.

Table 6: Expert-expert and expert-system accordance / discordance ratios for the trial set of 18 questions.

	System	Expert 2	Expert 3
All experts	4 / 14	-	-
Expert 1	0 / 14	-	-
Expert 2	3 / 15	3 / 15	-
Expert 3	3 / 15	3 / 15	3 / 15

CONCLUDING REMARKS AND FUTURE WORK

In this paper we introduce an extension of the Event Calculus suited to the representation of information extracted from outbreak reports within the framework of automatic following of epidemic spread.

The adequacy of Event Calculus to narratives' representation,[27] the structural characteristics of these texts, especially their centering on spatiotemporally located events, have pushed us to use this formalism rather than those issued from the Situation Calculus,[27,28] or from action-based models.[29,30,31,32,33] In its original version or its extensions, this calculus is unable to easily represent the joint spatiotemporal location of events or the events' aggregation with respect to their occurrence location (required for merging information from several reports). Others models based on chronicles have been developed for representing spatiotemporal situation described in narratives,[34] but in an objective of situation recognition and not of modeling and information summarization as in our case.

Our extension allows a representation that is very close to the narrative, centered on event occurrences and keeping the event relationship network. It also allows to group together several report contents for building a global description of an outbreak occurrence, considering an epidemic as a complex event, which results from the aggregation of the events that are reported (spatiotemporal abstraction mechanism).

However, if this representation seems to be adequate for the purpose of representing information extracted from outbreak reports, its ontology (a modeling language ontology) cannot be considered as those of epidemiology (a domain ontology), although it may

or must be a part of. Further work in this direction must be made to get a system that is able to proceed to an automatic modeling of outbreaks.

The system developed in this work cannot be considered as a complete information extraction/representation system. It only focuses on the final step of a complete natural language processing system that remains to be completed. However, we expect that STEEL would help us solving the problem of location and event identification[6,14,35] during natural language processing of outbreak reports.

ACKNOWLEDGMENTS

This research was supported in part by the French Ministry of Research and Technology (ACI "Télémédecine et Santé" n° 2000/120), by a grant from the Délégation Générale à l'Armement of the French Defense (DGA n° 00.34.030.00.470.75.65) and by the ADIMI Association for Medical Informatics.

References

[1] Woodall J. Official versus unofficial outbreak reporting through the Internet. International Journal of Medical Informatics 1997;47:31-4.

[2] Damianos L, Day D, Hirschman L et al. Real users, real data, real problems: the MiTAP system for monitoring bio events. In: Proceedings of BTR2002: Unified Science & Technology for Reducing Biological Threats & Countering Terrorism; The University of New Mexico, Albuquerque, New Mexico; March 2002.

[3] Grishman R, Huttunen S, Yangarber R. Information extraction for enhanced access to disease outbreak reports. Journal of Biomedical Informatics 2002;35:236–46.

[4] Roux M, Ledoray V. Understanding of medico-technical reports. Artificial Intelligence in Medicine 2000;18:149–72.

[5] Buckeridge DL, Graham J, O'Connor MJ, Choy MK, Tu SW, Musen MA. Knowledge-based bioterrorism surveillance. Proc AMIA Symp. 2002;:76-80.

[6] Huttunen S, Yangarber R, Grishman R. Diversity of scenarios in information extraction. In: Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC 2002); Las Palmas, Canary Islands, Spain; May 2002.

[7] Schokkenbroek C. News Stories - Structure, time and evaluation. Time and Society 1999; 8(1):59-98.

[8] Bell A. The discourse structure of news stories. In: A Bell and P Garrett, editors. Approches to Media Discourse. Oxford: Blackwell; 1998. p. 64-104.

[9] Kowalski R, Sergot M. A logic-based calculus of events. New Generation Computing 1986;4:67–95.

[10] Cervesato I, Montanari A. A calculus of macro-events: progress report. In: A Trudel, SD Goodwin (eds), 7th International Workshop on Temporal Representation and Reasoning (TIME'00); Cape Breton, Nova Scotia, Canada; 7-9 July 2000. IEEE Computer Society Press; 2000. p. 47-58.

[11] Galton A. Toward an integrated logic of space, time, and motion. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93); Chambéry, France; 28 August-3 September 1993. p. 1550-5.

[12] Bennett B. Space, time, matter and things. In: Proceedings of the international conference on Formal Ontology in Information Systems (FOIS'2001); Ogunquit, Maine, USA. New York, NY, USA: ACM Press; 2001. p. 105-16.

[13] Knight B, Peng T, Ma J. Reasoning about change over time: actions, events, and their effects. In: The Fourth Symposium on Logical Formalizations of Commonsense Reasoning (CS98); Queen Mary and Westfield College, London, UK; 7-9 January 1998. p. 183-97.

[14] Setzer A, Gaizauskas R. Annotating events and temporal information in newswire texts. In: Proceedings of the Second International Conference On Language Resources And Evaluation (LREC-2000); Athens, Greece; 31 May- 2 June 2000. p. 1287-93.

[15] Shanahan M. Solving the frame problem. Cambridge, Massachussets: MIT Press; 1997.

[16] Combi C, Shahar Y. Temporal reasoning and temporal data maintenance in medicine: issues and challenges. Computers in Biology and Medicine 1997;27(5):353-68.

[17] Allen JF. Towards a general theory of action and time. Artificial Intelligence 1984;23:123-54.

[18] Galton A. A critical examination of Allen's theory of action and time. Artificial Intelligence 1990;42:159-88

[19] Steedman M. Temporality. In: J Van Bentham and A ter Meulen, editors, Handbook of logic and language. Amsterdam: Elsevier;1997. p. 895-938.

[20] Shahar Y. A framework for knowledge-based temporal abstraction. Artificial Intelligence 1997;90:79-133.

[21] Sadri F, Kowalski R. Variants of the Event Calculus. In: L Stirling Ed),Proceedings of the International Conference on Logic Programming,

Kanagawa, Japan, June 1995, The MIT Press, p. 67-81.

[22] Pratt I, Francez N. Temporal prepositions and temporal generalized quantifiers. Linguistic and Philosophy 2001;24(2):187-222.

[23] Asher N, Vieu L. Toward a geometry of common sense: a semantic and complete axiomatization of mereotopology. In: Proceedings of IJCAI'95; Montréal, Québec, Canada; 20-25 August 1995. p. 846-52.

[24] Hazarika SM, Cohn AG. Qualitative spatio-temporal continuity. In: D Montello, editors. Proceedings of COSIT'01; Morro Bay, California, USA; September 2001. Volume 2205 of Lecture Notes in Computer Sciences. Berlin: Springer-Verlag; 2001. p. 92-107.

[25] Cohn AG, Hazarika SM. Continuous transitions in mereotopology. In: Commonsense-2001: 5th Symposium on Logical Formalizations of Commonsense reasoning; New York, USA; 2001. p. 71-80.

[26] Knauff M. The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. Spatial Cognition and Computation 1999;1:261-90.

[27] Cervesato I, Franceschet M, Montanari A. A guided tour through some extensions of the event calculus. Computational Intelligence 2000;16(2):307-47.

[28] Miller R, Shanahan M. Narratives in the Situation Calculus. Journal of Logic and Computation 1994; 4(5 - Special Issue on Actions and Processes):513-30.

[29] Miller R. Situation Calculus specifications for Event Calculus logic pograms. In: Proceedings of the 3rd International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR'95), 1995, Springer Verlag (LNAI 928), p. 217-30.

[30] Kakas A, Miller R. A simple declarative language for describing narratives with actions. Journal of Logic Programming 1997;31 (1-3) (Special Issue on Reasoning about Action and Change):157-200.

[31] Kakas AC, Miller R, Toni F. ε-RES : reasoning about actions, events and observations. In : T. Eiter, W. Faber, M. Truszczynski (Eds), 6th Int. Conf. LPNMR 2001, Vienna, Austria, 17-19 September 17-19 2001, p. 254-66.

[32] Baral C, Gelfond M. Reasoning about effects of concurrent actions. Journal of Logic Programming 1997; 31:85-118.

[33] Baral C, Son TC, Tuan L. A transition function based characterization of actions with delayed and continuous effects. In: D Fensel, F Giunchiglia, DL McGuinness, MAe Williams, Proceedings of the Eights International Conference on Principles and Knowledge Representation and Reasoning (KR-02), Toulouse, France, 22-25 April 2002, p. 291-302.

[34] Sansonnet JP, Gerard S. A spatio-temporal model for the representation of situations described in narrative texts. In: D Christodoulakis (eds), Natural Language Processing - NLP 2000, Second International Conference, Patras, Greece, 2-4 June 2000, Springer, Lecture Notes in Computer Science 1835, p. 176-84.

[35] Filatova E, Hovy E. Assigning Time-Stamps to Event-Clauses. In: Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing, ACL-2001, Toulouse, France, 6-11 July, p. 88-95

Using Semantic Dependencies for Consistency Management of an Ontology of Brain-Cortex Anatomy

Olivier Dameron [a,b,c] Bernard Gibaud [b] Mark Musen [a]

[a]Stanford Medical Informatics, Stanford University, USA (http://smi.stanford.edu)
[b]Laboratoire IDM, Université de Rennes, France (http://idm.univ-rennes1.fr)
[c]INRIA, France (http://www.inria.fr)

Abstract

In the context of the Semantic Web, ontologies have to be usable by software agents as well as by humans. Therefore, they must meet explicit representation and consistency requirements. This article describes a method for managing the semantic consistency of an ontology of brain-cortex anatomy. The methodology relies on the explicit identification of the relationship properties and of the dependencies that might exist among concepts or relationships. These dependencies have to be respected for insuring the semantic consistency of the model. We propose a method for automatically generating all the dependent items. As a consequence, knowledge base updates are easier and safer.

Our approach is composed of three main steps: (1) providing a realistic representation, (2) ensuring the intrinsic consistency of the model and (3) checking its incremental consistency. The corner stone of ontological modeling lies in the expressiveness of the model and in the sound principles that structure it. This part defines the ideal possibilities of the ontology and is called realism of representation*. Regardless of how well a model represents reality, the* intrinsic consistency *of a model corresponds to its lack of contradiction. This step is particularly important as soon as dependencies between relationships or concepts have to be fulfilled. Eventually, the* incremental consistency *encompasses the respect of the two previous criteria during the successive updates of the ontology.*

Keywords*: Ontology, consistency, dependencies*

Context

A symbolic model allows developers to represent general knowledge about a domain and the meaning that is commonly associated with it. This knowledge can be used by itself (e.g., for teaching), or indirectly as a reference to process specific facts (e.g., to assist queries or data retrieval). In the latter case, symbolic models are perceived as a key feature to provide software assistance for tasks that now require domain-aware intervention by a human. As interoperability of these software applications is desirable, shared conceptual models, and specifically ontologies, play a major role in a Semantic Web context [1]. Since these models are to be usable by software, they must meet explicit representation and consistency requirements.

For medical applications, anatomy provides a common reference used to reason about pathology or localization of functional activity [2, 3]. The Foundation Model of Anatomy (FMA) [4] and Galen [5] are two major conceptual models that provide a symbolic representation of human anatomy. However, neither of them provides a satisfactory representation of brain-cortex anatomy. The major sources of neuroanatomical knowledge are paper-based atlases [6, 7] and terminological systems such as Neuronames [8].

We are working on an ontology of brain-cortex anatomy. Our goal is more to formalize existing knowledge than it is to propose new anatomical concepts or relationships. Our model has been described in previous publications [9, 10]. It comprises 304 concepts and 1254 relationships that represent the organization of anatomical structures. Because the brain surface presents complicated folding patterns, typical anatomical structures are gyri (the bulges of cerebral matter, similar to hills), the sulci (the hollow foldings, similar to valleys) and lobes (sets of gyri).

Our model's taxonomy hierarchy is composed of three levels. First, the generic level contains concepts such as `Lobe` or `Sulcus`, and is mainly used to define the domain and range of the relationships. Second, the abstract level represents a prototypical brain hemisphere, and contains concepts such as `Frontal Lobe` or `Central Sulcus`. Third, the lateralized level is used to represent left/right asymmetries, and contains concepts such as `Left Frontal lobe`.

For mereology, the model identifies several relationships such as *`hasDirectAnatomicalPart`*, *`hasAnatomicalPart`*, *`hasSegment`* and their properties, inspired from previous theoretical works [11, 12].

The model also represents neighborhood relationships such as the separation of two cortical structures by a sulcus, anatomical continuity, and sulci connection.

In this context, managing the semantic consistency of the ontology has been one of our main concerns.

This work has consisted in checking that the model reflects reality, and that the relationship properties are respected. The last point has led us to identify dependencies among relationships. This article describes some of these dependencies and proposes an original method for making sure that they are respected during successive updates. This method consists of automatically generating all the dependent relationships, which also make the knowledge base maintenance easier.
The "Realism of representation" section describes our efforts to keep our model as close as possible to reality. The "Intrinsic consistency" section describes the identification of dependencies among relationships, and their representation by implication rules which can be used to generate a self-consistency of a version of the ontology. The "Incremental consistency" section describes how to make sure that the successive updates are all self-consistent and provide the expected modifications.

Realism of representation

The adequacy of the model with respect to some reality is a core aspect of ontological modeling. It ensures that the definitions and the propositions derived from the model are acceptable. Since reality is hard to define and can be relative, canonical knowledge [4], (*i.e.*, derived from generalization and synthesis of previous observations) provides at least a gold standard [13]. For our work on brain anatomy, reference atlases [6, 7] and discussion with an expert provided the base of the canonical knowledge.
The correspondence between the concepts of a symbolic model and the elements of some reality is achieved through an *interpretation function* [14], which maps every concept to its individual instances. The structure of the model defines the possible interpretation functions. A lax model would allow interpretation functions that associate concepts that do not match generally admitted knowledge to concrete situations. Conversely, a restrictive model would dismiss the interpretation functions that associate desired concepts to concrete situations. This section describes how we tried to make our symbolic model as restrictive as possible with regard to anatomical variability.

All reality An ontology is a simplified view of some reality. However, an ontology has to comply with all the situations of the domain of study. For instance, our model of brain anatomy has to cope with a precentral sulcus composed of two segments for one individual, as well as a precentral sulcus composed of four segments for another one.
For our brain anatomy ontology, the main difficulties lay in left/right asymmetries between the two hemispheres, as well as in inter-individual variability. The acknowledgment of this variability and its explicit representation in our model is particularly apparent in part/whole as well as in topological relationships, where a distinction has to be made between mandatory and possible relationships. Necessary conditions are represented by the existential operator ($\exists$). Possible conditions are represented by the universal operator ($\forall$). For instance, "the precentral sulcus (precs) must have a superior segment (sup-precs) and an inferior segment (inf-precs), and can have an intermediate (int-precs) and a marginal segment (marg-precs)" is represented by "all the segments of precs are sup-precs or inf-precs or int-precs or marg-precs; and there is a sup-precs; and there exists an inf-precs". In addition, existence probabilities for concepts as well as for relationships are specified whenever possible[1].
Modeling all reality is pretty easy by reducing the constraints. Therefore, lax models are favored here.

Only reality Ideally, an ontology must not allow developers to describe things other than those in the reality being modeled. A model of anatomy that would allow a brain hemisphere to have any number of lobes, or two frontal lobes, cannot be considered as a good model. Therefore, the model has to enforce enough constraints in order to reject any bad interpretation of the reality. We took this point into account for specialization, composition and topological relationships.
In the taxonomic hierarchy, the distinction between the generic, abstract, and lateralized levels, as well as the consideration that the concepts of a same level are mutually exclusive (e.g., a lobe cannot be both a frontal lobe and a parietal lobe) conform to this principle.
For mereological relationships, both the cardinality constraints and the partitioning principle that requires that anatomical structures have no common part also play important roles. For instance, we do not simply state that "a hemisphere is composed of five lobes; frontal lobe is a lobe; parietal lobe is a lobe; temporal lobe is a lobe, occipital lobe is a lobe and limbic lobe is a lobe", as most symbolic models of anatomy would do. We stated that a hemisphere has five direct anatomical parts that include exactly one frontal lobe, exactly one parietal lobe, etc.; these five lobes are mereologically mutually disjoint.
For topological relationships, representation using binary relationships that a sulcus separates two cortical structures, just as a river separates two regions, could lead to erroneous inferences. Figure 1 illustrates such situations. If we use a binary relationship to represent that a sulcus `S` is a boundary of a cortical structure

[1] Mainly from Ono's Atlas [6].

(e.g., G_1) as shown in the middle column, then we are unable to infer correctly that S separates G_1 from G_3 but not from G_2. The bottom of Figure 1 shows another typical situation where some erroneous separations cannot be ruled out. Therefore, we had to use a ternary *separates* relationships (right column of Figure 1).

Intrinsic consistency

There are important dependencies among the relationships in our model of brain anatomy. The various dependencies we could identify are described in the "Dependencies between relationships" subsection. These dependencies can be seen as consequences of the properties of the relationships.
These dependencies could be modeled by implication rules. Examples of such rules are provided in the "Consistency rules" subsection.

Dependencies between relationships

Specialization dependencies Specialization-related dependencies occur between a general concept and a more specific one. Such dependencies are similar to those of inheritance for object-oriented models. Although very simple, such dependencies still have to be taken into account.
In our model of anatomy, specialization dependencies occur between the three taxonomic levels of our model. For instance, a Sulcus (generic level) is filled with cerebro-spinal fluid. Therefore, the Central Sulcus (abstract level) which is subsumed by Sulcus, is also filled with cerebro-spinal fluid, and so are the Left Central Sulcus and the Right Central Sulcus (lateralized level).
Specialization dependencies can also take place between relationships. Thus, the existence of the *hasDirectAnatomicalPart* relationship between two anatomical structures implies that they are also linked by the broader *hasAnatomicalPart* relationship. Similarly, if a Sulcus *isBranchOf* another one, both of them also have to be linked by the *isConnectedTo* relationship.

Dependencies between mereological relationships The dependencies between part/whole relationships are mainly consequences of the taxonomy of mereological relationships (Fig 2) and of the transitive property of some of them. Particularly, the *isDirect...* relationships are non-transitive sub-relationships of transitive ones. This is a standard practice both in programming and in knowledge representation. For example, the Orbital Pars of Inferior Frontal Gyrus *isDirectAnatomicalPartOf* Inferior Frontal Gyrus. *isDirectAnatomicalPartOf* is a sub-relation of *isAnatomicalPartOf*. Therefore, the latter also holds between the two cortical structures. Similarly, Inferior Frontal Gyrus *isDirectAnatomicalPartOf* Frontal Lobe. It follows that Inferior Frontal Gyrus *isAnatomicalPartOf* Frontal Lobe. As the *isAnatomicalPartOf* relationship is considered to be transitive (whereas *isDirectAnatomicalPartOf* isn't), it also must hold between Orbital Pars of Inferior Frontal Gyrus and Frontal Lobe.
The spatial extensions of anatomical structures constitute another example of dependencies (Fig 3). Indeed, there is a mereological hierarchy between the spatial extensions of an anatomical structure (Fig 4). This hierarchy combines with the mereological hierarchy of anatomical structures, as mereological relationships between anatomical structures implies mereological relationships between their spatial extensions (Fig 5). For instance, the VisibleCorticalZone of a cortical anatomical structure *isSubAreaOf* the ExtendedCorticalZone of the same structure. This is true for the PreCentral Gyrus as well as for the Frontal Lobe. But since the former *isAnatomicalPartOf* the latter, the VisibleCorticalZone (respectively ExtendedCorticalZone) of PreCentral Gyrus *isSubAreaOf* the VisibleCorticalZone (respectively ExtendedCorticalZone) of Frontal Lobe. This example shows that dependencies can occur between relationships such as *isVisiblePartOf* and *isAnatomicalPartOf* that are not sub-relations of each other.

Dependencies between topological relationships The dependencies between topological relationships are mainly due to the taxonomy of these relationships. For instance, if a sulcus separates two cortical structures, then these structures also have to be contiguous. The duality between the configuration of the sulci and that of the gyri is another example of dependencies. However, these dependencies are hard to model and have not yet been taken into consideration.

Combined dependencies Of course, it is also possible to combine the three previous kinds of dependencies which makes it harder to categorize them. These combinations are particularly interesting because they involve dependency patterns that are more complex than simple sub-relations. For instance, if

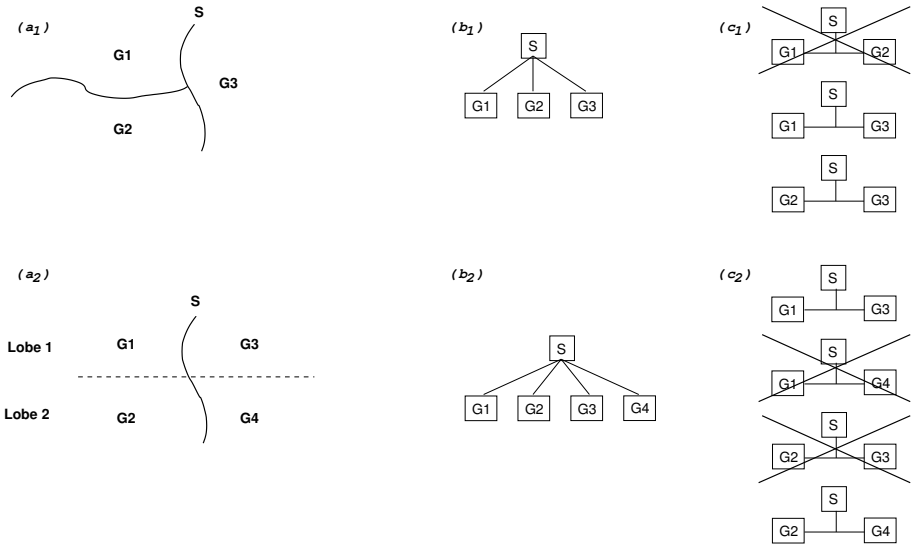


Figure 1: Example of situations where a ternary relationship is necessary to infer if two cortical structures are separated by a sulcus. (a_1) and (a_2) illustrate two configurations involving a sulcus S and some gyri $G_1 ... G_n$. (b_1) and (b_2) model the corresponding separation relationships by binary relationships. (c_1) and (c_2) show all the separation relationships that are inferred from (b). The erroneous ones, such as S separates G_1 and G_2 for (c_1), are crossed out.

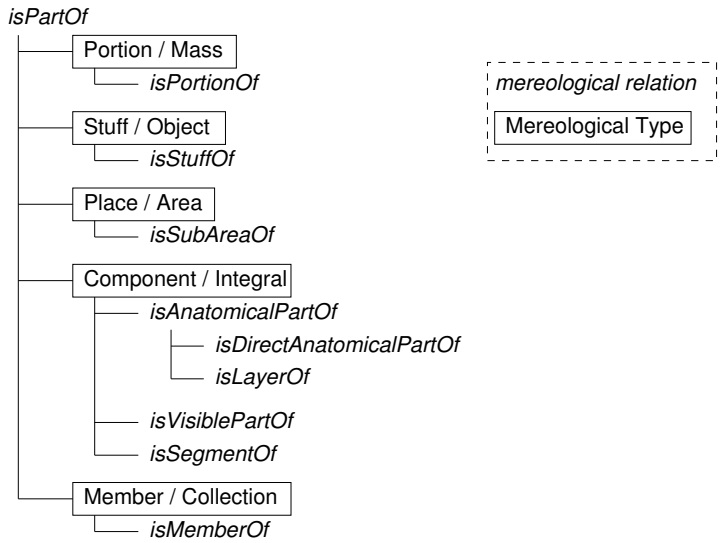


Figure 2: Taxonomical hierarchy of the mereological relationships used for brain-cortex anatomy. The mereological types have been identified by theoretical works on mereology [11] and have different properties.

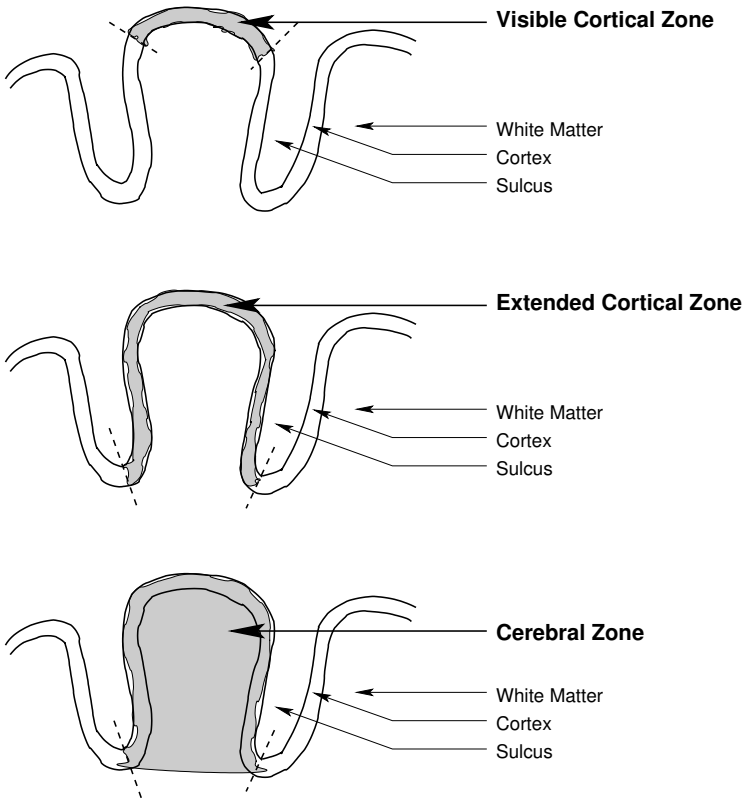


Figure 3: Possible spatial extensions of a gyrus. The cerebral zone is referred to in surgical procedure. The extended cortical zone is referred to in functional activity studies. The visible cortical zone is referred to when teaching neuroanatomy.

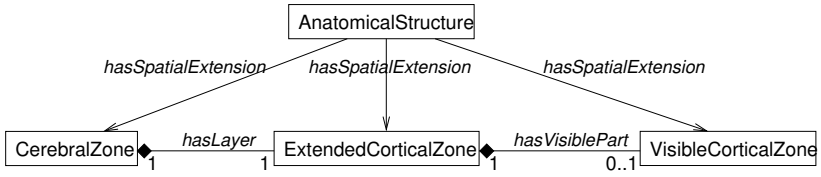


Figure 4: Mereological relationships among the spatial extensions of an anatomical structure.

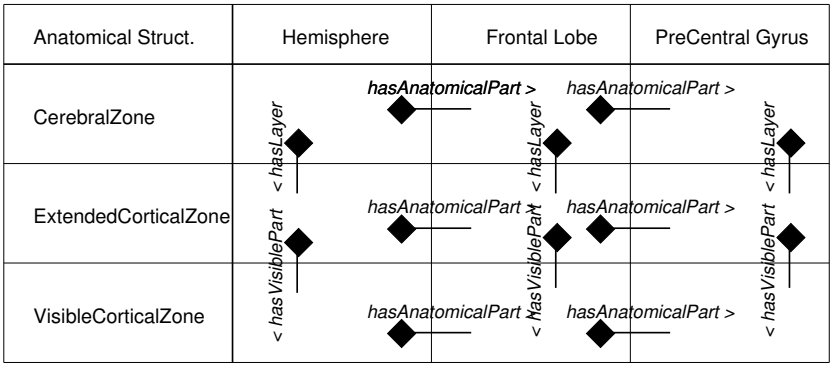


Figure 5: Dependencies between the mereological hierarchy of anatomical structures and that of their spatial extensions. The compound is on the diamond side of mereological relationships

two cortical structures are separated by a *segment* of a sulcus, they are also separated by this sulcus (e.g., `PreCentral Gyrus` and `Superior Frontal Gyrus` are separated by `Superior PreCentral Sulcus`; therefore, they are also separated by `PreCentral Sulcus`), thus combining mereological and topological dependencies.
Another example of dependency combining both mereological and topological relationships also involves relationship properties. Two cortical structures are anatomically continuous if and only if their visible parts are externally connected. If one of the two cortical structures is an anatomical part of a whole, but the other is not a part of this whole, then the visible part of the whole can be proved to be externally connected to the visible part of the second anatomical structure. Figure 6a shows a schema of such a dependency. For example, `PreCentral Gyrus` and `Opercular pars of Inferior frontal Gyrus` are anatomically continuous. Since the `Opercular pars` is an anatomical part of `Inferior Frontal Gyrus` and since `Inferior Frontal Gyrus` and `PreCentral Gyrus` are mereologically disjoint (they do not have any common part), they also have to be anatomically continuous. In addition, this inferred relationship can be used iteratively to apply the same principle (which is equivalent to using the transitive property of *`isAnatomicalPartOf`*). Figure 6b to 6d illustrate the successive application of this principle.
This approach can be extended to anatomical contiguity or the separation of two cortical structures by a sulcus. Thus, the fact that `Central Sulcus` *`separates`* `Frontal Lobe` and `Parietal Lobe` can be seen as a consequence of the fact that `Central Sulcus` separates `PreCentral Gyrus` (a part of `Frontal Lobe`) and `PostCentral Gyrus` (a part of `Parietal Lobe`).
Finally, specializing abstract level concepts into lateralized concepts also generates dependencies.

Consistency rules

The previous dependencies can be represented as implication rules. Such rules, along with the relationship properties, constitute knowledge about anatomical knowledge. They belong to a level separate from that concepts and relationships.

The implications can form the basis of an inference engine that automatically generates all the dependent concepts and relationships.
We maintain only an *abstract restricted model* composed of :

- the concepts of the abstract level (i.e., non lateralized, such as `Central Sulcus` and `Frontal Lobe`);
- all the independent relationships
- a restricted base of asymmetry-specific facts, such as the different existence probabilities for the left and right intermediate precentral sulcus.

Typically, it consists in representing taxonomic relationships, direct mereological relationships, and topological relationships among the smallest parts.
The *extended abstract model* is generated automatically. This step consists of inferring all the dependent relationships among composed anatomical structures. 59.7% of the relationships from the extended abstract model are automatically created [10].
The *extended lateralized model* is generated by applying specialization rules for lateralization to the extended abstract model. These rules:

- create the lateralized concepts as subconcepts of those defined on the abstract level (e.g., `Left Frontal Lobe` and `Right Frontal Lobe` are subsumed by `Frontal lobe`);
- add consistency statements (e.g., `Left Frontal Lobe` and `Right Frontal Lobe` are taxonomically disjoint, and `Frontal Lobe` is equivalent to `Left Frontal Lobe` or `Right Frontal Lobe`);
- generate all the required relationships (e.g., from the statement "`Frontal Lobe` *`hasAnatomicalPart`* `PreCentral Gyrus`", we would infer that `Frontal Lobe` *`hasAnatomicalPart`* `Left PreCentral Gyrus` (respectively right) and that `Left Frontal Lobe` (respectively right) *`hasAnatomicalPart`* `Left PreCentral Gyrus` (respectively right)).

Incremental consistency

Managing *incremental consistency* consists in making sure that intrinsic consistency is still respected after an update of the knowledge base, and that the result meets the "realism of representation" requirement. It can be reduced to answering the following questions :

1. Does every concept and relationships that we wanted to add belong to the model? For instance, if we add a part for a gyrus, we want this structure to be a part of every anatomical concept the gyrus is a part of.

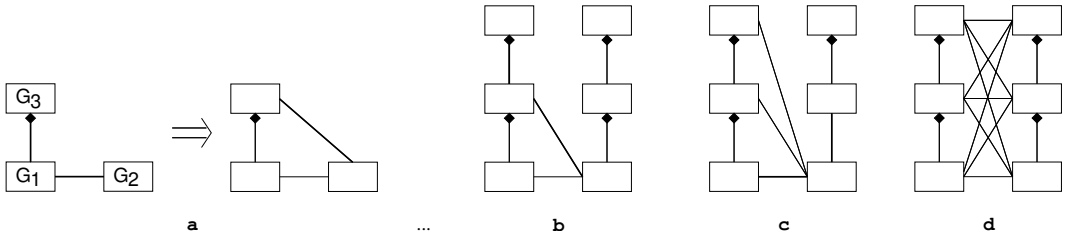


Figure 6: Relationships depending on `hasAnatomicalPart` (the compound is on the diamond-side of the line) and on topological relationships of continuity, contiguity or separation by a sulcus (in the latter case, only the two cortical structures are represented on the schema). Example: for a, G_1=`Opercular Pars of Inferior Frontal Gyrus`, G_2=`PreCentral Gyrus`, G_3=`Inferior Frontal Gyrus`. The separation of G_3 and G_2 depends on that of G_1 and G_2 by `PreCentral Sulcus`.

2. Did the consistency rules generate any concepts or relationships that do not correspond with anything in canonical knowledge? For instance, a wrong inference rule will generate erroneous relationships.
3. Has every concept and relationship that we wanted to remove actually disappeared?
4. Did we remove from the model more than we should have? For instance, removing a relationship in order to fix the model has for consequence of removing all the dependent relationships, some of which being right.

Because this step consists in comparing the result with canonical knowledge, it has to be performed manually by a (human) domain expert. However, a simple tool has been developed to assist this task.

Every update of the knowledge base only takes place in the restricted abstract model. The abstract and lateralized extended models are then regenerated automatically. A simple XML Stylesheet helps the domain specialist to compare them with their previous versions. As a result, an HTML page is generated which explicitly represents in green all the concepts and relationships that have been added, and in red those which have been removed, similar to the *diff* command.

Discussion

As we are confronted with an increasing number of concepts and relationships, maintaining the ontology's consistency becomes more and more difficult. In addition, the growth of the model is complicated by the need to add a lot of integrity constraints to the model so that it is not too lax. Therefore, our approach seems to be more and more relevant.

Unfortunately, to our knowledge, none of the main symbolic models of anatomy such as the Digital Anatomist Foundational Model and Galen supports an explicit representation of the dependencies among concepts or relationships. This point is particularly important, since both of these ontologies have to handle concepts and relationships that number in the tens of thousands. Galen's *sanctioning statements* [15] are assertions preventing impossible situations (e.g., "fracture of the eyelid") or redundant ones (e.g., "the hand which is a part of the arm"). They play an important role in Galen consistency, but do not address semantic dependencies between relationships.

Although our approach has only been applied to a model of the brain cortex, it seems that the principle could be extended to any anatomical model. Moreover, it could also be extended to other domains. However, anatomical knowledge is rather stable. Other domains such as pathology or the study of brain functions are more likely to evolve, which would require in addition a management of obsolescence –something we haven't studied.

In addition to being used in specific domains, identification of semantic dependencies is also of particular importance when establishing mappings between domains. For instance, pathology located in a part of an anatomical structure may also need to be recognized as located in the anatomical structure overall. Schulz provides an interesting analysis of this kind of problems [16, 17]. These capabilities are needed considering the role of anatomy as a localization reference, and its use in application contexts that require automatic reasoning.

The dependencies identified in this article, and their usage to maintain semantic consistency of an anatomic model are beyond the scope of logical consistency-checking tools such as ConsVISor [13] or FaCT[2]. For instance, ConsVISor would not issue any warning if the central sulcus separates the precentral and postcentral gyri but not the frontal and parietal lobes.

This paper describes the management of consistency

[2] `http://www.cs.man.ac.uk/~horrocks/FaCT`

from the modeling point of view. It does not rely on any representation formalism. However, it turned out that the consistency rules could not be easily represented in ontology languages such as OWL [18]. Extensions such as RuleML[3] or SWRL[4] could provide very interesting future extensions. They would allow to represent explicitly some consistency constraints to map anatomy to pathology (e.g., to express that a tumor located in a part of an organ has also to be considered as a tumor located in the organ itself).

The functionality of the script used for managing the incremental consistency is similar to that of the *diff* command or of the PROMPT plugin for Protégé [19] (but less flexible). However, the usage of a specific modeling environment is beyond the scope of this article.

By automatically generating more than 59% of the relationships, our approach makes the task of the curator easier, less error-prone and hopefully less tedious. However, choosing the appropriate modifications in the abstract restricted model requires a good understanding of the existing dependency rules. Here again, it is possible to devise some tools for assisting the curator and detecting any principle violation. Eventually, if any problem is detected by the domain expert during the enforcing of incremental consistency, the curator will be in charge of determining if it comes from a modeling error or from an erroneous rule.

Conclusion

Our effort to identify explicitly the properties of the relationships we used, as well as the experience of building the ontology, allowed us to identify dependencies among concepts and relationships. The explicit representation of these dependencies is important for the representation of the semantics of the domain. In addition, it turned out that it can be helpfully used to assist in the management of the knowledge base and to ensure the model's semantic consistency.

The method we adopted consists in maintaining only a core set of independent concepts and relationships. All the dependent items are then automatically generated. A domain expert still have to manually screen the result in order to make sure that it is correct with regard to canonical knowledge.

Acknowledgments

Xavier Morandi assessed the knowledge base and shared his experience of neuroanatomy. Christine Golbreich provided numerous insights and constructive remarks for rules modeling.

[3] http://www.dfki.uni-kl.de/ruleml

[4] http://www.daml.org/2003/11/swrl

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, (5), 2001.

[2] J. Brinkley, B. Wong, K. Hinshaw, and C. Rosse. Design of an anatomy information system. *IEEE Computer Graphics*, 19(3):38–48, 1999.

[3] N.F. Noy, M.A. Musen, J.L.V. Mejino, and C. Rosse. Pushing the envelope: Challenges in a frame-based representation of human anatomy. *Data and Knowledge Engineering Journal*, page In Press, 2004.

[4] C. Rosse, J.L. Mejino, R. Modayur, B.R.and Jakobovits, K.P. Hinshaw, and J.F. Brinkley. Motivation and organizational principles for anatomical knowledge representation: The digital anatomist symbolic knowledge base. *Journal of the American Medical Informatics Association*, 5(1):17–40, Jan/Feb 1998.

[5] A.L. Rector, W.A. Nowlan, and the GALEN Consortium. The GALEN project. *Computer Methods and Programs in Biomedicine*, 45:75–78, 1994.

[6] M. Ono, S. Kubik, and C. Abernathey. *Atlas of the Cerebral Sulci*. Thieme Medical Publishers, Inc., 1990.

[7] J. Talairach and P. Tournoux. *Co-Planar Stereotactic Atlas of the Human Brain*. Georg Thieme Verlag, 1988.

[8] D.M. Bowden and R.F. Martin. Neuronames brain hierarchy. *Neuroimage*, 2:63–83, 1995.

[9] O. Dameron, B. Gibaud, A. Burgun, and X. Morandi. Towards a sharable numeric and symbolic knowledge base on cerebral cortex anatomy: lessons from a prototype. In *American Medical Informatics Association AMIA02 proceedings*, pages 185—189, 2002.

[10] O. Dameron, A. Burgun, X. Morandi, and B. Gibaud. Modelling dependencies between relations to insure consistency of a cerebral cortex anatomy knowledge base. In *MIE'03 Proceedings*, pages 403–408, 2003.

[11] A. Artale, E. Franconi, N. Guarino, and L. Pazzi. Part-whole relations in object-centered systems: An overview. *Data Knowledge Engineering*, 20(3):347–383, 1996.

[12] A Varzi. Parts, wholes, and part-whole relations: The prospects of mereotopology. *Data and Knowledge Engineering*, 20:259–86, 1996.

[13] K. Baclawski, M. Kokar, R. Waldinger, and P. Kogut. Consistency checking of semantic web ontologies. In I. Horrocks and J. Hendler, editors, *International Semantic Web Conference ISWC02 proceedings*, volume 2342 of *LNCS 2342*, pages 454–459. Springer-Verlag Heidelberg, 2002.

[14] M. Buchheit, F. Donini, and A. Schaerf. Decidable reasoning in terminological knowledge representation systems. Technical Report RR-93-10, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH Erwin-Schrödinger Strasse Postfach 2080 D67608 Kaiserslautern Germany, 1993.

[15] A.L. Rector, S.K. Bechhofer, C.A. Goble, I. Horrocks, W.A. Nowlan, and W.D. Solomon. The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, 9:139–171, 1997.

[16] S. Schulz, U. Hahn, and M. Romacker. Modeling anatomical spatial relations with description logics. In *AMIA 2000 Annual Symposium*, pages 779–783, 2000.

[17] S. Schulz and U. Hahn. Mereotopological reasoning about parts and (w)holes in bio-ontology. In C Welty and B Smith, editors, *Formal Ontology in Information Systems FOIS01*, pages 210–221, 2001.

[18] C. Golbreich, O. Dameron, B. Gibaud, and A. Burgun. Web ontology language requirements w.r.t expressiveness of taxononomy and axioms in medecine. In *International Semantic Web Conference ISWC03 proceedings*, 2003.

[19] N.F. Noy and M.A. Musen. The prompt suite: Interactive tools for ontology merging and mapping. *Journal of Human-Computer Studies*, 59(6):983–1024, 2003.

Weaving the Biomedical Semantic Web with the Protégé OWL Plugin

Holger Knublauch Olivier Dameron Mark A. Musen

Stanford Medical Informatics, Stanford University, Stanford, CA (`http://protege.stanford.edu`)

Abstract

In this document we show how biomedical resources can be linked into a Semantic Web using Protégé. Protégé is a widely-used open-source ontology modeling environment with support for the Web Ontology Language (OWL). With the example domain of brain cortex anatomy we demonstrate how Protégé can be used to build an OWL ontology and to maintain ontology consistency with a description logic classifier. We also show how Protégé can be used to link existing Web resources such as biomedical articles and images into a Semantic Web.

INTRODUCTION

Biomedical Web resources in the existing internet are mainly optimized for use by humans. For example, researchers need to know the "correct" keywords to do a meaningful search using an online publications database. The vision of the Semantic Web [3] is to extend the existing Web with conceptual metadata that are more useful to machines, revealing the intended meaning of Web resources. This meaning could be used by software agents to perform tasks that are difficult with the current Web architecture. For example, an intelligent agent could retrieve semantically related publications, even if they don't contain the "correct" keyword.

Ontologies are a central building block of the Semantic Web. Ontologies define domain concepts and the relationships between them, and thus provide a domain language that is meaningful to both humans and machines. Ontologies are being defined for many biomedical domains, such as anatomy, genetics, and cancer research. The concepts from these ontologies can be used to annotate Web resources. The Web Ontology Language (OWL) [13] is widely accepted as the standard language for sharing Semantic Web contents. Protégé [4, 7] is an ontology development environment with a large community of active users. Protégé has been used for more than a decade to build large-scale biomedical applications. Rather recently, Protégé has been extended with support for OWL, and has become one of the leading OWL tools.

Our goal in this document is to help biomedical projects get started with Semantic Web technology. We first describe the architecture of a typical biomedical Semantic Web application from the domain of brain cortex anatomy. Then we give a short overview of Protégé and its OWL support. We describe how Protégé can be used to define domain classes and properties, and how to use features such as a classifier to maintain semantic consistency. We also briefly introduce the essential features of OWL and their representation in Protégé. Then we show how to link existing Web resources into the Semantic Web, so that they can be accessed by intelligent agents. We end this document with discussion and conclusions.

A BIOMEDICAL SEMANTIC WEB

The current Internet already contains vast amounts of biomedical information resources, such as research articles, images, clinical guidelines, and drug catalogues. Making these resources available in a more structured way is one of the goals of several large-scale ontology development efforts. For example, the goal of the National Cancer Institute's Thesaurus project [5] is to provide a well-defined conceptual model so that cancer-related resources can be structured in a machine-readable way. This conceptual model is an OWL ontology with tens of thousands of classes and dozens of properties.

For the purpose of this paper, we start with a less ambitious example ontology of brain cortex anatomy. Potential use cases of this ontology are teaching, decision support for clinical practice, sharing of neuroimaging data, or semantic assistance for data processing tools. The ontology defines concepts such as `FrontalLobe` and `LeftCentralSulcus`, and specialization, composition and spatial neighborhood relationships. In addition, the ontology also defines the logical characteristics of the concepts. For example, it states that a brain `Hemisphere` is composed of exactly five distinct lobes: one `FrontalLobe`, one `ParietalLobe`, one `TemporalLobe`, one `OccipitalLobe` and one `LimbicLobe`. These concepts and relationships are implemented as OWL classes and properties. They are stored in an OWL file which resides on a publicly accessible Web server. After the ontology has been published on the Web, other OWL ontologies, resources, agents, and services can

link to this file and use the ontology's concepts. For example, a Web repository of MRI scans could provide a collection of image metadata objects that would represent the attributes of the single scans (dimensions, resolution, contents), so that the best images for a specific topic can be retrieved automatically. If the image repository is loosely coupled and distributed over multiple hosts (e.g., multiple hospitals), then each of the servers could provide its own metadata objects. A user searching for a particular scan of a frontal lobe could then invoke an intelligent agent that would crawl through the various repositories to search for the best matches.

Another example of a Semantic Web application would be a context-sensitive search function for research articles. A publication database such as PubMed could provide a Web service that would refer to a conceptual model when providing metadata about articles. It could also rely on this conceptual model to guide and assist query processing. Users could invoke this Web service through a simple client application. The Web service could exploit the definitions from the ontology to widen or narrow the search into concepts that are substantially related to the terms the user has asked for. For example, it could deliver papers about glioma located in the precentral gyrus although the user has only asked for tumors of the frontal lobe, exploiting the background knowledge that a glioma is a kind of tumor and that the precentral gyrus is a part of the frontal lobe.

One of the advantages of shared conceptual models is that they can be reused in various contexts, even some that have not been imagined yet. Finally, the Semantic Web could even be used to point researchers and domain experts into new directions and to reveal cross-links between domains.

These examples illustrate the central role of *ontologies* in Semantic Web applications. Ontologies should adequately represent a domain and allow some kind of formal reasoning. They should be both understandable by humans and processable by software agents. Furthermore, since ontologies will evolve over time, they need to be maintainable. This demands for ontology modeling tools that provide a user-friendly view on the ontology and support an iterative working style with rapid turn-around times. Tools should also provide intelligent services that reveal inconsistencies and hidden dependencies among definitions.

PROTÉGÉ AND THE OWL PLUGIN

Since its beginning in the 1980's, Protégé has been driven by biomedical applications. Protégé started as a rather specialized tool for a specific kind of problem solving [4], but evolved into a very generic and flexible platform for many types of knowledge-based applications and tools from all kinds of domains.

Protégé can be characterized as an ontology development environment. It provides functionality for editing classes, slots (properties), and instances. One of its strengths is that it can automatically generate a user interface from class definitions, and thus can support rapid knowledge acquisition. Protégé supports database storage that is scalable to several million concepts, and provides multi-user support for synchronous knowledge entry.

The current version of Protégé (2.1) is highly extensible and customizable. At its core is a frame-based knowledge model [9] with support for metaclasses. These metaclasses can be extended to define other languages on top of the core frame model [10]. For these other languages, Protégé can be extended with back-ends for alternative file formats. Currently, back-ends for Clips, UML, XML, RDF, DAML+OIL, and OWL are available for download.

Protégé not only allows developers to extend the internal model representation, but also to customize the user interface freely. As illustrated in Figure 1, Protégé's user interface consists of several screens, called *tabs*, which display different aspects of the ontology in different views. Each of the tabs can be filled with arbitrary components. Most of the existing tabs provide a tree-browser view of the model, with a tree on the left and details of the selected node on the right hand side. The details of the selected object are typically displayed by means of *forms*. The forms consist of configurable components, called *widgets*. Typically, each widget displays one property of the selected object. There are standard widgets for the most common property types, but ontology developers are free to replace the default widgets with specialized components. Widgets, tabs, and back-ends are called *plugins*. Protégé's architecture allows developers to add and activate plugins arbitrarily, so that the default system's appearance and behavior can be completely adapted to a project's needs.

The OWL Plugin[1] [8] is a complex Protégé plugin with support for OWL. It can be used to load and save OWL files in various formats, to edit OWL ontologies with custom-tailored graphical widgets, and to provide access to reasoning based on description logic. As shown in figure 1, the OWL Plugin's user interface provides various default tabs for editing OWL classes, properties, forms, individuals, and ontology metadata. The following section explains how to use the Classes, Properties and Metadata tabs for the de-

[1] `http://protege.stanford.edu/plugins/owl`

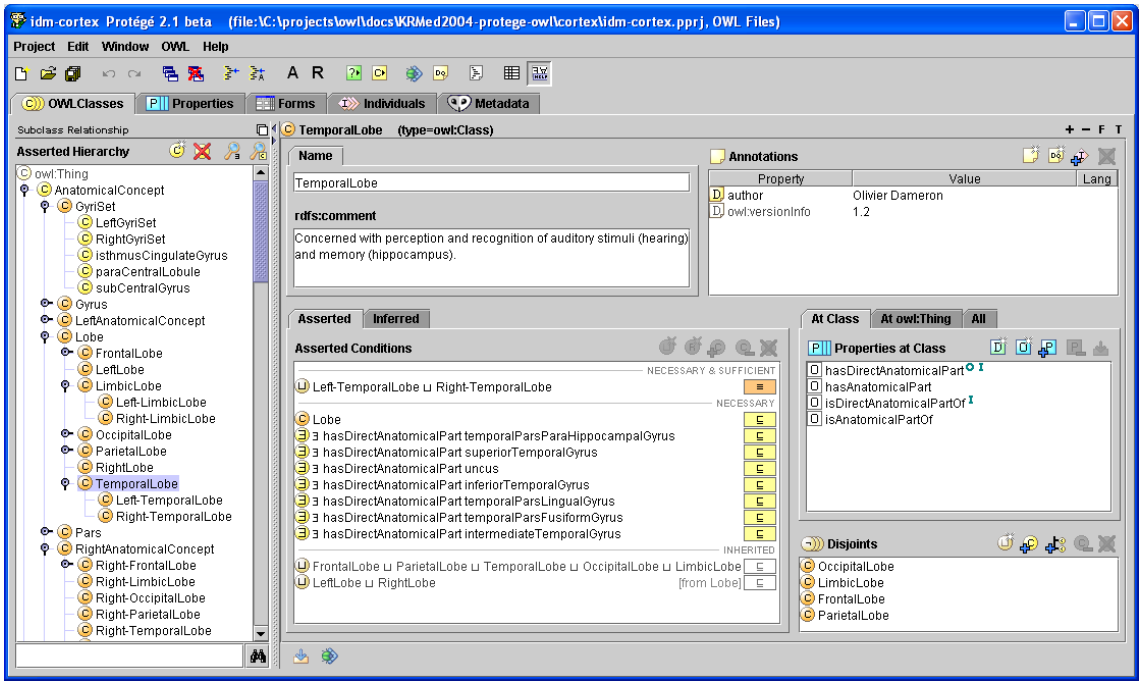


Figure 1: The class editor of the Protégé OWL Plugin.

sign of a biomedical ontology. The section after that introduces how to use the Individuals and Forms tabs for the acquisition of Semantic Web contents.

BUILDING OWL ONTOLOGIES WITH PROTÉGÉ

An OWL ontology can be regarded as a network of classes, properties, and individuals. *Classes* define names of the relevant domain concepts and their logical characteristics. *Properties* (sometimes also called slots, attributes or roles) define the relationships between classes, and allow to assign primitive values to instances. *Individuals* are instances of the classes with specific values for the properties. The Semantic Web can be regarded as a network of ontologies and other Web resources. OWL ontology concepts can have references to concepts in other ontologies. The basic mechanism for this capability is ontology import (i.e., an ontology can import resources from existing ontologies and create instances or specializations of their classes).

In our biomedical example ontology, we have a class called `CentralSulcus` which is defined as a kind of `AnatomicalConcept` that has a measured average depth. Individuals from this ontology would describe specific case data (e.g., a specific left central sulcus of an individual with the value of 23 mm for its depth). For the example ontology, we can import an existing ontology about units, and thus reuse the concepts from other files and support knowledge sharing. Let's take a look at how these elements can be defined in Protégé.

Classes

The most important view in the Protégé OWL Plugin is the OWLClasses tab (Figure 1). This tab displays the tree of the ontology's classes on the left, while the selected class is shown in a form in the center. The upper region of the class form allows users to edit class metadata such as name, comments, and labels, in multiple languages. The widget in the upper right area of the form allows users to assign values for *annotation properties* to a class. Annotation properties can hold arbitrary values such as author and creation date. Ontologies can define their own annotation properties or reuse existing ones such as those from the Dublin Core ontology. In contrast to other properties, annotation properties do not have any formal meaning for external OWL components like reasoners, but they are an extremely important vehicle for maintaining project-specific information. A typical use case for annotation properties in a biomedical field is to assign standardized identifiers such as ICD codes for concepts that describe a disease. Annotation properties, such as the predefined `rdfs:seeAlso`, can also be used to define cross-references between concepts. The OWL Plugin also uses annotation properties to store Protégé-specific information, and to manage "to-do" lists for ontology authors.

Properties

The *Properties* widget of the OWLClasses tab allows users to view and create relationships between classes. It provides access to those properties that could be used by the instances of the current class. The characteristics of a property are edited through the form shown in Figure 2. This form provides a metadata area in the upper part, displaying the property's name, annotations, and so on, similar to the presentation in the class form.

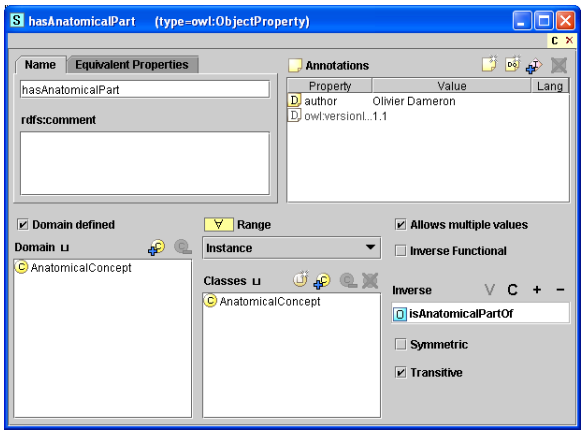


Figure 2: An OWL property form in Protégé.

The available choices in the *Range* drop-down box depend on whether the property is a *datatype property* with primitive values, or an *object property* with references to other classes. For datatype properties, Protégé supports enumerations of symbols (`owl:oneOf`), and all reasonable XML Schema datatypes, grouped into booleans, floats, integers, and string types. For example, the datatype property *`hasMeasuredDepth`* can only take floats as values. Object properties can store references to individuals or classes from the ontology. For example, the object property *`hasAnatomicalPart`* can only take instances of `AnatomicalConcept` as values.

Depending on whether a property is an object or a datatype property, Protégé provides widgets for other property characteristics, such as whether the property is symmetric or transitive. Symmetric properties describe bidirectional relationships (i.e., if A is related to B via property R_s, then B is also related to A). For example, the contiguity relationship is symmetric. A property R_t is transitive if when A is related to B by R_t and B is related to C by R_t, then (A is also related to C by R_t). Part/whole relationships such as *`hasAnatomicalPart`* are usually considered to be transitive.

The *Domain* widget can be used to restrict a property's domain (i.e., the list of classes where the property can be used). Domain restrictions are optional and often left blank in OWL ontologies, because they may slow down some reasoning processes. If a property does not have a domain restriction, then it can be used for instances of any class.

Specialization

OWL has its theoretical foundation in description logic [1]. In description logic, a class is a set of individuals. The concept corresponding to the set of all individuals is usually called *Top* ($\top$), or *Thing*. Whenever the set of the individuals of a class B is a subset of the set of the individuals of a class A, B is said to be a *subclass* of A (noted $B \sqsubseteq A$). B is also said to be a kind of A. All classes are subconcepts of $\top$.

In other words, superclasses define *necessary* conditions for class membership. Conversely, subclasses define *sufficient* conditions for class membership. For example, being a frontal lobe is a necessary condition for being a left frontal lobe: in order to be an instance of `LeftFrontalLobe`, an individual has to be an instance of `FrontalLobe` (and most certainly has to fulfill other requirements). Conversely, being a left frontal lobe is a sufficient condition for being a frontal lobe: every instance of `LeftFrontalLobe` is also an instance of `FrontalLobe` (but there may be other instances of `FrontalLobe` that are not instances of `LeftFrontalLobe`).

It is really important to keep in mind that a subconcept is a subset of individuals. Indeed, it is a common mistake to mix specialization and composition hierarchies. However, defining `UpperLobeOfLung` as a subconcept of `Lung` is erroneous because a lobe of a lung is not a kind of lung, but a part of a lung. Correct subconcepts for lung could be `LeftLung` and `RightLung`.

The specialization principle also implies inheritance of the properties. For instance, if we say that every `Sulcus` has an *`averageDepth`* and that `CentralSulcus` is a subclass of `Sulcus`, then every `CentralSulcus` also has an *`averageDepth`*. Because subclasses are more specific than their superclasses, the range of a subclass may itself be a subclass of the range of the superclass. This is called *property restriction*. For example, we can say that every `Sulcus` has a side in the class `Side`, and that every `LeftSulcus` (subclass of `Sulcus`) has a side `LeftSide` (subclass of `Side`).

In Protégé, the tree widget of the OWLClasses tab is organized according to the subclass hierarchy. We can see that `owl:Thing` (which represents $\top$) is the root of the tree. Protégé users can browse, view, and edit the classes from the tree, create new subclasses, and

move classes easily with drag-and-drop. Direct superclasses are also listed in the Conditions widget, which is described next. The OWL Plugin also allows to navigate and edit ontologies according to other relationships between classes, in particular to visualize the part-of relationships that are so common in biomedical domains.

Logical Class Characteristics

The *Conditions* widget of the OWLClasses tab allows to fully take advantage of OWL's description logic support, and to express conditions on the classes based on property restrictions and other expressions. The syntax used for OWL expressions in Protégé is summarized in table 1.

The key point here is to understand that an expression involving a property and its range such as "∃ *property* Concept" or "∀ *property* Concept" represents a set of individuals, and therefore can be interpreted as a concept. For example, (∃ *hasPart* Lobe) is the set of all the individuals related to at least one instance of Lobe by the *hasPart* relationship (they could also be related to instances of other concepts). Conversely, (∀ *hasPart* Lobe) is the set of all the individuals which are exclusively related to instances of Lobe by the *hasPart* relationship (or which are related to nothing by this relationship). Similarly, the union and intersection of two sets are also sets and can be interpreted as classes. For example, (LeftAnatomicalPart ⊓ Gyrus) represents the set of all left anatomical parts that are at the same time gyri, and (LeftGyrus ⊔ RightGyrus) represents the set of individuals that are instances of either concept. The ¬ operator can be used to define a class of any individual except those from a given class. For instance, ¬LeftSide is the set of all the individuals that are not instance of LeftSide. Finally, OWL also allows to define a class by exhaustively enumerating its instances.

The logical symbols used by the Protégé OWL Plugin are widely used in the description logic community [1]. Their major advantage is that they allow to display even complex class expressions in a relatively compact form. As shown in Figure 3, Protégé provides a convenient expression editor with support for either mouse or keyboard editing. However, some domain experts, especially from rather non-technical domains such as biomedicine, may require some training before they get used to these symbols. For these users, Protégé provides an English prose explanations of an OWL expression when the mouse is moved over it. Our collaborators are also working on alternative editors which support a rather template-based editing metaphor. Protégé's generic form architecture allows to quickly assemble alternative editors into the environment.

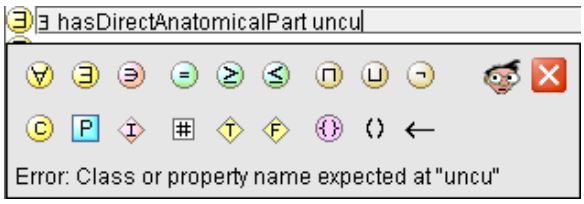


Figure 3: Protégé provides a comfortable editor for arbitrary OWL expressions.

The formal definitions of the OWL primitives can be exploited by reasoners. They compute the specialization relationships (inheritance) between the classes based on their logical definitions. This reasoning support has shown to be a very valuable feature during ontology design, particularly in biomedical domains ([5, 11]). Ontology designers can periodically invoke a reasoner to see whether the logical class definitions meet the expectations, and to make sure that no inconsistency arise.

Necessary conditions. As mentioned above, a necessary condition for an individual to be an instance of a class is to be an instance of all the superclasses of this class. In addition to saying that a class is a subclass of its superclasses, such as FrontalLobe is a subclass of Lobe, necessary conditions allow the specify the properties that the class has to fulfill. This is an important activity when building an ontology, because, we don't want to limit ourselves to saying that a frontal lobe is a kind of lobe; we also want to represent what is specific to the frontal lobe, as opposed to the other lobes. For example, the frontal lobe has to be delimited by the central sulcus, as well as by the lateral sulcus. Therefore, to the original condition FrontalLobe ⊑ Lobe, we can add the two following necessary conditions "FrontalLobe ⊑ (∃ *isDelimitedBy* CentralSulcus)" and "FrontalLobe ⊑ (∃ *isDelimitedBy* LateralSulcus)". These conditions can also hold for other concepts, but an individual that fails to fulfill these conditions cannot be an instance of FrontalLobe.

Necessary and sufficient conditions. Necessary conditions can be interpreted as subset-superset relationships between sets of individuals. Similarly, we may want to represent that two classes have exactly the same instances (they are mutual subclasses of the other). For example, as the left and the right frontal lobe are two kinds of frontal lobe, we have

OWL element	Symbol	Key	Example expression in Protégé
owl:allValuesFrom	∀	*	∀ *hasPart* Lobe
owl:someValuesFrom	∃	?	∃ *hasDirectAnatomicalPart* RectusGyrus
owl:hasValue	∋	$	*hasColor* ∋ yellow
owl:minCardinality	≥	>	*hasSide* ≥ 1 (at least one value)
owl:maxCardinality	≤	<	*hasSide* ≤ 2 (at most two values)
owl:cardinality	=	=	*hasSide* = 1 (exactly one value)
owl:intersectionOf	⊓	&	LeftAnatomicalConcept ⊓ Gyrus
owl:unionOf	⊔	|	LeftGyrus ⊔ RightGyrus
owl:complementOf	¬	!	¬LeftSide
owl:oneOf	{ ... }	{ }	{yellow green red}

Table 1: Protégé uses traditional description logic symbols to display OWL expressions. Property names such as *hasSide* appear in italics. A common naming convention is to use uppercase names such as Lobe to represent classes, while individuals like yellow should be written in lower case.

the following condition: (LeftFrontalLobe ⊔ RightFrontalLobe) ⊑ FrontalLobe. But we also want to say that every frontal lobe is either a left or a right frontal lobe. Therefore, we use a necessary and sufficient condition (LeftFrontalLobe ⊔ RightFrontalLobe) ≡ FrontalLobe, which basically says that if you have a frontal lobe, then it is either a left or a right one (⊒); and that if you have a left or a right frontal lobe, then it is a frontal lobe (⊑). Classes with necessary and sufficient conditions are called *defined* classes (represented by orange icons in Protégé), while classes with only necessary conditions are called *primitive* (yellow icons). The Conditions widget allows to edit either type of conditions, and to copy or move expressions between blocks.

The open world assumption. Description logic make the so-called *open world assumption*, that is what is not said denotes a lack of knowledge (whereas in other contexts such as databases, what is not said is assumed to be false). A direct consequence is that if we don't say explicitly that two classes such as LeftFrontalLobe and RightFrontalLobe are disjoint, then it is perfectly valid for them to have individuals in common. The *Disjoints* widget, in the lower right corner of the OWLClasses tab allows users to represent axioms to control this aspect.

Classification and Consistency Checking

One of the major strengths of description logic languages like OWL is their support for intelligent reasoning. In our context, *reasoning* means to infer new knowledge from the statements asserted by an ontology designer. *Reasoners* are tools that take an ontology and perform reasoning with it. The OWL Plugin can interact with any reasoner that supports the standard DIG interface, such as Racer [6]. Since these reasoners are separate tools we will not discuss their details in this paper, but focus on their application oriented utility. During ontology design, the most interesting reasoning capabilities from these tools are classification and consistency checking.

Classification. Classification is used to infer specialization relationships between classes from their formal definitions. Basically, a classifier takes a class hierarchy including the logical expressions, and then returns a new class hierarchy, which is logically equivalent to the input hierarchy. As illustrated in Figure 4, Protégé can display the classification results graphically. After the user has clicked the classify button, the system displays both the asserted and the inferred hierarchies, and highlights the differences between them.

For example, we defined LeftFrontalLobe as any frontal lobe located in the left hemisphere (LeftFrontalLobe ≡ (FrontalLobe ⊓ LeftAnatomicalConcept)). Therefore, it appears as a direct child of the last two concepts in the asserted hierarchy (Figure 4). Similarly, we also defined LeftLobe as any lobe located in the left hemisphere (LeftLobe ≡ (Lobe ⊓ LeftAnatomicalConcept)). Because the definition of LeftFrontalLobe doesn't mention LeftLobe, these two concepts don't appear to be related. However, after classification, the reasoner infers from FrontalLobe ⊑ Lobe that LeftFrontalLobe is also a subclass of Leftlobe. Note: we could as well have defined LeftFrontalLobe ≡ (FrontalLobe ⊓ LeftLobe), but then we wouldn't have known that it is also a LeftAnatomicalConcept until the reasoner have found out.

This reasoning capability associated with description logic is of particular importance because it allows the

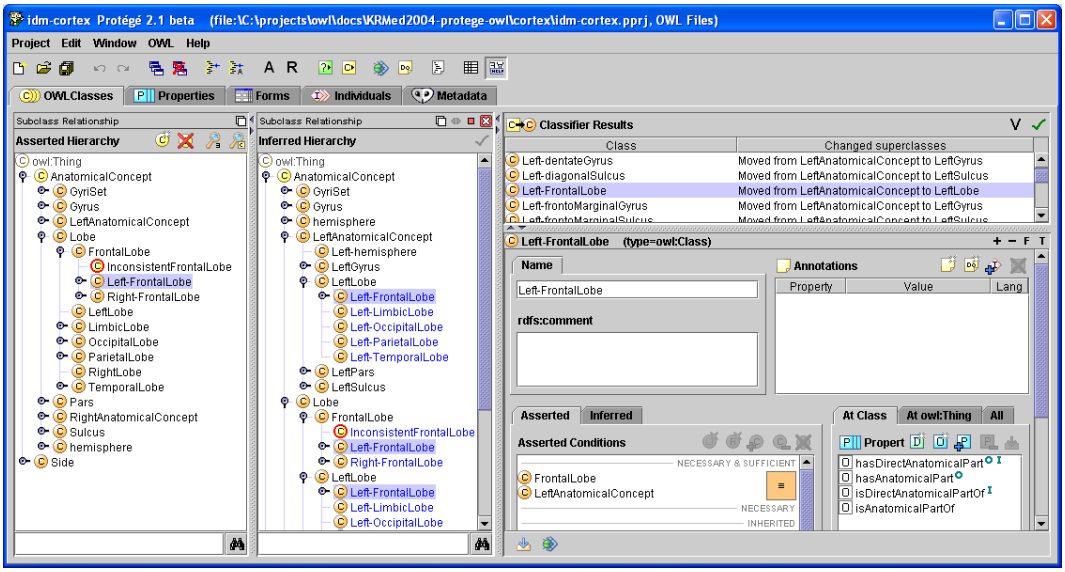


Figure 4: Protégé provides access to description logic classifiers and can display both the asserted and the inferred class relationships.

user to provide intensional definitions for the classes. The specialization relationships become consequences of these definitions, and allow constraints inheritance. Without reasoning capabilities, the approach of creating an ontology is more extensional. It would require to explicitly state every specialization relationships between the concepts (e.g., in the previous example between `LeftFrontalLobe` and `LeftLobe`). This support is especially valuable in the domain of biomedicine, with its deeply nested hierarchies and multi-relationships between almost every part of the anatomy [1, 12]. Using OWL, ontology designers could just add a new concept by describing its logical characteristics, and the classifier would automatically place it in its correct position. Furthermore, it would report the side-effects of adding a new class.

Consistency checking. In addition to providing automatic classification, reasoning capabilities can be exploited to detect logical inconsistencies within the ontology. We could introduce a class `InconsistentFrontalLobe`, which is both a `LeftFrontalLobe` and a `RightFrontalLobe`. Since the last two concepts are defined to be disjoint, the reasoner reports that no individual can be an instance of this class. Clearly, these consistency checks can help tremendously in the construction and maintenance of large biomedical terminologies [12].

OWL Full and OWL DL An important issue with reasoning in OWL is that many reasoners are not able to handle the full expressivity of OWL. The OWL specification distinguishes between OWL Full and OWL DL to indicate which language elements are typically tractable for reasoners. Ontologies that use OWL Full elements such as metaclasses cannot be classified. Protégé allows users to edit some OWL Full concepts and provides features to help convert the ontology into OWL DL when a classifier is to be used. However, since OWL Full ontologies can state anything about anything, Protégé does not support the complete OWL Full syntax.

LINKING BIOMEDICAL RESOURCES INTO THE SEMANTIC WEB

This section demonstrates how to use OWL to link biomedical resources into the Semantic Web. In our scenario, OWL ontologies provide the vocabulary for describing the contents of images and scientific articles.

In order to describe biomedical images, we have defined a small image ontology, which basically only defines a single class `Image`, and defines four properties for each image: the integer properties *`hasWidth`* and *`hasHeight`* provide the dimensions of the image, the property *`hasURI`* stores a reference to the image's location, and the property *`hasContents`* can link an `Image` to an OWL class, such as those defined in the brain cortex ontology. These content concepts can later be used by intelligent agents for search purposes. Protégé can now be used to create a new ontology `cortex-images.owl`, which imports the cortex

ontology and the images ontology. The new ontology basically contains instances of the `Image` class, and uses the classes from the cortex ontology as contents values. Whenever concepts are imported from another ontology, Protégé displays them with a prefix such as `cortex:`.

Protégé provides excellent support for the acquisition of instances. As illustrated in Figure 5, the OWL Plugin makes this functionality available through the *Individuals* tab. For each class in an ontology, Protégé generates forms with appropriate widgets to acquire instances of the class. The Individuals tab shows the classes, their instances, and a form for the selected instance. By default, this form will contain default widgets, such as a numeric text field for integer properties and a clickable list for object properties. For example, Protégé has selected a list widget with create, add and remove buttons for the *`hasContents`* property. However, for the *`hasURI`* property, the system has selected a simple text field widget, which is not optimized for displaying images.

Fortunately, Protégé provides a *Forms* tab, which can be used to customize the forms. The Forms tab allows users to move and resize the widgets, and to replace widgets with other suitable ones. In our example, we have replaced the default text field widget for *`hasURI`* with an image widget, so that a preview of the image can be shown below the URI. Protégé's open architecture allows users to add arbitrary Java components as widgets, if the catalogue of default widgets is not sufficient. With a little bit of programming, we could provide a widget that allows users to select an image, and then fills the values of width and height automatically.

After the instances/individuals have been edited, they can be exported onto a Web server, so that agents can find and process them. A simple search agent would crawl through multiple image repositories, and analyze the image ontologies using an OWL parsing library such as Jena[2]. Supplied with a search concept such as `FrontalLobe`, an agent could then retrieve and filter images by their semantic proximity. A very similar approach can be used to implement a repository of scientific articles.

DISCUSSION AND CONCLUSION

Our main goal in this paper was to introduce the Protégé OWL Plugin, and to show that it provides a promising platform for biomedical ontology and Semantic Web projects. The OWL Plugin pioneers user-friendly components for building and reasoning with description logic ontologies. While researchers from the description logic community have managed to create deeply studied maps of their theoretical terrain, we believe it is now time to put languages such as OWL into practice, and thus reveal the strengths and weaknesses of these languages for particular domains in everyday use. Some issues of how to handle description logic in the development of large clinical terminologies have already been discussed by others (e.g., [5, 12, 11]). However, more work is necessary, in particular in training biomedical domain experts to use the rich semantics of OWL.

Some of the advantages of OWL are already obvious. Descriptions logic rely on a well defined semantics which makes modeling not only the structure, but also the meaning of a domain possible. As opposed to other formalisms such as frames [9], description logic allow users to provide intensional definitions for the concepts. As a consequence, ontologies are more compact, less error-prone, and easier to maintain. The precise semantics of description logic makes it possible to perform automatic reasoning. The intensional definitions of the concepts can be exploited by classifiers. Therefore, when adding a new class, one doesn't have to worry anymore about putting it in the right place in the taxonomic hierarchy. Moreover, multiple inheritance is automatically detected and dealt with. Classifiers can detect any logical inconsistencies in a class definition, that would prevent it of having instances. Eventually, reasoners can infer the correct relationships when combining ontologies of related domains, or extending an ontology with context-specific features. This point favors the sharing of common semantic references and their reuse in various contexts. Therefore, we expect OWL to play a key role not only for the Semantic Web, but also for the evolution and sharing of biomedical knowledge.

A final note about other ontology modeling tools. Given the short history of the Semantic Web, there are few other tools available with OWL support. One of the most popular ontology editors beside Protégé is OilEd [2]. From the beginning on, OilEd has been optimized for reasoning with description logic, and has been successfully used for various biomedical ontology projects. However, OilEd's authors never intended it as a full ontology development environment, but rather as a platform for experiments. As a result, OilEd's architecture is neither scalable to really large ontologies, nor sufficiently flexible to support customized user interface widgets. Furthermore, it suffers from a rather complicating user interface for editing logical expressions. The developers of Protégé and the OilEd team have recently joined forces in a transatlantic project called CO-ODE, which leads to a growing number of extensions for the Protégé OWL

[2] `http://jena.sourceforge.net/`

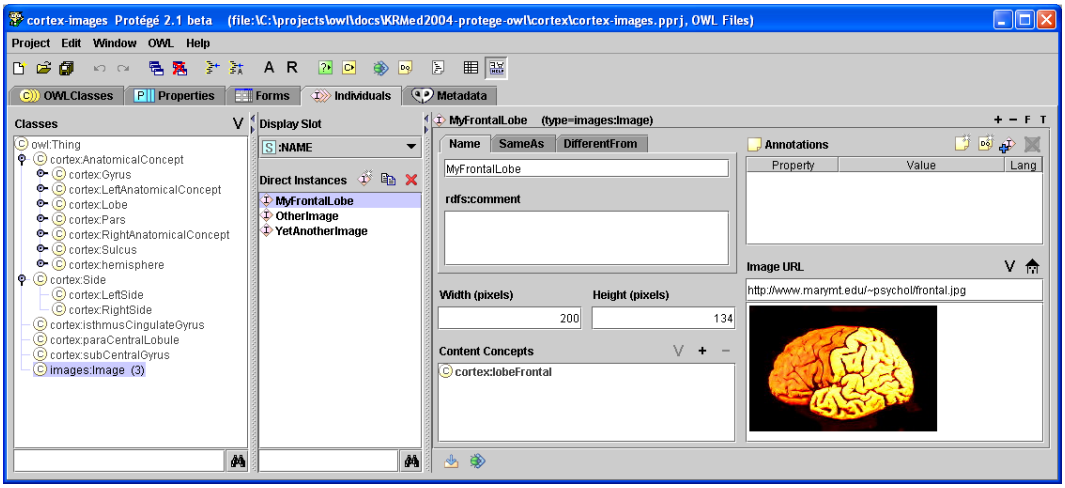


Figure 5: Protégé generates user interfaces to acquire individuals of ontology concepts. This can be used to annotate Web resources such as images for a clinical online repository.

Plugin. Many other groups from around the world are also developing Protégé plugins, including tools which can be used to edit OWL classes and relationships in a visual UML-style diagram. Other large-scale Protégé plugins are being optimized for the OWL Plugin. With its large and rapidly growing community of thousands of users, Protégé has the potential to maintain its position as one of the leading open-source ontology development environments for the Semantic Web.

Acknowledgements

This work has been funded by a contract from the US National Cancer Institute and by grant P41LM007885 from the National Library of Medicine. Olivier Dameron is funded by INRIA. Additional support for this work came from the UK Joint Information Services Committee under the CO-ODE grant. Several colleagues and students at SMI were involved in the development of the OWL Plugin, in particular Ray Fergerson and Prashanth Ranganathan. Our partners from Alan Rector's team at the University of Manchester have made very valuable contributions.

References

[1] Franz Baader, Diego Calvanese, Deborah McGuineness, Daniele Nardi, and Peter Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, 2003.

[2] Sean Bechhofer, Ian Horrocks, Carole Goble, and Robert Stevens. OilEd: a reason-able ontology editor for the Semantic Web. In *14th International Workshop on Description Logics*, Stanford, CA, 2001.

[3] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284:34–43, 2001.

[4] John H. Gennari, Mark A. Musen, Ray W. Fergerson, William E. Grosso, Monica Crubézy, Henrik Eriksson, Natalya F. Noy, and Samson W. Tu. The evolution of Protégé-2000: An environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, 58(1):89–123, 2003.

[5] Jennifer Golbeck, Gilberto Fragoso, Frank Hartel, James Hendler, Bijan Parsia, and Jim Oberthaler. The national cancer institute's thesaurus and ontology. *Journal of Web Semantics*, 1(1), 12 2003.

[6] Volker Haarslev and Ralf Moeller. RACER user's guider and reference manual. `http://www.cs.concordia.ca/~faculty/haarslev/racer`, 2003.

[7] Holger Knublauch. An AI tool for the real world: Knowledge modeling with Protégé. JavaWorld, June 20, 2003.

[8] Holger Knublauch, Mark A. Musen, and Alan L. Rector. Editing description logics ontologies with the Protégé OWL plugin. In *International Workshop on Description Logics*, Whistler, BC, Canada, 2004.

[9] Natalya F. Noy, Ray W. Fergerson, and Mark A. Musen. The knowledge model of Protégé-2000: Combining interoperability and flexibility. In *2nd International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, Juan-les-Pins, France, 2000.

[10] Natalya F. Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W. Fergerson, and Mark A. Musen. Creating Semantic Web contents with Protégé-2000. *IEEE Intelligent Systems*, 2(16):60–71, 2001.

[11] Alan Rector. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In *2nd International Conference on Knowledge Capture (K-CAP)*, Sanibel Island, FL, 2003.

[12] Alan L. Rector. Clinical terminology: Why is it so hard? *Methods Inf. Med.*, 4–5(38):239–52, 1999.

[13] Michael K. Smith, Chris Welty, and Deborah L. McGuinness. OWL Web Ontology Language Guide. `http://www.w3.org/TR/owl-guide/`, 2003.

Symbolic modeling of structural relationships in the Foundational Model of Anatomy

José L.V. Mejino, Jr. M.D., and Cornelius Rosse, M.D., D.Sc.
Structural Informatics Group, Department of Biological Structure,
University of Washington, Seattle, WA 98195

Email contact: *mejino@u.washington.edu*

ABSTRACT

The need for a sharable resource that can provide deep anatomical knowledge and support inference for biomedical applications has recently been the driving force in the creation of biomedical ontologies. Previous attempts at the symbolic representation of anatomical relationships necessary for such ontologies have been largely limited to general partonomy and class subsumption. We propose an ontology of anatomical relationships beyond class assignments and generic part-whole relations and illustrate the inheritance of structural attributes in the Digital Anatomist Foundational Model of Anatomy. Our purpose is to generate a symbolic model that accommodates all structural relationships and physical properties required to comprehensively and explicitly describe the physical organization of the human body.

Keywords: Ontology; Knowledge representation; Spatial reasoning; Mereotopology; Partonomy; Anatomy

INTRODUCTION

The main objective of the terminologies correlated by UMLS is to serve as repositories of terms that can be reused with consistency by a variety of applications.[1] In general, most of the current biomedical and educational applications are designed to present hard-coded, didactic information, or they support low-level, look-up functions with no, or at best limited, capabilities for inference. The semantic structure of today's controlled medical terminologies (CMTs) as well as of biomedical ontologies seems adequate for the needs of such contemporary applications. Next-generation applications, however, will have to incorporate increasing levels of intelligence in order to meet the demands of the evolving environment in education, biomedical research and the practice of the various health professions. Such knowledge-based applications call for the representation of much deeper and richer knowledge than that retrievable from today's CMTs and ontologies. Since most of these projects primarily target clinical medicine, they are deficient in basic science concepts necessary to support reasoning. Moreover, since relationships between concepts constitute an important dimension of knowledge, next-generation knowledge sources must model comprehensively not only the concepts but also the relationships that characterize a particular field of basic science. Therefore, there is a need to generate enabling knowledge sources at least in those domains that generalize to diverse fields of education, biomedical research and clinical practice. Anatomy is such a fundamental domain.

We are developing the Foundational Model of Anatomy (FMA)[2-4] as an evolving resource for knowledge-based applications that will require anatomical information. Our intent is that the FMA should serve as a reference ontology for biomedical informatics[4] by furnishing a representation of anatomical entities and relationships necessary for the symbolic modeling of the structure of the human body at the highest level of granularity. The FMA explicitly represents declarative anatomical knowledge currently constrained to the human species in computable form, which should also be understandable by humans. It is intended as a reusable and generalizable resource for any biomedical application that requires anatomical information.

We first give a brief account of the ontological structure of the FMA to put in perspective the modeling of structural relationships in terms of a high level scheme, which we call the Anatomical Structural Abstraction (ASA). We then describe the components of this scheme and their interactions with one another.

ONTOLOGICAL FEATURES OF THE FMA

The elements of a disciplined modeling approach for establishing the FMA, described in greater detail elsewhere,[4] consist of declared foundational principles, a high level scheme for representing anatomical concepts and relationships, and a knowledge modeling environment that implements the principles and the inheritance of definitional and non-definitional attributes. Of these elements we only comment in this paper on the high level scheme for the FMA and, in the next section, the scheme for the ASA.

The high level scheme of the FMA specifies the concept domain and scope of the symbolic model and defines its main components:

$$\boldsymbol{FMA = (AT, ASA, ATA, Mk)} \qquad (1)$$

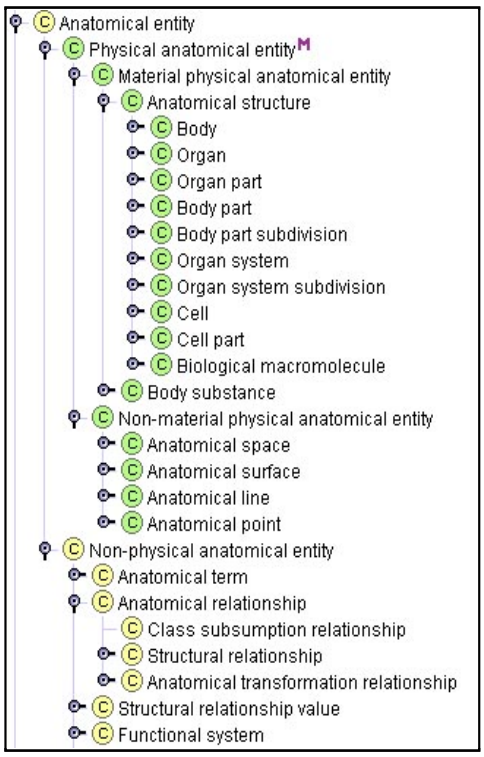


Figure 1. High level classes in the Anatomy Taxonomy (**AT**) displayed in Protégé-2000.

AT, the Anatomy taxonomy, assigns anatomical entities as class concepts in an Aristotelian-type hierarchy; **ASA**, the Anatomical Structural Abstraction, includes structural relationships among the entities represented in the **AT** and is the subject of this report; **ATA**, the Anatomical Transformation Abstraction, is based on relationships that describe the morphological and physical transformation of anatomical entities during pre- and postnatal development (not yet instantiated); and **Mk** refers to *Metaknowledge*, which comprises the principles and sets of rules, according to which the relationships are represented in the model's other three component abstractions.

Figure 1 shows a portion of the **AT** to illustrate some of its the high level classes, including anatomical relationships.

Our previous reports[2,3,5-10] are primarily concerned with the classification of physical anatomical entities (material objects, body substances, spaces, surfaces, lines and points), which constitutes the **AT**. In this communication our objective is to illustrate the importance of anatomical relationships among these entities for the symbolic modeling of structural knowledge, a dimension unique to anatomy among the biomedical sciences.

ANATOMICAL STRUCTURAL ABSTRACTION

High Level Scheme

Many treatises on mereotopology make extensive reference to human anatomy[11,12] but they all stop short of implementing in a comprehensive system the theories they propose and illustrate. Since the purpose of the FMA is to represent the physical organization (i.e., anatomical structure) of the human body, we have implemented more than a million of explicit structural relationships in the FMA. This knowledge base population task was guided by the specification of knowledge elements that describe this organization in terms of structural relationships and physical properties. We conceptualized these knowledge elements as the high level scheme of the **ASA**, which consists of two taxonomies that complement the **AT** and a number of interacting networks made up of different classes of relationships.[3,13]

$$\mathbf{ASA = (Dt, PPt, Bn, Pn, , SAn)} \qquad (2)$$

Dt, Dimensional taxonomy, is a type hierarchy which represents dimensional entities of zero to three dimensions and shape classes of 3D entities, and distinguishes between real and virtual dimensional entities. **PPt**, Physical Properties taxonomy, describes physical state properties of anatomical entities, such as mass, temperature, viscosity and density, which determine or affect the structural organization of anatomical entities. Both taxonomies are

represented in terms of which the Boundary network (**Bn**), Partonomy network (**Pn**) and Spatial Association network (**SAn**) may be described at an abstract level. Elaboration of **PPt** is beyond the scope of this paper and is discussed in the context of the symbolic representation of physiologic function as an extension of the FMA[14]. The subsequent sections explain and illustrate the interacting networks.

Boundary Network

Although parthood relationships predominate in anatomical reasoning and knowledge representation, the specification of boundaries is prerequisite for the demarcation of parts. The practical application of boundary information is critical in the segmentation of images and volumetric datasets, tasks that the FMA supports[5]. We define a boundary as a `Non-material physical anatomical entity`* of two or fewer dimensions that delimits or demarcates anatomical entities from one another that are of one dimension higher than the bounding entity. Thus the FMA specifies the `Internal surface of stomach` (a 2D entity) as the boundary of the `Cavity of stomach` (a 3D entity), as well as that of the `Wall of stomach` (3D). Should it become desirable for educational applications, for instance, to accept `Wall of stomach` as the boundary of the cavity, the appropriate modifications would need to be introduced in the particular application ontology derived from the FMA reference ontology.

We model the relationship between bounded and bounding entities by the inverse relations -*bounds*- and -*bounded by*-. The boundary network arises by a progression along the boundaries of an entity in a decreasing order of dimension: `Right ventricle` (3D) -*bounded by*- `Surface of right ventricle` (2D) -*bounded by*- `Line of right coronary sulcus, Line of anterior interventricular sulcus, Line of posterior interventricular sulcus` (1D) -*bounded by*- `Crux of heart, Apex of heart` (0D). The boundary network of the `Right ventricle`, moreover, also interacts with the **Bn** of the `Left ventricle` and `Right atrium`.

Modeling of anatomical boundaries presents a complex challenge in terms of fiat and real boundaries defined by Smith[11], which we have not yet implemented in the FMA. We distinguish between real and virtual boundaries. A real boundary of an anatomical structure corresponds to its surface, which is a `Non-material physical anatomical entity` in the AT. A virtual boundary is a `Non-physical anatomical entity`, such as the imaginary plane that demarcates the esophagus from the stomach (`Plane of gastoesophageal junction`), or the `Plane of pelvic inlet`, which demarcates the abdominal cavity from the pelvic cavity.

Partonomy Network

Although some knowledge modelers may regard an entity's boundary as a kind of parthood, we make a distinction between boundary and parthood. In the FMA, parthood relations are allowed only for entities of the same dimension. For example, `Cavity of stomach` (3D entity) -*has part*- `Cavity of pyloric antrum` (3D entities); `Internal surface of stomach` -*has part*- `Internal surface of pyloric antrum` (2D entities). Such a generic part relation suffices for describing spaces, surfaces

* Classes represented in the **AT** appear in the text in `New Courier` font.

and lines, as well as body substances (e.g., blood, semen), but greater specificity is called for when representing the parts of anatomical structures. Based on the work of Winston et al.[15] several authors have proposed a classification of parts, but cognates of the generic part relation are implemented, apart from the FMA, only in the anatomy (common reference) module of GALEN[16]. We have elaborated on such earlier proposals and developed a taxonomy of part-whole relationships[17] for guiding the representation of anatomical parts in the FMA. In addition we have defined distinct partitions for decomposing anatomical structures, and also enhanced the specificity of parthood by attributing part relations[17.]

Elaboration of Part Relations

When we address partonomy pertaining to instances of the class `Anatomical structure`, specifications must be introduced in the generic part-whole relationship because anatomical structures can be and have been decomposed based on several different contexts. The taxonomy of anatomical part relations, shown in Figure 2, illustrates such contexts. For instance, the stomach can be decomposed into its fundus, body and pyloric antrum (to name but a few of such parts), in one context and, as already mentioned, into its wall and cavity, in another context. We regard the former as a spatial partition into "*regional*" parts, whereas the latter is a compositional partition into "*constitutional*" parts. Constitutional parts are genetically determined, whereas regional parts are defined not only by genetically regulated developmental processes (e.g., lobe of lung, cortex of kidney, finger), but also by arbitrary landmarks or coordinates, such as used for demarcating the thoracic and abdominal parts of the aorta and the fundus of the stomach from adjacent parts of the corresponding wholes.

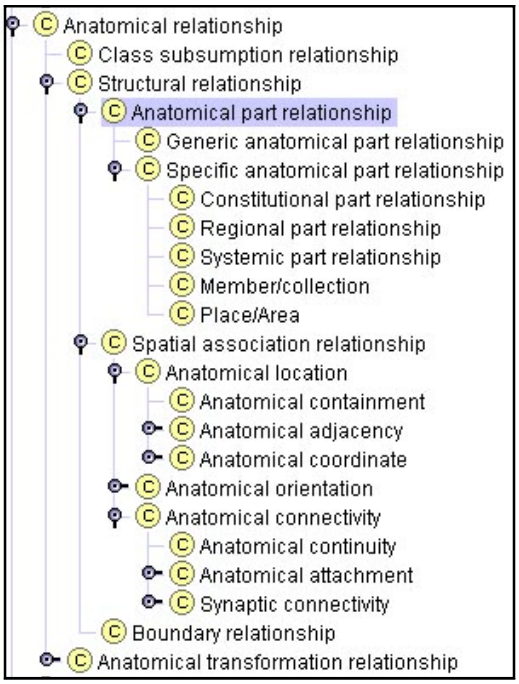


Figure 2. Classes of anatomical part-whole relationships represented in Protégé-2000.

As illustrated in Figure 3, we represent this distinction by associating the attributes *anatomical* or *arbitrary* with regional parts, and do so for anatomical structures at all levels in the **AT**. Figure 4 applies this scheme to the stomach. Furthermore, these attributes provide the basis for the different views of regional partitions, as in the case of the liver, where its traditional partition into lobes based on *arbitrary* landmarks constitutes an arbitrary kind of regional view, while another partition based on the distribution of the tributaries of the hepatic veins or branches of the hepatic artery constitutes an *anatomical* regional view. Both views, and in the case of some other organs, more than two such views, are current in clinical and educational discourse.

Although inherent 3D shape is a defining attribute of instances of the class `Anatomical structure`, the nature of continuities established between anatomical structures is such that certain parts of one structure overlap or become shared by another. The tracheobronchial tree and right and left lungs each meet the definition of `Organ`. However, since a part of the tracheobronchial tree is embedded in the right and left lungs, a distinction

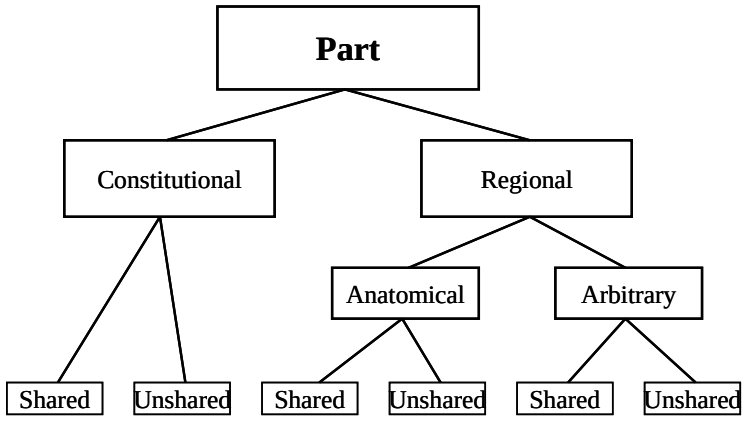


Figure 3. Taxonomy of part-whole relationships for subclasses of `Anatomical structure`.

needs to be made between the parts of the tree that are shared and unshared. Instances of the class that form branching trees (e.g., `Vascular tree`, `Neural tree`) and serous sacs (e.g., `Pleural sac`, `Peritoneal sac`) always share some of their parts with instances of another organ subclass. The attributes *shared* and *unshared* can be associated with constitutional as well as with regional parts and these attributes can specify partonomic relationships at any level of the **AT.**

Figure 3 illustrates these meronymic enhancements that are accordingly inherited by the concepts subsumed by the class `Anatomical structure`.

In our opinion, accurate and comprehensive representation of the structural organization of the body requires the level of specificity we are implementing in the FMA for partonomic relations. Indeed, all these knowledge elements are explicitly or implicitly embedded in scholarly treatises of anatomy, as well as in anatomical discourse. An ontological representation of parthood, however, also demands that clear distinctions be made between part relations and other relations, such as boundary and containment (see below).

Distinction of Part and Other Structural Relations

In addition to boundary, containment relations, included in the Spatial Association network, may also be conflated with partonomic relations. While context in natural language usually circumvents confusion and ambiguity, we believe both boundary and containment need to be distinguished explicitly in an anatomical reference ontology. Therefore we have formulated two rules, which enforce these distinctions[17].

As already illustrated in the sections on the boundary and partonomy networks, the rule of *Dimensionality Consistency* distinguishes between boundary and partonomy relationships in the FMA. The rule of *Containment/Part Distinction* constrains the *-contains-* relationship to the class

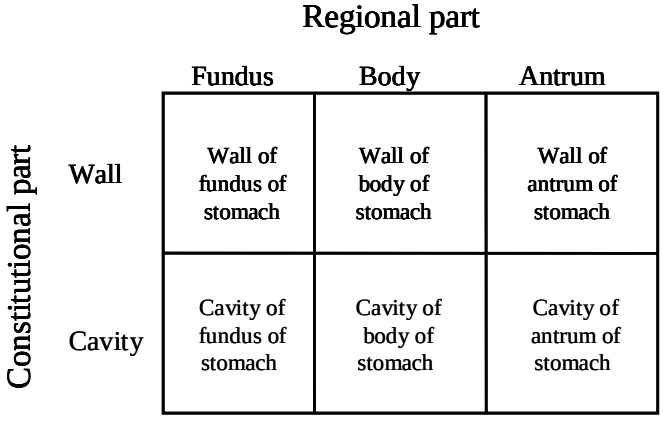


Figure 4. Table columns represent the *arbitrary* regional parts of the stomach and table rows, the constitutional parts.

Anatomical space, and its inverse, *-contained-in-*, to Body substance and Anatomical structure. Therefore, in accord with this rule, the following are valid assertions: Tibialis anterior *-contained in-* Anterior compartment of leg; Anterior compartment of leg *-part of-* Leg; Tibialis anterior *-part of-* Leg. Although this example suggests transitivity across containment and part relations, another example negates such an assumption: Urine *–contained in-* Cavity of urinary bladder; Cavity of urinary bladder *–part of-* Urinary bladder; but Urine *-part of-* Urinary bladder is an invalid assertion. Thus, in anatomical context, keeping containment and part relations independent of one another, serves the purpose of specificity and clarity.

Spatial Association network

In addition to boundary and parthood, the FMA also represents topological relationships that are important for describing the structure of the body. These relations constitute the Spatial Association network (**SAn**) component of the ASA, which itself consists of a number of subnets corresponding to the descendants of the Spatial association relationship class shown in Figure 2. The descendants of this relationship class represent three topological axes or viewpoints in terms of which anatomical spatial associations may be conceptualized:

SAn = (Location, Orientation, Connectivity) *(3)*

Location. Topology deals extensively with location, and the relation *-has location-* is used ubiquitously to describe the positioning of not only anatomical structures relative to one another, but also to associate disease processes with anatomical entities that they affect (e.g., hepatitis *-has location-* liver). However, the modeling of the structural arrangement of anatomical entities in the body calls for greater specificity. Therefore the relation *-has location-*, as such, is not used in the FMA at all; rather it serves as the type for three specific location relationships, which are explicitly implemented in the model (Figure 2). We specify location relationships between anatomical entities as *Containment*, *Adjacency* or *Qualitative coordinate*. For the current purpose enough has been said about containment in relation to its conflation with the part relation; here we elaborate on adjacency and qualitative coordinates.

Adjacency. We consider anatomical entity A to be adjacent to entity B if A and B have no overlapping (shared) boundaries and parts, and no other anatomical entity is interposed between them. The adjacency relationship is symmetrical and is valid for entities of the same dimension. Using an example first as an approximation to illustrate the relationship: lung *-adjacent to-* diaphragm; inferior surface of lung *-adjacent to-* superior surface of diaphragm. The modeling in the FMA is more accurate than this assertion implies; it takes into account the interposition of the pleural sac between the lung and the diaphragm: Right lung *-surrounded by-* Right pleural sac; Basal part of right pleural sac *-adjacent to-* Basal part of right lung, Right dome of diaphragm.

The example illustrates a number of challenges for modeling adjacency relationships: 1. Adjacency may be viewed at different levels of granularity in different contexts: the first approximation hides a number of inaccuracies and ontological inconsistencies, although it may be acceptable for the representation of anatomical knowledge at an elementary and crude

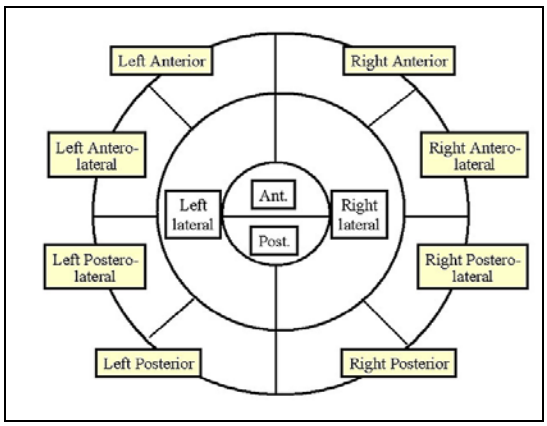


Figure 5. Qualitative radial coordinate system for the **Dt** shape class 'conventional cylinder'.

level; the second one describes the arrangement of the related entities without ignoring elements of reality that may not be meaningful to some users, and this is the objective of the FMA; 2. Adjacency assertions must be constrained to anatomical entities subsumed by the same **AT** subclasses of `Anatomical structure`, which specify levels of structural organization: `Biological macromolecule`, `Cell part`, `Cell`, `Tissue`, `Organ part`, `Organ`, which correspond to the granular partitions of the body proposed by Bittner and Smith[18]. Thus, organs: `Right lung`, `Right pleural sac`; organ parts: `Basal part of right pleural sac`, `Basal part of right lung`, `Right dome of diaphragm`; 3. Adjacency relationships must be qualified by such descriptors as -*surrounded by*- and its inverse -*surrounds*-, or by qualitative anatomical coordinates that describe vectors of directionality, illustrated by the following example.

The esophagus, or a part of it, inherits its shape from the **Dt** class `Conventional hollow cylinder`. This shape specifies the set of adjacency relationships that is allowed for this shape class. Figure 5 shows these relationships graphically in terms of a qualitative radial coordinate system. In Figure 6 the qualitative coordinate system for cylinder is superimposed and centered on the esophagus in a section of the male Visible Human at the level of the eighth thoracic vertebra. In Figure 7 the adjacencies of `T8 part of the esophagus` are represented symbolically in terms of these qualitative coordinates. Although some of these adjacency relationships remain constant, others change from one vertebral level to the next. The AT of the FMA represents each vertebral level of the esophagus as a discrete subzone, which permits the symbolic modeling of the changing adjacency relationships of the esophagus as it "passes" from the neck to the abdomen.

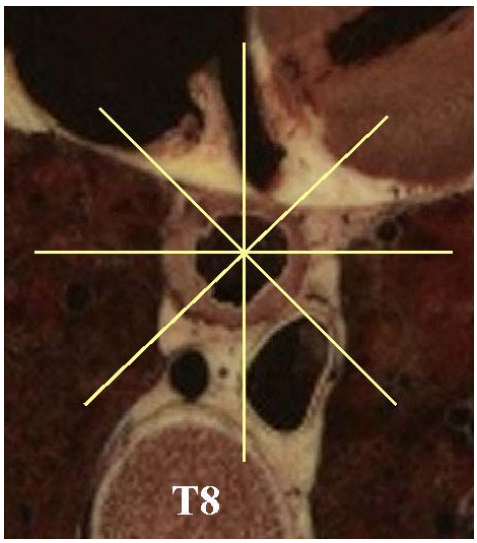


Figure 6. Coordinate system of conventional cylinder superimposed on T8 part of esophagus.

It deserves mention that the qualitative coordinates anterior, posterior, lateral, mentioned in Figures 5 and 7, as well as others (e.g., superior, inferior) are standard directional terms defined in relation to the orientation of the body in the so called "*anatomical position*"; they remain constant regardless of the position the body assumes.

The spatial knowledge captured by the adjacency relationships shown in Figure 7 is of importance to a student dissecting the esophagus for the first time and also to a surgeon planning to remove a lymph node adjacent to the esophagus through a mediastinoscope. The FMA can provide knowledge of adjacency relationships appropriate for applications developed for each of these types of users. Moreover,

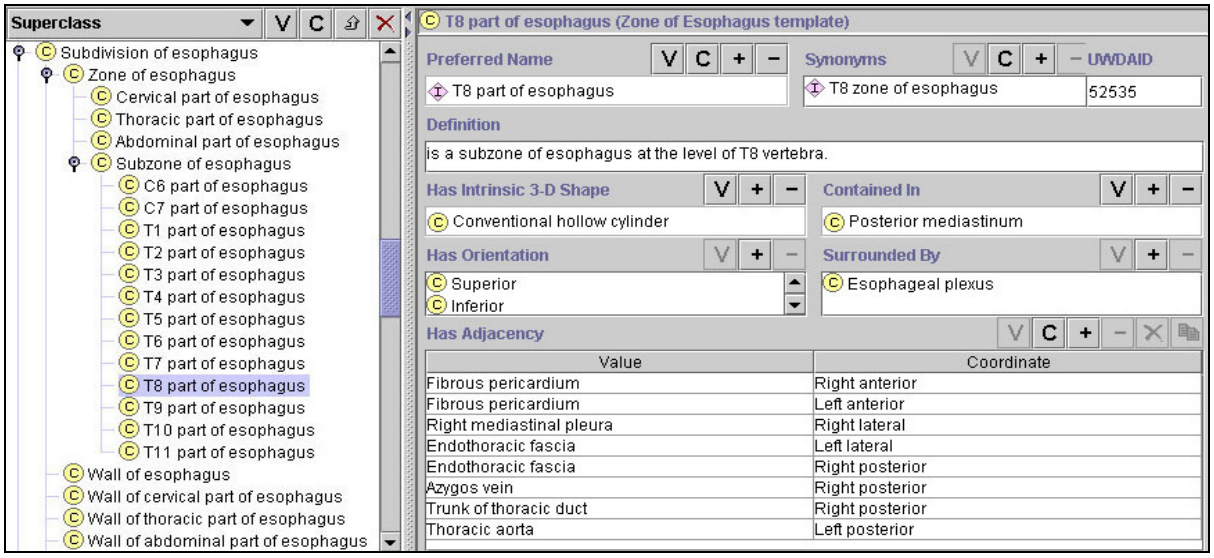


Figure 7. Frame-based representation in Protégé-2000 of `T8 part of esophagus` in **At** in the left pane and its attributes in the right pane.

since we can represent inverse values for these relationships, and make inferences based on their transitivity, the FMA could support inference required for answering user-generated spatial queries at different levels of complexity.

Figures 5 and 6 invite comment about the relative usefulness of geometric and qualitative coordinates for representing such structural attributes as location and adjacency. The relationships expressed in terms of qualitative coordinates could be derived from the quantitative geometric matrix of the Visible Human data set, for example. These geometric coordinates, however, would have to be expressed as qualitative coordinates in order to make them intelligible in anatomical discourse. Geometric coordinates are valid only for one instance, whereas anatomical qualitative coordinates describe relationships that hold true in all members of a species. Only those structures can be referenced by geometric coordinates that are visible with a particular imaging modality. Qualitative coordinates, on the other hand, can describe the relationship of invisible structures to visible ones, as illustrated in Figure 7 by the esophageal plexus, fibrous pericardium and mediastinal pleura; none of these structures can be identified in the image of the anatomical section. Moreover, inference required for reasoning about structural relationships within the body must make use of qualitative coordinates. Therefore, the symbolic representation of location relationships in terms of qualitative coordinates is an important component of the FMA.

In summary, location of an anatomical structure may be described in terms of containment (e.g., `Right lung` -*contained in*- `Right half of thoracic cavity`); adjacency (e.g., `Right lung` -*surrounded by*- `Right pleural sac`) and qualitative anatomical coordinates, such as those illustrated for `T8 part of the esophagus.`

Orientation. Since a defining attribute of entities subsumed by the class `Anatomical structure` is inherent shape, their orientation within the body can be specified, largely in terms of shape and the qualitative coordinates of their parts or boundaries that demarcate them from other structures. Figure 8 illustrates orientation information entered in the FMA for the `Esophagus`, the shape of which is the dimensional entity `Hollow cylinder`. The

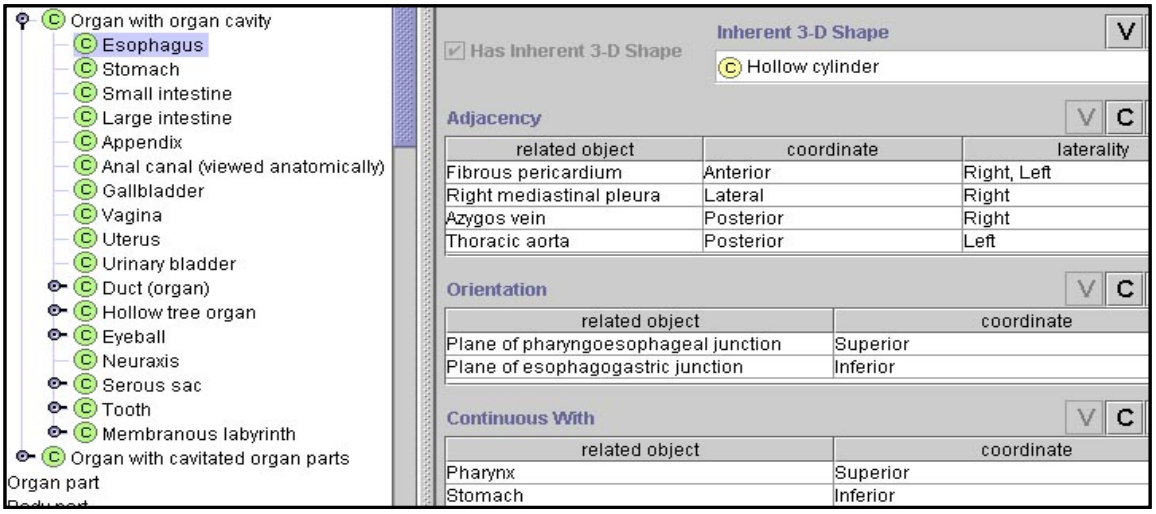


Figure 8. Spatial Association network (SAn) slots *–adjacency-*, *-orientation-* and *–continuous with-* in the frame of the `Esophagus` displayed in Protégé-2000.

orientation of the esophagus is defined by the virtual `Plane of pharyngoesophageal junction` and `Plane of gastroesophageal junction`, which demarcate the esophagus from the pharynx and the stomach respectively. The orientation of the esophagus is specified by the qualitative coordinates superior and inferior for these two planes, respectively, which serve as coordinate and vector reference in the context of the human anatomical position. In other instances, it is necessary to declare right or left laterality coordinates. For example, in describing the orientation of the cone-shaped `Heart`, we use `Apex of heart` and `Base of heart` as the entities of reference and specify their location by qualitative coordinates (inferior and left lateral for the apex and posterior for the base). Orientation is treated much less specifically in conventional anatomical discourse than in geometric modeling. However, there is a need for coordinating symbolic modeling in the FMA with geometric modeling and this will require, for example, that we define axes of anatomical structures for specifying orientation also in the FMA.

Connectivity. Among anatomical structures only cells floating free in blood and other body substances or locked in the lacunae of hyaline cartilage can be considered unconnected to other structures. Even cells that move about in loose connective tissue, or on epithelial surfaces, or through epithelia form adhesions with the substrates on or through which they move. With the few notable exceptions, all anatomical structures are connected to one another through a variety of continuities and junctions. Connections exist horizontally and vertically across all levels of structural organization or granular partitions, which accounts for the material integrity of the human body or that of any biological organism. Perhaps the greatest attention has been paid to inter- and intracellular junctions, which, like junctions at higher levels, have a specific structure that distinguishes them from one another. Therefore in the FMA, we classify these junctions as anatomical structures, rather than relationships. In this section we are concerned with the connectivity relationship, rather than the material entities that establish the physical connection between two or more structures.

As in the case of location, we consider connectivity a relation type or class and explicitly implement in the FMA only its cognates: *Continuity*, *Attachment* and *Synaptic connectivity*.

Continuity. We regard continuity as a symmetrical connectivity relationship between two or more anatomical entities asserted by the relationship *-continuous with-*. We regard A as *-continuous with-* B if no real boundary exists between corresponding constitutional parts of A and B. For example, in these terms, continuity exists between a main arterial, venous and nerve trunk on the one hand, and their respective branches on the other. We also sanction the assertion `Esophagus` *-continuous with-* `Stomach`, because constitutional parts of their wall (mucosa, submucosa, muscularis) are not demarcated by a real boundary. `Esophagus` and `Stomach` qualify as different organs because of the distinct structural attributes they exhibit in terms of shape and the characteristic arrangement of their constitutional parts (the structure and morphology of their mucosa and organizational pattern of muscle layers in their wall).

As illustrated in Figure 8, we attribute each continuity relationship with a qualitative coordinate, in order to distinguish continuities with more than one structure. Such attributed continuities also need to be declared between regional parts of an organ, which may or may not be associated with a structural change in the constitutional parts of its different regions. For example, we need to assert that continuity exists between the fundus and the body of the stomach, but there is no continuity between the fundus and the pyloric antrum, all of which are regional parts of the stomach. The FMA does not accommodate negation or disjunction; therefore the lack of continuity with an entity must be inferred from its absence among the values of the *-continuous with-* slot in the frames of two entities.

Continuity between arbitrary regional parts of an anatomical structure may be taken for granted. However even such continuities need to be explicitly represented, since it needs to be asserted that the thoracic part of the esophagus is continuous *superiorly* with its cervical part, and continuous *inferiorly* with the abdominal part of the esophagus. Listing continuities without their attributes would omit an element of structural knowledge.

The FMA also represents continuities between anatomical spaces, surfaces and lines as well as between anatomical structures. The modeling of these continuities, however, presents less of a challenge than that of anatomical structures.

Attachment. We regard attachment as an asymmetrical connectivity relationship between two or more anatomical entities asserted by the inverse relationships *-attached to-* and *-receives attachment of-*, which are constrained to selected subclasses of `Anatomical structure`. We regard A as attached to B, and B as receiving the attachment of A, if A and B are subsumed by different subclasses of `Anatomical structure` and if A intermingles at least one of its constituent parts with a constituent part of B. For example, the patellar ligament [subclass of `Ligament(organ)`] is attached to a narrow area along the lower margin of the patella and to a tuberosity at the upper end of the tibia [the two bones are subsumed by subclasses of `Bone(organ)`]. All these anatomical structures have their own real boundaries, but at its proximal and distal ends, the stout ligament comes into intimate contact with circumscribed areas of each bone, where extensions of its collagen fiber bundles (so called Sharpey's fibers) penetrate the bone and intermingle with each bone's own matrix. The ligament may be separated from the bone only by severing Sharpey's fibers.

Similar attachments occur between membranes and bones (e.g., the circumference of the tympanic membrane is attached to bones of the skull forming the external auditory meatus), membranes and viscera (e.g., visceral pleura is attached to the lung proper intermingling its loose connective tissue on its non-serous surface with the fibrous stroma of the lung), and also between muscles and bones.

Muscle attachments are qualified with respect to whether the bone to which they attach moves or remains stable in the normal course of the muscle's action. Therefore, each site of a muscle's attachment is attributed as either the origin or the insertion .

Synaptic connectivity. We regard synaptic connectivity as a specialized attachment relationship occurring in neural and neuromuscular synapses. It is also implemented as an attributed relationship that identifies the connection between the parts of synapsing structures like the axon and the dendrite or the neuromuscular junction.

The included figures which illustrate various relationships that in aggregate constitute the ASA are all based on Protégé-2000, the frame-based ontology authoring and editing environment[19]. The next section enlarges on aspects of this implementation, which is a critical element of the disciplined modeling process through which we have and continue to populate the Foundational Model of Anatomy.

IMPLEMENTATION

We consider the evolution of the FMA from an earlier controlled vocabulary and elaborate in some detail about the representation of attributes and relationships using the Protégé-2000 modeling environment.

UWDA and FMA. In its initial iteration the FMA was called the University of Washington Digital Anatomist (UWDA) vocabulary and was developed as an anatomical enhancement of UMLS[1]. Populating the UWDA we were less concerned with the richness of anatomical relationships than with the comprehensiveness of the classification of anatomical entities. The authoring tool we developed was designed to generate parallel hierarchies (directed acyclic graphs) based on *is-a*, *part-of*, *branch-of* and *tributary-of* relationships. As we populated subclasses of `Organ part` in the *is-a* hierarchy, for example, we also aligned the concepts along the transitive *part-of* relationship in another hierarchy. However, such a link-centric view and representation of anatomy proved to be inadequate once we began to appreciate the complexity of relationships that were necessary for comprehensively describing the anatomy of the body. The need for such a comprehensive, reusable resource led to formulating the FMA as an ontology of the physical organization (structure) of the human body.

Close to 70,000 FMA concepts are still accessible through the UWDA vocabulary of UMLS, providing a comprehensive controlled terminology for macroscopic, microscopic and neuro-anatomy. Our current work entails the instantiation of the **ASA** networks of these concepts. The association of such multi-dimensional relationships with anatomical concepts called for a node-centric view of anatomy, which was beyond the capacity of the link-centric representation we implemented. The frame-based knowledge acquisition system Protégé-2000[19] has the requisite expressivity and scalability for comprehensively modeling anatomical relationships encompassed by the **ASA**. The same will be true for **ATA** relationships, once we begin the implementation of developmental transformations.

Modeling the ASA in Protégé-2000

Protégé-2000 has been adapted to meet current and evolving needs of the FMA[19]. It is being enhanced by customized active user-interface components as we encounter new challenges in modeling[20].

We regard the FMA as an ontology of concepts and relationships which are represented as frames in Protégé-2000. These frames are data structures, which, through their slots, specify the types of information to be associated with a concept in the **AT**. The values

for some of these slots are derived from the **AT** and others from two additional taxonomies: the Dimensional taxonomy (**Dt**) and Physical Properties taxonomy (**PPt**). A fourth taxonomy, the 'Anatomical entity metaclass' hierarchy assures the selective inheritance of the attributes of the entities represented in the **AT**. The 'Anatomical entity metaclass' hierarchy provides templates for all the **AT** classes. Each template is a frame composed of a set of slots; each slot corresponds to a defining or associative attribute manifested by the entities subsumed by a particular **AT** class. The templates become elaborated by new attributes that are introduced as slots when a new class in the **AT** subsumes entities that exhibit the new attribute.

The frames of **AT** classes are assigned as instances of metaclasses (or templates) and therefore inherit the templates slots of their respective metaclasses, These slots now become own slots of the instances of classes, the values of which are unique to the instances.

DISCUSSION

The Foundational Model of Anatomy is the largest and most comprehensive ontology for the anatomy domain, which encompasses in one continuous information space anatomical structures at all levels of biological organization from macromolecules to cells, tissues, organs, organ systems and body regions. Our purpose in this communication is to illustrate the implementation of a theory expressed by the high level schemes of the FMA and its **ASA** component. This theory concerns the computable symbolic representation of the structural and topological arrangement of the body's constituents. We have emphasized the critical role such relationships play in the modeling of this arrangement. They provide the basis on which spatial reasoning (inference) can be supported[21,22].

The FMA continues to evolve, in particular through the instantiation of its **ASA** component, the main topic of this communication. Although the FMA and **ASA** model a broad segment of declarative structural knowledge in great detail, there remain numerous gaps that must still be filled and other areas that must be refined. However, we consider the most significant feature of the FMA to be not so much its contents as its semantic structure. This structure, reflected in the high level conceptualization coupled with the practical implementation of the ontology, was established through an evolving disciplined approach to populating the knowledge base[4].

A salient feature of our approach is the deliberate constraining of the modeling to a structural context. Structure provides the foundation for all other types of biological information. We believe that the logical and consistent organization of biological structure is a prerequisite for the representation of other biological fields. Therefore we regard the FMA as a *reference ontology* for biological structure. By this assertion we mean that in its "native" format the FMA may not precisely meet the needs of any particular user group. However, developers of applications designed to address particular problems and tasks should be able to filter and derive from the FMA the anatomical information they need. With this motivation in mind, we provide access to the FMA through the Internet and make it available to those whose need for anatomical information goes beyond the mere reuse of anatomical terms.[23]

We believe that even more important is the role the FMA can play as a reference ontology for other disciplines and domains by providing a template for other symbolic models. First examples of such a use of the FMA are the anatomy of non-human species[24] and physiological function[14]. It is our hope that ontology developers in other domains will follow.

Acknowledgment
This work was supported in part by contract LM03528 and grant LM06822, National Library of Medicine and the DARPA contract Virtual Soldier Project.

REFERENCES

1. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32:281-91.
2. Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and organizational principles for anatomical knowledge representation: the Digital Anatomist symbolic knowledge base. J Am Med Inform Assoc 1998;5:17-40.
3. Rosse C, Shapiro LG, Brinkley JF. The Digital Anatomist Foundational Model: principles for defining and structuring its concept domain. Proc AMIA Symp 1998;820-4.
4. Rosse C, Mejino JLV. A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics* 2004.
5. Brinkley JF, Wong BA, Hinshaw KP, Rosse C. Design of an anatomy information system. IEEE Comp Graphics Applic 1999;3:38-48.
6. Michael J, Mejino JLV, Rosse C. The role of definitions in biomedical concept representation. Proc AMIA Symp 2001; 463-467.
7. Martin RF, Mejino JLV, Bowden DM, Brinkley JF, Rosse C. Foundational model of neuroanatomy: its implications for the Human Brain Project. Proc AMIA Symp 2001; 438-442.
8. Mejino JL, Rosse C. The potential of the Digital Anatomist Foundational Model for assuring consistency in UMLS sources. Proc AMIA Symp 1998;825-9.
9. Mejino JL, Rosse C. Conceptualization of anatomical spatial entities in the Digital Anatomist Foundational Model. Proc AMIA Symp 1999;112-6.
10. Agoncillo AV, Mejino JL, Rosse C. Influence of the Digital Anatomist Foundational Model on traditional representations of anatomical concepts. Proc AMIA Symp 1999;2-6.
11. Smith B. Mereotopology: a theory of parts and boundaries. Data & Knowledge Engineering 1996;20:287-303.
12. Schulz S, Hahn U. Mereotopological reasoning about parts (w)holes in bio-ontologies. In Proceedings of FOIS'01 New York: ACM Press, 2001. P. 198-209.
13. Neal PJ, Shapiro LG, Rosse C. The Digital Anatomist structural abstraction: a scheme for the spatial description of anatomical entities. Proc AMIA Symp 1998;423-7.
14. Cook DL, Mejino JLV Jr, Rosse C. Evolution of a Foundational Model of Physiology: Symbolic Representation for Functional Bioinformatics**.** To appear in Proceedings of MedInfo 2004.
15. Winston ME, Chaffin R, Herrman D. A taxonomy of part-whole relations. Cognitive Sci. 1987; 11:417-444.
16. Rogers J, Rector A. GALEN's model of parts and wholes: experience and comparisons. Proc AMIA Symp. 2000:714-718.
17. Mejino JLV Jr, Agoncillo AV, Rickard KL, Rosse C. Representing complexity in part-whole relationships within the Foundational Model of Anatomy. Proc AMIA Symp 2003;450-454.
18. Bittner T, Smith B. A theory of granular partitions. In: Foundations of geographic information science, London: Taylor & Francis, 2003.

19. Noy NF, Fergerson RW, Musen MA. The knowledge model of Protégé-2000: combining interoperability and flexibility. In: Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW-2000) 2000. Juan-les-Pins France,Springer.
20. Noy NF, Mejino JLV, Musen MA, Rosse C. Pushing the envelope: challenges in frame-based representation of human anatomy. Data & Knowledge Engineering. In Press.
21. Distelhorst G, Srivastava V, Rosse C, Brinkley JF. A prototype natural language interface to a large complex knowledge base, the Foundational Model of Anatomy. Proc AMIA Symp 2003;200-204.
22. Detwiler LT, Chung E, Li A, Mejino JLV, Agoncillo AV, Brinkley JF, Rosse C, Shapiro LG. A Relation-Centric Query Engine for the Foundational Model of Anatomy. In *Proceedings, MedInfo 2004*, San Francisco, CA. To appear.
23. Foundational Model of Anatomy. http://fma.biostr.washington.edu/.
24. Travillian RS, Rosse C, Shapiro LG. An approach to the anatomical correlation of species through the Foundational Model of Anatomy. Proc AMIA Symp 2003;669-673.

Towards a Computational Paradigm for Biomedical Structure

Stefan Schulz
Department of Medical Informatics
Universitätsklinikum Freiburg
Stefan-Meier-Str. 26, D-79104 Freiburg, Germany
stschulz@uni-freiburg.de

Udo Hahn
Text Knowledge Engineering Lab
Universität Freiburg
Werthmannplatz 1, D-79085 Freiburg, Germany
hahn@coling.uni-freiburg.de

Abstract

The symbolic representation of the physical structure of living organisms needs an ontologically well-founded and logically sound approach so that formal reasoning can adequately be supported. We describe a set of canonical relations and attributes necessary for the description of biological structures. Based on these epistemological primitives, we sketch how a broad range of organisms can be represented by cascading theories which are ordered by various dimensions, such as granularity, development, species and canonicity. We thus aim at a rational reconstruction and non-redundant representation of biological structure notions.
Keywords: *Biological Ontologies*

Introduction

Formally founded descriptions of the physical composition of biological entities have attracted increasing attention in the last few years, as their pivotal role in biomedical ontologies has been increasingly recognized [2, 13, 16, 14].

In order to achieve a comprehensive formal representation of living systems, the first step would be to construct a multi-purpose reference ontology of *biological structure*. Such an approach should ideally cross the boundaries between species, because even organisms with largely different phenotypes show surprising similarities at a genetic level. Hence, knowledge about one organism should be re-usable in order to understand other organisms [19]. In terms of sheer coverage, a large amount of biomedical terms are already represented by the UMLS [18], the Gene Ontology [5] and a continuously increasing number of "anatomies", developed within the Open Biological Ontologies (OBO) framework. [17]. However, all of these systems are committed to a highly selective view of biological structure in terms of developmental stages, granularity and species-specific structure. Each species anatomy is being built from scratch, although the rough architecture of organisms exhibits considerable similarities between species and developmental stages.

Focusing on the anatomy of the heart, Fig. 1 shows a synopsis of several OBO models, together with the Foundational Model of Anatomy (FMA) [16, 15]. Abstracting away from terminological differences (e.g., *circulatory system* vs. *cardiovascular system*), we recognize a number of commonalities between diverse organisms. For instance, the *heart* is always part of the *circulatory system*. Except in the case of flies and in the early developmental stages of the mouse, *hearts* have *chamber(s)* and *valves*. The difference between *heart atriums* and *ventricles* exists in fish as well as in mice and humans.

With the exception of the FMA, which is based on strict principles and is moving towards a formally founded redesign, the anatomies of the other species, as well as the (theoretically) species-independent Gene Ontology, are no more than controlled vocabularies with thesaurus-like relations, which in some cases do not even make consistent use of the *part-of* relation and provide largely incomplete taxonomic links [1]. Consequently, the decision as to whether a deduction such as *cell has-part nucleolus* is valid or not and how it should be interpreted, assuming a model which asserts, e.g., *cell - has-part - cell nucleus* and *cell nucleus - has-part - nucleolus*, is left to the user, because there is no ontological commitment to either the algebraic properties of *has-part*, e.g. transitivity, or the dependency status of *has-part cell nucleus* (Is every nucleolus part of a cell nucleus, or does every cell nucleus have a nucleolus as part?).

This may be acceptable when the use of these vocabularies is limited to manual, expert-level gene annotation or document retrieval tasks. However, anticipating their use for knowledge-intensive ap-

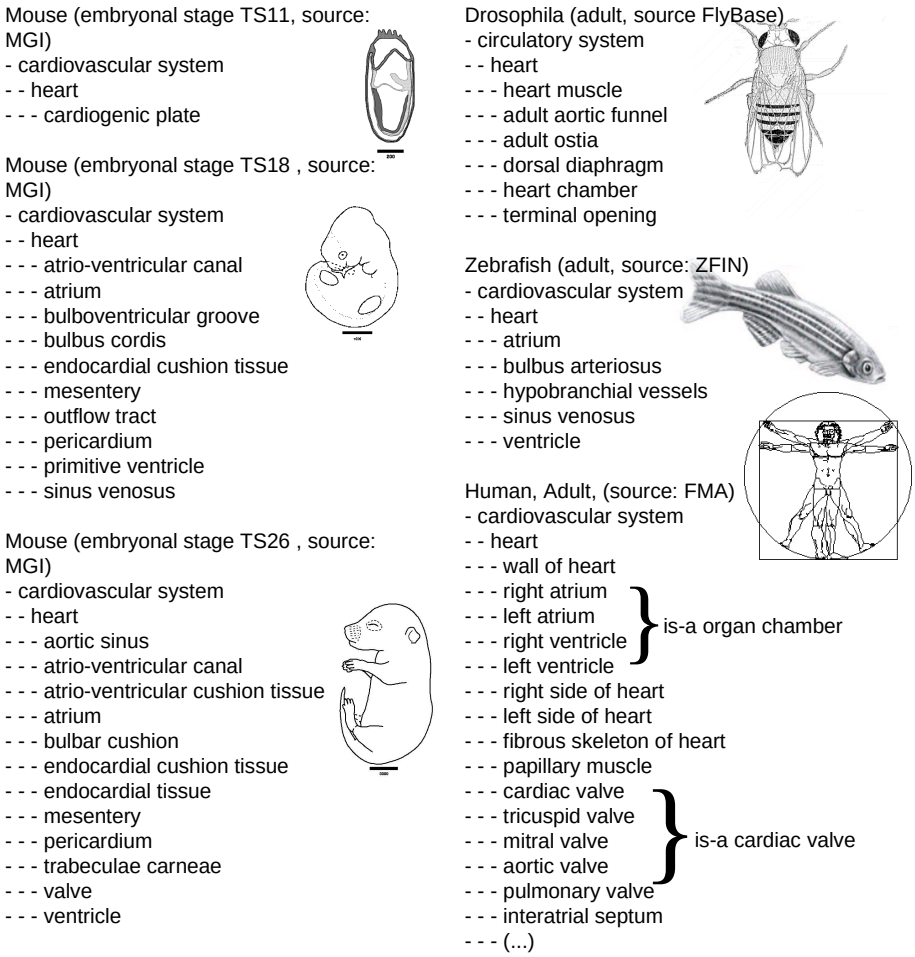


Figure 1: Comparative Heart Anatomy (only part-whole links) from OBO Biological Ontologies and the Foundational Model of Anatomy

plications, the informal approach leads to an impasse, because semantically vague, and even inconsistent assertions about concepts may cause a broad range of arbitrary invalid and, thus, unwarranted deductions.

Therefore,we argue for a domain description in terms of a set of formal axioms which allow valid and correct inferences. Complex concept descriptions built from a set of formally founded conceptual relations may be a good starting point for a formally adequate treatment of biological structures. In the following sections, we focus on various aspects of conceptual modeling of biological structure in its broadest sense, aiming at a multi-purpose foundational ontology.

Relations and Attributes

An ontological analysis of any domain should be guided by generally shared principles. According to Gangemi *et al.* [4], this first requires selecting a set of foundational (formal) relations, then defining the ground axioms for these relations, establishing constraints across basic relations and defining a set of formal properties induced by these formal relations. Then a set of basic categories is introduced, and the relevant kinds of domain entities (concept classes and instances) are classified according to the basic categories. Finally, the dependencies and interrelations among basic categories have to be studied. In this paper, we limit ourselves to an overview of adequate foundational relations and attributes. Rather than proposing a single canonical formalism, we outline alternative axiomatizations, their consequences and intricacies. Our selection of relations comprises some of the (informal) relations provided by the is comprised of some of the (informal) relations provided by the UMLS semantic network, completed by additional ones, considered relevant for describing biological structure.

Foundational Relations

Domain entities can be ordered according to (strict) partial orders which are characterized by a set of ordering relations. Strict partial orders are transitive, antisymmetric and irreflexive, whereas partial orders are transitive, antisymmetric and reflexive. Additional constraints may stem from type restrictions on the domain and the range of a relation. Of paramount importance is the distinction between classes (universals, concepts, sorts of things, e.g., *"Left Hand"*) and individuals (particulars, instances, concrete things in the world, e.g., *"my left hand"*). Because we have found that relations (such as *part-of*) are commonly asserted between concept classes – and have therefore a different semantics than their cognate relations between individuals – we stick to the following naming convention: lower case relation names are used for relations between individuals, and upper case names characterize relations between concept classes. Accordingly, we write concept (class) names with upper case initials, and instance names with lower case initials.

Taxonomy The taxonomic *Is-A* relation, a partial order, [26] relates specific classes to conceptually more general classes, e.g. *Mitral Valve Is-A Atrioventricular Valve* or *Alanin Is-A AminoAcid*. More specific classes inherit all properties from more general classes. The definition of a class illuminates its distinctive characteristics in relation to already defined (more general) classes, following the Aristotelian principle of *genus* and *differentiae*. Whereas the genus assigns an entity to a class, the differentiae distinguish the entity from other entities also assigned to that class. For example, *Left Hand* has *Hand* as its genus and its laterality attribute *left* as differentiae. Taxonomies can have either a monohierarchical (single parent), or a polyhierarchic (multiple parent) structure. In the Foundational Model of Anatomy (FMA) [16, 11], e.g., huge taxonomies are represented as strict monohierarchies. The relation *Is-A* must not be mixed up with the relation *instance-of* which relates individuals with the classes they belong to, e.g., *my left hand instance-of Left Hand*. Unfortunately, the relation *instance-of* is often used inadequately in biomedical ontologies, e.g. *Muscle System instance-of Organ System* in the FlyBase vocabulary.

Mereology At least for the life science domain, not only taxonomic relations (*Is-A* and *instance-of*) but also mereological relations (basically, *part-of* vs. *has-part*) are of outstanding and equal importance for the design of any ontology describing biological structure. In classical (i.e., axiomatic) mereology [22, 3] generic parthood is treated as a partial order. Common conceptualizations in the biological domain, however, suggest that the assumption that *part-of* be reflexive must be abandoned.[1] The most obvious distinction between *Is-A* and *part-of* relates to the fact that the first one is maintained between classes, whereas the second one is maintained only between individuals.

[1] Otherwise, any instance of "stomach" would be an instance of "stomach part", with the consequence that the class "*partial* resection of stomach" would include "*total* resection of stomach".

As an example, *my left thumb* is *part-of my left hand*, but the class *Thumb* is certainly not *part-of* the class *Hand*. However, "being part-of a hand" is a property of any instance of *Thumb*. This distinction has been largely ignored in our domain. As a result, the meaning of mereological relations asserted between a pair of concepts, such as *Part-Of(CellNucleus, Cell)*, is ambiguous, allowing the possibility for conflicting interpretations to evolve: The Gene Ontology [5] interprets *Part-Of* as "can be a part of, not is always a part of" which frequently leads to unexpected conclusions [25]. In contrast, the Foundational Model of Anatomy (FMA) [16] conceptualizes *part-of* in a very strict manner: *A Part-Of B* means that any instance of *B* has an instance of *A* as part, and any instance of *A* is part of an instance of *B* [23]. This interpretation imposes a mutual dependency between parts and wholes and, therefore, may be too rigid in many cases. For example, we may want to express that any instance of a cell nucleus is part of a cell, but certainly not any instance of a cell has a cell nucleus. Certainly, we also may want to instantiate non-standard organisms which lack certain body parts. As far as other models of organisms referred to in the introductory section are concerned, especially mouse, zebrafish and drosophila anatomy, there is no commitment at all to the proper semantics of *Part-Of*.

A mereological relation between concepts (classes of individuals), therefore, cannot be interpreted unambiguously, unless we make clear statements on the existence of a whole with respect to its parts, as well as the existence of a part with respect to its whole. Taking into account the (supposed) intended meaning of mereological relations between concepts, we define, similar to [23], *Part-Of* and *Has-Part* on the basis of *part-of* and *has-part*, using *inst-of* as the membership relation between an individual and a class:

$$Part\text{-}Of(A,B) =_{def} \forall x : inst\text{-}of(x,A) \Rightarrow \exists y : inst\text{-}of(y,B) \wedge part\text{-}of(x,y) \quad (1)$$

$$Has\text{-}Part(A,B) =_{def} \forall x : inst\text{-}of(x,A) \Rightarrow \exists y : inst\text{-}of(y,B) \wedge has\text{-}part(x,y) \quad (2)$$

Location The locative relation [3], characterized by the relation pair *location-of* vs. *has-location*, is another partial order between individuals. It relates a spatial entity with another spatial entity or a material object, e.g., *brain has-location cranial cavity*. Wherever locative relations are asserted between concept classes, we define *Location-Of* and *Has-Location* similar to *Part-Of* and *Has-Part* in Formula (2) and (3). A crucial decision is whether to keep mereological and topological aspects separated, or to subscribe to a more simplified mereotopological view in which spatial objects coincide with the region they occupy. As an example, is a *bacterium* after being ingested by a *cell* (e.g., a *macrophage*) part of this *cell*? If not, do its components (e.g., molecules) become parts of the original structure after decomposition? Without any doubt, both the bacteria and its components are *located* within that cell. Similarly, is a hollow space a part of its host or part of the exterior space (cf. [20])? Is a boundary a part of the entity it bounds?

In a restricted domain such as biology, the distinction between mereology and topology may seem arbitrary and inconsistent. Here, *Part-Of* may imply *Has-Location*, and connection can be expressed in terms of mereology [3]. In this case, mereological relations would be mere subrelations of locative ones [21]. This may, however, complicate the conceptualization of detached parts, which one could still consider to be included in the notion of part. For example, a metastasis of a tumor may still be considered a part of the primary tumor which is, however, not located in the primary site (the alternative would be to consider it related to the primary tumor by a relation such as *has-origin*). This example makes clear how important it is in biology to clarify the meaning of part, where at least three conceptualizations co-exist: The locative one (a heart chamber is part of a heart), the functional one (an axon is part of a motor neuron), and the one motivated by origin (a metastasis is part of a tumor, an epithelium in a sputum sample is part of the respiratory mucosa).

Other Foundational Relations. **Branching** relations (*has-branch*, *branch-of*) define tree-like structures which typically describe pathways for the flow of matter or information in higher organisms (blood, lymphatic vessels and nerves), but which may also constitute the building principle of an organisms such as a plant or coral. There are several ways to conceptualize branching relations. In the FMA, a tree consists of a trunk and many generations of branches. Each branch is considered a subtree of a higher order tree, and each branch also has its own trunk. Thus *branch-of* can be interpreted either as a subtree or as a continuity relationship between two or more trunks. A subtree branch has a part relation to the higher order tree; two trunks have a branch relationship if

they are continuous with one another end to side or if a trunk terminates by bifurcating or trifurcating into subsidiary trunks. Consequently, branching relations cannot be subsumed by mereological relations because, generally speaking, a branch is not considered part of its trunk. To further illustrate this, any instance of *Aorta*, as the trunk of a systemic arterial tree, does not mereologically include any instance of its ramifications such as *Left Common Iliac Artery*, or *Femoral Artery*. Wherever branching relations are asserted between concept classes, we define *Has-Branch* and *Branch-Of* similar to *Part-Of* and *Has-Part* in Formula (2) and (3).

The development of the individual (**ontogeny**) and the development of the species (**phylogeny**) accordingly form the relation pairs *has-developmental-form*/ *developmental-form-of*, and *Has-Descendent* vs. *Descends-From*. Both are strict partial orders. In an embryo, e.g., *splanchnic mesenchyme* is a developmental form of *cardiogenic cords*, which – across some other steps – is a developmental form of *primitive heart*. According to the above comments, the inter-concept relations (*Has-Developmental-Form* / *Developmental-Form-Of*) have to be introduced when two concept classes are to be linked in terms of ontogeny.

All phylogeny relations, in contrast to the ontogeny relations, are maintained between concept classes, and not between individuals. As an example, *Homo Habilis Has-Descendant Homo Erectus* and *Homo Erectus Has-Descendant Homo Sapiens*. For any given instance of homo sapiens, there is no specific instance of any other hominid species, so there is no correlate of this relation at the level of individuals. Phylogenetic relations are maintained between organism concepts as well as between anatomical structure concepts (e.g., *Wing Descends-From Forelimb*).

There are other relations which are not partial orders but to which a foundational status can be equally ascribed. Topology provides, in addition to mereology, an important ontological organization principle. In formal approaches to topology, the basic relation, *connects*, is symmetric and relates two entities in space [3]. There are different kinds of connection, e.g. external connection (touching) or partial overlap [12]. Biological and common-sense notions of connection vary widely, so it may be advisable to talk about continuity, contiguity or attachment. If we stay closer to formal topology, we need the relation *externally connects*, which describes the touching of two objects without the sharing of parts, corresponding to the relation *continuous-with* in the FMA. For example, an *endocardium* is externally connected to a *myocardium*. If we allow boundaries (see below), another important relation pair is *bounds* vs. *bounded-by* [10], which is irreflexive, intransitive and antisymmetric (e.g., a *heart* is bounded by a *surface-of-the-heart*). Again, all of these relations exclusively relate individuals. Therefore, new concept-to-concept relations have to be defined (e.g. *Connects*), wherever they occur, in concept class definitions, similar to formulae (2) and (3). Note that the algebraic properties of these relations may differ: *continuous-with* is symmetric, but *Continuous-With* is not: In an individual neuron, its *cell body* is connected to its *axon* and vice versa. This contrasts with what we observe at the level of concept classes: Although each *axon* is connected to some *cell body*, not every *cell body* is connected to an *axon*.

General Attributes

In contrast to relations (e.g., *has-part, Is-A*), ordinary attributes such as *has-dimension*, *has-inherent-shape* can only be filled once. Important attributes are the (geometric) dimension, the distinction between solid and holes, as well as the distinction of count, collection and mass entities. All biological structure (in a strict sense) has a spatial dimension, which ranges from volumes, surfaces, lines to points. Quite naturally, the notion of a boundary comes into play. Any boundary must have exactly one dimension less than the entity it bounds. This restricts the domain and the range of the bounding relation pair *bounds* / *bounded-by*. Upper-level concepts, such as *Volume, Surface, Line, Point*, divide the domain of spatially relevant biological concepts into four disjoint partitions, because each biological structure entity has exactly one defining dimension [10]. Bounding structures can also be divided into so-called fiat and bona fide boundaries. According to [24] and [9], *bona fide* boundaries are those which have a structural correlate, e.g. the surface of the body, or the inner surface of a cell membrane. *Fiat* boundaries are 'artificial' boundaries, e.g., the *Medioclavicular Line*, or the *Sagittal Plane* in gross anatomy.

The next fundamental ontological distinction between three-dimensional objects is between "hollow spaces" and "solids". Examples for hollow spaces are the cranial cavity, the right atrium, the lumen of a bronchiole or the hollow space in a protein molecule. Nearly all biological objects have hollow space as parts (It is, therefore, not plausible to consider them as parts of the exterior space,

such as in formal topology). A possible axiomatization is that solids must have solids and may have hollow spaces as parts, whereas hollow spaces can only have hollow spaces and boundaries as parts [20].

Biological structures can occur as single, countable entities (e.g., a *liver*, a *tooth*, or a *cell*), but also as collections of uniform objects (e.g., *mitochondria*), or as stuff, e.g. a portion of *blood* or *water* [6]. Collection entities can be viewed either as sets of their constituents, or as their mereological sum. In the latter case, the relation between a collection and its elements boils down to a sort of *has-part*. As an example, the concept class *Leukocytes* denotes all possible mereological sums of individual *leukocytes*. Mass concepts can be treated as collections as well, because they are collections of small particles (cells, molecules, atoms). Whether to classify an item as a mass or collection is essentially a matter of perspective .

Non-foundational Relations between Concepts

There are some relations in the FMA, UMLS or in OPENGALEN which do not have a foundational status: Subrelations of *part-of*, such as *segment-of*, *layer-of*, *shared-part-of*, *arbitrary-part-of*, or *constitutes*, for which transitivity no longer holds [14, 9], can often be derived from the foundational ones by domain or range restrictions. For instance, *layer-of* requires an anatomical layer as domain and a physical entity as range. Or, *constitutes* has a mass or material as domain. The relation *shared-part-of*, on the other hand, can be inferred from the fact that an entity is part of more than one other entity. E.g., an *aorta* is part of a *trunk* and part of a *systemic arterial tree*. Other relations that can be inferred are *innervation* (*nerve* whose *endings* are connected to a *muscle*) and *insertion* (*tendon* connected to a *bone*).

The relation *Is-Conceptually-Disjoint* relates two concept classes which do not have any instance in common. This is the default situation in strict monohierarchies where all classes which do not subsume one another are mutually disjoint. In polyhierarchic taxonomies a class may have more than one taxonomic parent. As an example, *Pancreas* may be modeled as being both an *Endocrine Organ* and an *Exocrine Organ*, and an *Amino Acid* both as an *Organic Acid* and an *Organic Amine*. Most pairs of concept classes, however, are mutually exclusive: An *organ* cannot be a *cell*, and a *nucleotide* cannot be a *lipid*. In order to prevent unintended models, these concepts (or any parent of them) must be linked via the relation *Is-Conceptually-Disjoint*.

An analogous situation can be observed in a mereological ordering. Most arbitrary physical entities are spatially disconnected, e.g., there is no pair of respective instances that share any parts, e.g., a *hand* with a *foot*, or an *eye* with a *mouth*. Mereological disconnectedness between concepts can be asserted whenever the following condition is assumed to hold:

$$\begin{aligned} &\mathit{MereologicallyDisconnected}(A,B) =_{def} \\ &\quad \forall x,y : \mathit{inst\text{-}of}(x,A) \wedge \mathit{inst\text{-}of}(y,B) :\Rightarrow \\ &\qquad \neg\exists z : \mathit{part\text{-}of}(z,x) \wedge \mathit{part\text{-}of}(z,y) \end{aligned}$$

Theories

The vast domain of life science requires a decomposition of the whole domain into local theories, both in terms of granularity [8] as well as scope [7]. We define a theory as a set of formal axioms which describe a restricted (local) domain. We propose a lattice of theories which is designed along four parameters, viz. *granularity* (G), *species* (S), *development* (D) and *canonicity* (C).

Granularity. The conceptualization of biology is coined by our cognition. Macroscopic anatomy is restricted to the naked eye's view, histology requires a light microscope, our notions of cell biology are formed by the electron microscope, and knowledge of molecular biology and genetics is gathered using chemical and physical techniques. Along these lines, granularity issues have a large impact on high-level properties. In a very coarse-grained view, one may even consider classifying a microscopically thin membrane, such as a *basement membrane*, as a two-dimensional boundary, thus completely neglecting its spatial extension. Besides the sortal difference (degrees of dimensionality are mutually disjoint), this also has an impact on the connection of neighboring structures. What may be defined as externally connected to the naked eye will appear disconnected under the microscope. A low granularity may also encompass abstractions in terms of neglecting structural differences of kinds of objects (concept classes). Cell populations, such as *Leukocytes*, e.g., may be further classified into *Lymphocytes*, *Granulocytes* and others. A distinction of *Lymphocytes* into *B-* and *T-Lymphocytes*, and the latter into *T4-* and *T8-lymphocytes* will be required only in fine-grained theories, e.g. needed for the description of the pathology of immunodeficiency. In a complete ontological account of living organisms, granularity

ranges from populations, on the one hand, to atoms and subatomar particles, on the other hand.

Species. The universe of life includes millions of species. Hence, the domain of human anatomy is an extremely restricted one. Mediating domains are those of vertebrates or mammals. According to the classification of organisms, which is the prototype of a taxonomic order, properties can be introduced at any level of the classificatory tree and propagate across that tree. Under a simplifying view, *heart* is a muscular organ which has a cavity and is part of a *circulatory tract*. These properties hold true for *chordates*, *arthropodes* and some other phyla. As far as the hearts of more specific organisms are concerned, additional properties are required, e.g., a certain number of *ventricles* and *valves*, the presence of *blood* or *hemolymph*, different locations of *pacemaker cells* (see Fig. 1). Additionally, we have to consider intra-species variations such as gender or race.

Development. Organisms traverse a life cycle from birth to death. Each developmental stage has its own characteristics. Even distantly related organisms, such as humans and flies, exhibit a high degree of similarity in the first embryologic stages. The existence of many parts of an organism is restricted to certain stages. For example, in mice embryos, an *ectoderm* exists only in the so-called Tanner stages TS9 – TS19, and there is no heart before the Tanner stage TS11. Other body parts (e.g. the heart, cf. Fig. 1) appear in a certain embryologic stage and perdure in all subsequent steps of the life cycle.

Canonicity. Here we introduce the notion of *canonicity*, as the well-formedness of biological structure, and define it as the degree by which a biological object corresponds to its canonical, i.e., idealized form. We suggest an ordinal scale with five levels of canonicity, cf. Table 1. The higher the canonicity level, the more axioms have to be applied. All axioms introduced in a lower level are propagated to all higher levels. Axioms which describe structural modifications specific to a concrete disorder, e.g *Stomach Has-Part Ulcer* are not considered in this framework.

- *Level 1* introduces those axioms which hold even with lethal structural modifications or postmortem degeneration, such as *Erythrocytes has-part Hemoglobin*, *Bone Has-Part Calcium Carbonate*, *Heart Ventricle Part-Of Heart*, *Leather Has-Part Collagen* (but not *Heart Valve Part-Of Heart* because it could be an isolated heart valve for transplantation);

- *Level 2* introduces, additionally, all those axioms which hold for the description of biological structures organized in an *organism*, irrespective of living or dead, e.g., the axiom *Heart Valve Part-Of Heart* is introduced at this level, as well as *Cell Nucleus Part-Of Cell*;

- In *Level 3* all those axioms are added which hold in living organisms, in addition to dead organisms, e.g., *Aorta Location-Of Blood*, or *Vertebrate Body Has-Part Head* (but not yet *Gastrointenstinal Tract Has-Part Stomach*, because most individuals survive a total remotion of the stomach).

- *Level 4* introduces, additionally, all those axioms which characterize a healthy organism, e.g. *Hand Has-Part Thumb* and *Gastrointenstinal Tract Has-Part Stomach*. However, it still allows anatomical variations when they have no impact on the function of the organism.

- *Level 5* finally completes the set of axioms needed for the description of the "ideal" organism. Here enter, e.g., many cardinality constraints (e.g., in human: 32 *teeth*), one *spleen*, three *lobes* of the *right lung*.

A theory can be expressed by a node in the lattice of the four axes *viz.* G, S, D, and C. Hereby the values of granularity (G), development (D), canonicity (C) are located on an ordinal scale, the values of species (S) are given by the nodes of the classification of organisms. Each node of this classificatory tree introduces properties which are inherited by its subsequent nodes. As an example, *Fish Heart* Is-A *Vertebrate Heart* or *Drosophila Eye* Is-A *Arthropode Eye*. This means that *Fish Heart* inherits all properties from *Vertebrate Heart*, and *Drosophila Eye* inherits all properties from *Arthropode Eye*. The same mechanism can be observed with canonicity. All properties that structures of low canonicity have in common (e.g., *Tissue* consisting of *Cells*) are inherited by the more canonic structures. No such inheritance rules apply to the variables *development* (D) and *granularity* (G).

Taking the *heart* as prototypical example, we will now demonstrate practical inferences which are supposed to be drawn from a biological ontology based on our assumptions:

- A heart with four chambers is not compatible, e.g., with any theory characterized by *S = fish*, or by *S = human & D = 4-week-embryo*.

Level	1	2	3	4	5
Theory	any amount of matter, if of biological origin	any living or dead organism	any living organism	living organism without pathologic modifications	ideal organism
Set of Axioms	n_1	n_2 $n_1 \subset n_2$	n_3 $n_2 \subset n_3$	n_4 $n_3 \subset n_4$	n_5 $n_4 \subset n_5$

Table 1: Ordinal Scale of Canonicity

- Let us assume the relation *connects* which is maintained between the right and the left ventricles. We then may exclude most non-mammals (since they have no right and left ventricle), but we may also exclude anatomical hearts of adult mammals (because they have a septum between the two ventricles). This scenario is compatible with the theories *D = embryo & S = mammal* as well as with *S = mammal & C = 3*.

- Given the theory *D = adult & S = vertebrate & C = 5*, every instance of heart implies the location of blood, and every instance of blood has an instance of erythrocytes as part. Assuming that *has-part* implies *location-of*, and that *location-of* is transitive, we are able to infer that in this theory every instance of *Heart* is the location of an instance of *Erythrocytes*, as well.

Conclusions

In this paper we defined requirements for ontologies of biological structure. We introduced a set of canonical relations and attributes required for the description of biological structure, and discussed their semantics as well as algebraic properties.
Finally, we sketched an architecture by which terminological knowledge about the anatomy of a broad range of organisms, developmental stages as well as malformations and pathological modifications, can be expressed. A central element is the decomposition into theories, which help organize the hierarchies and the axioms in terms of granularity, developmental stage, species, and canonicity. We claim the following advantages in using this approach:
(i) Redundancies are avoided. As an example, most axioms that describe the species *mice*, *humans*, and *dogs* are identical and therefore can reasonably be encoded into a more general theory (such as the one of *vertebrates*). In turn, the more general theory inherits the shared properties of more specific theories, e.g., the ones pertaining to mice, humans or dogs. In a similar vein, attributes that healthy and pathologically modified organisms have in common are described in the non-canonical theory from which the canonical theory inherits the shared properties.
(ii) Adequate theories for a specific application can be selected. It is neither computationally tractable nor useful to export the whole knowledge of biology into a formalism in which logical operations can be performed, e.g., by a terminological reasoner. For example, if we need to reason about a TS12 mouse embryo, we select the adequate intersection of theories to access the axioms we really need. Some of these axioms are inherited from the *mammal* theory, others from the theory of the *vertebrates*, and still others come from the theory of the *chordates*. Some axioms are encoded in the subtheory of a developmental stage of the vertebrates, and, last but not the least, there are some axioms which are only specific to the TS12 mouse embryo.
(iii) The intersection of arbitrary theories has variable extensions. There are many cases with no extensions. The compatibility of theories can be checked by formal reasoning devices. As an example, a heart with one ventricle in a theory restricted by *S = human* and *D = adult* is not compatible with *C = canonical*.

References

[1] J. Aitken, B. Webber, and J. Bard. Part-of relations in anatomy ontologies: A proposal for RDFS and OWL formalisations. In *Proceedings of the Pacific Symposium on Biocomputing 2004*, pages 166–177. Hawaii, USA, January 6-10, 2004.

[2] K. Campbell, A. Das, and M. Musen. A logical foundation for representation of clinical data. *Journal of the American Medical Informatics Association*, 1(3):218–232, 1994.

[3] R. Casati and A. Varzi. *Parts and Places. The Structures of Spatial Representation*. Cambridge, MA: MIT Press, 1999.

[4] A. Gangemi, N. Guarino, C. Masolo, and A. Oltramari. Understanding top-level ontological distinctions. In *Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing*, pages 26–33. Seattle, USA, August 4-5, 2001.

[5] Gene Ontology Consortium. Creating the Gene Ontology resource: Design and implementation. *Genome Research*, 11(8):1425–1433, 2001.

[6] P. Gerstl and S. Pribbenow. Midwinters, end games and body parts: A classification of part-whole relations. *International Journal of Human-Computer Studies*, 43:865–889, 1995.

[7] C. Ghidini and F. Giunchiglia. Local models semantics, or contextual reasoning = locality + compatibility. *Artificial Intelligence*, 127(2):221–259, 2001.

[8] J. Hobbs. Granularity. In *Proceedings 9th Intl. Joint Conference on Artificial Intelligence*, pages 432–435. Los Angeles, CA, 18-23 August, 1985.

[9] J. Mejino, A. Agoncillo, K. Rickard, and C. Rosse. Representing complexity in part-whole relationships within the foundational model of anatomy. In *Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association*, pages 71–75. Washington, D.C., November 8-12, 2003.

[10] P. Neal, L. Shapiro, and C. Rosse. The DIGITAL ANATOMIST structural abstraction: A scheme for the spatial description of anatomical entities. In *Proceedings 1998 AMIA Annual Fall Symposium*, pages 423–427. Orlando, FL, Nov. 7-11, 1998.

[11] F. N. Noy, M. Musen, J. Mejino, and C. Rosse. Pushing the envelope: Challenges in a frame-based representation of human anatomy. Technical Report SMI-2002-0925, Stanford University, 2002.

[12] D. Randell, Z. Cui, and A. Cohn. A spatial logic based on regions and connection. In *Principles of Knowledge Representation and Reasoning. Proceedings of the 3rd International Conference*, pages 165–176, 1992.

[13] A. Rector, A. Gangemi, E. Galeazzi, A. Glowinski, and A. Rossi-Mori. The GALEN model schemata for anatomy: Towards a re-usable application-independent model of medical concepts. In *Proceedings of the 12th Conference of the European Federation for Medical Informatics*, pages 229–233. Lisbon, Portugal, 1994.

[14] J. Rogers and A. Rector. GALEN's model of parts and wholes: Experience and comparisons. In *Proceedings of the 2000 Annual Symposium of the American Medical Informatics Association*, pages 714–718. Los Angeles, CA, November 4-8, 2000.

[15] C. Rosse and J. Mejino. A reference ontology for bioinformatics: the Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 2004. In press.

[16] C. Rosse, J. Mejino, B. Modayur, R. Jakobovits, K. Hinshaw, and J. Brinkley. Motivation and organizational principles for anatomical knowledge representation: The DIGITAL ANATOMIST symbolic knowledge base. *Journal of the American Medical Informatics Association*, 5(1):17–40, 1998.

[17] OBO. *Open Biological Ontologies (OBO)*. http://obo.sourceforge.net/, 2004.

[18] UMLS. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2003.

[19] XSPAN. *Cross Species Anatomy Network (XSPAN)*. http://www.xspan.org, 2004.

[20] S. Schulz and U. Hahn. Mereotopological reasoning about parts and (w)holes in bio-ontologies. In *Formal Ontology in Information Systems. Collected Papers from the 2nd International FOIS Conference*, pages 210–221. Ogunquit, Maine, USA, October 17-19, 2001.

[21] S. Schulz and U. Hahn. Parthood as spatial inclusion – evidence from biomedical conceptualizations. In *Principles of Knowledge Representation and Reasoning. Proceedings 9th International Conference*. Whistler, Canada, June 2-5, 2004.

[22] P. Simons. *Parts: A Study in Ontology*. Oxford: Clarendon Press, 1987.

[23] B. Smith and C. Rosse. The role of foundational relations in the alignment of biomedical ontologies. In *Proceedings 11th World Congress on Medical Informatics*. San Francisco, CA, Sept., 2004.

[24] B. Smith and A. Varzi. Fiat and bona fide boundaries. *Philosophy and Phenomenological Research*, 60(2):401–420, 2000.

[25] B. Smith, J. Williams, and S. Schulze-Kremer. The ontology of the Gene Ontology. In *Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association*, pages 609–613. Washington, D.C., November 8-12, 2003.

[26] C. Welty and N. Guarino. Supporting ontological analysis of taxonomic relationships. *Data & Knowledge Engineering*, 39(1):51–74, 2001.

Representing the MeSH in OWL: Towards a Semi-Automatic Migration

LF. Soualmia[1,2], C. Golbreich[3], SJ. Darmoni[1,2]

[1] CISMeF & L@STICS, Rouen University Hospital, 76031 Rouen, France
{Lina.Soualmia, Stefan.Darmoni}@chu-rouen.fr
[2]PSI Laboratory - FRE CNRS 2645, INSA-Rouen, 76131 Mont-Saint Aignan, France
[3] Laboratoire d'Informatique Médicale, University Rennes 1, 35033 Rennes, France
Christine.Golbreich@univ-rennes1.fr

Abstract

Due to the numerous health documents available on the Web, information retrieval remains problematic with existing tools. This paper is positioned within the context of the CISMeF project (acronym of Catalogue and Index of French-speaking Medical Sites) and of a future Semantic Web. In CISMeF the resources are described using a set of metadata based on a structured terminology which "encapsulates" the MeSH thesaurus in its French version. Now, the objective is to migrate the CISMeF terminology, and thus the MeSH thesaurus, to a formal ontology, so as to get a more powerful search tool. The paper presents the very first stage and results of this ongoing project, aiming at migrating the MeSH to OWL. It reports on the first steps, which have presently been done, concerning the automatic transformation of the terminology into OWL-DL. First, the CISMeF terminology has been "formalized" in OWL. Then, the resulting OWL "ontology" has been imported under the Protégé editor which makes possible to check its consistency and its classification in using Racer. Finally, the paper concludes on the current results and encountered difficulties, and gives future work perspectives.

INTRODUCTION

The amount of health information available on Internet is considerable. Information retrieval remains problematic: users are now experiencing huge difficulties in finding precisely what they are looking for, among the tons of documents available online. Generic search engines (e.g. Google) or generic catalogues (e.g. Yahoo) cannot solve this problem efficiently and offer a selection of documents that turns out to be either too large or ill-suited to the query. Free text word-based (or phrase-based) search engines typically return innumerable completely irrelevant hits requiring much manual weeding by the user and might miss important information resources. Free text search is not always efficient and effective: the sought page might be using a different term (synonym) that points to the same concept; spelling mistakes and variants are considered as different terms; search engines cannot process HTML *intelligently*, the most the most widespread language on the Web.

This paper is positioned within the context of the CISMeF[1] project (acronym of Catalogue and Index of French-speaking Medical Sites) and of a future Semantic Web[2]. The CISMeF catalogue was developed since 1995 to assist the health professionals, the students and the general public in their search for health information on the Web. CISMeF is a quality-controlled health gateway, cataloguing the most important and quality-controlled sources of institutional health information in French in order to allow end-users to search them quickly and precisely.

In CISMeF the resources are described using a set of metadata elements based on a structured terminology which "encapsulates" the MeSH[3] (Medical Subject headings) thesaurus in its French version. The present work follows that done in[4] and aims at migrating the CISMeF terminology, and thus the MeSH thesaurus, to a formal ontology, so as to get a more powerful search tool[5]. Every year the MeSH thesaurus is modified and new concepts are added. As the rapid evolution of medical knowledge and the very dynamic nature of web information require frequent updates, a formal knowledge representation also contributes in maintaining a consistent terminology, by detecting the inconsistencies that might result from updates or modifications. We chose the OWL DL sublanguage[6] to represent the CISMeF terminology, as being the W3C standard and also as it provides powerful reasoning services based on Description Logics.

The paper presents the very first stage and results of an ongoing project aiming at "formalizing" the MeSH in OWL. The long term goal is to migrate the existing terminology to a formal representation in OWL and to enhance it. The paper main contribution concerns the modeling choices underlying the automatic migration process used for migrating MeSH to OWL. Section 2 introduces the CISMeF catalogue in which these experimentations are carried out. Modeling choices underlying the automatic transformation towards OWL are detailed in section 3. Section 4 presents the results of the consistency checking and classification of the

OWL "ontology", after its import under the Protégé editor[7], using Racer[8]. Section 5 draws conclusion from the results and gives future work perspectives.

THE CISMeF TERMINOLOGY AND USE FOR RESOURCES INDEXING

The CISMeF catalogue describes and indexes a large number of health information resources (*n=13,198*) and has three main topics: guidelines for health professionals, teaching material for students in medicine, and consumer health information. A resource is any support that may contain health information : it can be a Web site, Web pages, documents, reports and teaching material. Metadata based on a terminology "ontology-oriented" are used to describe the resources.

CISMeF Terminology. The catalogue resources are indexed according to the CISMeF terminology, which is based on the French version of the MeSH concepts provided by the INSERM (National Institute of Health and Medical Research). The MeSH thesaurus in its 2003 version includes approximately 22,000 keywords (e.g. *abdomen, hepatitis*) and 84 qualifiers (e.g. *diagnosis, complications*). These concepts are organized into hierarchies from the most general to the most specific concept. For example, the keyword *hepatitis* is more general than the keyword *hepatitis viral A*. The qualifiers are used to specify which particular point of view of a keyword is addressed. For example the association of the keyword *hepatitis* and the qualifier *diagnosis* (noted *hepatitis/diagnosis*) restricts *hepatitis* to its *diagnosis* aspect. Qualifiers are also organized into hierarchies.

The heterogeneity of Internet health resources and the great specificity of MeSH keywords, which makes it difficult to refer broadly to a medical specialty, led the CISMeF group to introduce two new concepts, namely *metaterms* and *resource types*. Metaterms (*n*=66) concern medical specialties. The *resource types* (*n*=127) describe the nature of the resource e.g. *teaching material*, *clinical guidelines*. The keywords and qualifiers in CISMeF are thus clustered according to *metaterms.* Each one is related to one or several metaterms. The metaterms and resource types enhance information retrieval into the catalogue. In fact, meta-terms have been created to optimize information retrieval in CISMeF and to overcome the relatively restrictive nature of MeSH keywords. For instance, the queries 'guidelines in cardiology' and *'databases in psychiatry'* where *cardiology* and *psychiatry* are only MeSH keywords get few or no answers. Introducing *cardiology* and *psychiatry* as metaterms is an efficient strategy to get more results because instead of exploding one single MeSH tree (e.g. *psychiatry* as a MeSH keyword), using metaterms results in an automatic expansion of the queries by exploding other related MeSH or CISMeF trees as well as the current tree (e.g. *psychiatric hospital* as a MeSH keyword or *mental health dispensary* as a resource type will be exploded in the case of the *psychiatry* query).

The CISMeF terminology and the catalogue resources are stored in a relational database (Oracle 8.i). The CISMeF terminology has the same structure as a terminological ontology[9]:

- The vocabulary, that describes major terms of the medical domain, is well known by the librarians and the health professional.
- Each concept has:
  - a preferred term (Descriptor) to express it in natural language.
  - a set of properties.
  - a natural language definition that allows to differentiate it from the concepts it subsumes and those that it subsumes.
  - a set of synonyms.
  - a set of constraints to apply on the qualifiers. For example the qualifier *'Complications'* could only be used for the *'Diseases'* arborescence.
  - a set of equivalences. For example the association '*Hepatitis/chemically induced'* is equivalent to the keyword *'Hepatitis, toxic'*.

Many ways of navigation and information retrieval are possible into the catalogue. *Simple search* which is based on the subsumption relationships is the most often used. If the query, a given word or expression, can be matched with an existing term, then the result of the query is the union of the resources indexed by the term, and by the terms it subsumes, directly or indirectly, in all the hierarchies it belongs to. For example a query on *Hepatitis* will return as answer all the resources related to *Hepatitis* and also those related to *Hepatitis A, Hepatitis B*...etc. If the query cannot be matched, then the search is done over the other fields of the metadata. If it fails, a full-text search is carried out.

But although quite powerful, this kind of search requires a good knowledge of the medical domain, and exhibits some limitations.

Indeed, the consistency of this terminology has not yet been studied and some defaults may arise. For example, in the *'Anatomy'* tree, some keywords are hierarchically organized according to a 'specialization' relationship, while in fact they are related by the *'part of'* relationship. As a consequence, a query on *'headache*' also returns documents on *'mouth pain'*, *'eye pain'* and *'ear pain'* among others.

Another problem in query processing concerns the associations between keyword/qualifier. A query on "*hepatitis/diagnosis*" is processed in CISMeF as a conjunction of two queries one on "*hepatitis"* and

one on "*diagnosis"*. Thus, when exploded, this query returns also resources on "*lumbago/diagnosis*" and resources on *"lumbago/radiography"* since *"radiography"* is subsumed by *"diagnosis"*.

The descriptions are incomplete. For example, the keyword *"abdominal neoplasm"* is defined as a "*neoplasm"* and not as an *"abdominal disease"* whereas "*stomach neoplasm"* is defined as "*neoplasm"* and a "*stomach disease"*. The term "*abdominal disease"* does not exist in the MeSH. Therefore some improvements are now investigated. Because of its size, automatic tools are needed. A formal representation may be promising, in particular to verify the terminology consistency and the overall classification.

Metadata. The notion of metadata appeared before Internet but its interest has growth with the number of electronic publications and digital libraries. « The Semantic Web dream is of a Web where resources are machine understandable and where both automated agents and humans can exchange and process information.[1]». The solution proposed by the W3C is to use metadata to describe the data contained on the Web and to add semantic markup to Web resources that describes their content and functionalities, from the vocabulary defined in ontologies. Metadata are data about data or in the Web context, data describing Web resources. When properly implemented, metadata shall unambiguously describe resources, so enhancing information retrieval.

In CISMeF we use several sets of metadata. Among them there is the Dublin Core[10] (DC) metadata set, which is a 15-element set, intended to aid discovery of electronic resources. The resources indexed in CISMeF are described by eleven of the elements of Dublin Core: *author, date, description, format, identifier, language, editor, type of resource, rights, subject* and *title.* DC is not a complete solution; it cannot be used to describe the quality or location of a resource. To fill these gaps, CISMeF uses its own elements to extend the DC standard. Eight elements are specific to CISMeF: *institution, city, province, country, target public, access type, sponsorships* and *cost.* The user type is also taken into account. CISMeF defined two additional fields for the resources intended for the health professionals: indication of the *evidence-based medicine* and the *method* used to determine it. In the teaching resources eleven elements of the IEEE 1484 LOM (Learning Object Metadata) "Educational" category are added.

The metadata format was the HTML language in 1995. In 2000, in order to allow interoperability with other platforms the XML language became the metadata format. Since December 2002, the format used is RDF a basic Semantic Web language, within the EU-project MedCIRCLE framework[11] in which CISMeF is a partner. This project was initiated to qualify the quality of health information and to guide consumers to trustworthy health information. The vocabulary of the HIDDEL (High Information Description Disclosure Evaluation Language) metadata is contained in an ontology (represented in RDF Schema) and the resources are described in RDF according the concepts of the HIDDEL ontology.

AUTOMATIC MIGRATION TO OWL

There are several works concerning the UMLS® [24] and its Semantic Network representation with a formal language[12-15], but as far as we know, the MeSH formalization (a component of the UMLS metathesaurus), has not yet been studied. MeSH suffers from its size, its numerous inconsistencies and ambiguities concerning the medical concepts. For example, *'diagnosis'* is defined as a medical specialty and also a qualifier. In previous works MeSH has partly been enhanced by introducing new concepts in CISMeF[1] but it now appears not sufficient. An advantage of using description logics is to benefit of advanced inference services (satisfiability, subsumption, classification, consistency checking, instanciation, realization and retrieval), which can contribute to maintain a consistent terminological system and to improve results of queries thanks to inferences.

This section reports on the first stage of a general process aiming at the migration and enhancement of the MeSH.

Modeling principles. A first modeling principle was to "clean" the MeSH taxonomy, in distinguishing between the *'part-of'* and the *'is-a'* relationships (the *Anatomy, Biological Sciences* and *Geographic Locations* hierarchies are processed separately).

The second one was to clearly distinguish between the different notions: specialty, keyword, and qualifier. For example the specialty *"diagnosis"* is distinguished from the qualifier *"diagnosis"* because they denote different notions (resp. "*virology"*).

The third one concerns the elicitation of the qualifiers domain. Qualifiers cannot be associated to all the keywords. It is a MeSH restriction. For example, the qualifier "*diagnosis"* can be associated to the keyword "*diseases*" (and thus to

1 Ian Horrocks, IEEE Intelligent systems March / April 2002

```
Descripteur Francais: HEPATITE CHRONIQUE
Descripteur Americain: Hepatitis, Chronic
Code Cat MESH: C06.552.380.350
Synonymes Français: HEPATITE CHRONIQUE ACTIVE
Synonymes Américains: Chronic Hepatitis
                      Cryptogenic Chronic Hepatitis
                      Hepatitis, Chronic, Cryptogenic
Derives Americains: Hepatitis, Chronic Active
                    Active Hepatitides, Chronic
                    Active Hepatitis, Chronic
                    Chronic Active Hepatitides
                    Chronic Active Hepatitis
                    Chronic Hepatitides
                    Chronic Hepatitides, Cryptogenic
                    Chronic Hepatitis, Cryptogenic
                    Cryptogenic Chronic Hepatitides
                    Hepatitides, Chronic
                    Hepatitides, Chronic Active
                    Hepatitides, Cryptogenic Chronic
                    Hepatitis, Cryptogenic Chronic
MESH definition: A collective term for a clinical and pathological syndrome which has several causes
and is characterized by varying degrees of hepatocellular necrosis andinflammation. Specific forms of
chronic hepatitis include autoimmune hepatitis (HEPATITIS, AUTOIMMUNE), chronic hepatitis B;
(HEPATITIS B, CHRONIC), chronic hepatitis C; (HEPATITIS C, CHRONIC), chronic hepatitis D; (HEPATITIS
D, CHRONIC), indeterminate chronic viral hepatitis, cryptogenic chronic hepatitis and drug-related
chronic hepatitis (HEPATITIS, CHRONIC, DRUG-INDUCED).
NLM: D006521
```

Figure 1. Concept definition in the MeSH text file provided by the INSERM

all its descendants), but not to the "*geographic locations"*. These restrictions on the qualifiers were formalized to check whether a qualifier is not wrongly associated to a keyword, and can be viewed as defining the domains of the qualifiers.

The fourth one concerns multiple hierarchies. A keyword in the MeSH may belong to several trees. In this case, for the moment, the keyword is associated to the intersection of its direct super-concepts.

Finally, since the objective is to remain as much as possible compatible with the original MeSH indexing, for each resource, the related MeSH concepts used for its indexing, serve to define a new concept of the ontology used for the resource new formal indexing. This new concept is defined from the conjunction of the original ones and will be used to define the individuals.

From Text Files to a Database. Each year the MeSH text files (Fig1.) are first processed using a awk script on a Unix platform to inform the table TB_MeSH in the CISMeF database which contains the following items: *Descripteur Français, Code Cat MeSH* and *NLM*. The other fields are not yet taken into account (e.g. MeSH definition) because they are in English. Nevertheless, the definitions are under translation into French in the context of the VuMeF project[16].

The *Code Cat MeSH* indicates in which hierarchy the descriptor is located and refers to a level position. A descriptor may belong to many hierarchies. This information is very useful to represent the hierarchies. For example one can deduce that *Hepatitis, Chronic* (C06.552.380.350) is subsumed by *Hepatitis* (C06.552.380) with a difference of level of 1. In practice, a join is done on the tables TB_MeSH and TB_MC, which contains all the descriptors used in the catalogue (n= 9,765), to update the terminology and also to compute all the existing links between descriptors and the levels in the hierarchies.

From the Database to a Terminological Knowledge Base. OWL-DL is a Description Logics (DL) language[17]. DL structures the domain knowledge at two levels: a terminological level (TBox or ontology), containing the classes of domain objects (concepts), with their properties (roles) and an assertional level (ABox), containing individuals (instances). In our case the ontology contains the specialties, keywords, qualifiers and resource types OWL-DL classes. Instances represent the indexed resources (to be soon included in the ABox under construction).

A DL system not only stores terminologies in a formal logic-based language, but also provides reasoning services. Main reasoning tasks concern satisfiability (existence of a model of the ontology), subsumption (supporting the classification of a concept in the hierarchy), and instance recognition (enabling to identify for a particular individual the most specific concepts it is an instance of).

The CISMeF terminology, is automatically transformed from the previous relational database into an OWL ontology, in using Java and SQL queries. The construction is a Top-Down construction, going from the Top concept to the specialties, and then to the keywords and resource types. The qualifiers hierarchy is modeled separately. The objective is to automate as much as possible all the process. As in [18] the illegal characters (- : , &) and spaces of the original descriptor names were replaced by underscores. All accented characters (*e.g.*, *"éèêë"*) were replaced by non-accented (*"e"*) ones. Names that began with numbers were prefixed with underscores. For

example, "*11-hydroxycorticostéroïdes*" is renamed by "*_11_hydroxycorticosteroides*".

Representing the terminology in OWL

- ***OWL classes***

The keywords, metaterms and resource types, are represented as OWL classes. When two concepts have the same label but correspond to distinct notions, they are prefixed by *mt_* when it is a speciality, *tr_* when it is a resource type, *qu_* when it is a qualifier.

The specialties are, for the moment, represented as primitive concepts, without any OWL definition. Each specialty from the CISMEF specialty table, is automatically transformed into such a concept, for example, the specialty '*cardiology'* is represented by the OWL class:

```
<owl:Class rdf:ID="mt_cardiology">
```

- ***OWL hierarchies structuration***

The 'is-a' relations from the "cleaned" MeSH terminology are represented thanks OWL subsumption axioms. First, the keywords and resource types who are direct sons of the specialities are described. Then their descendants are progressively added level by level. For example '*accident domestique'* is a sub-concept of '*accidents'*:

```
<owl:Class rdf:ID="accident_domestique">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#accidents" />
  </rdfs:subClassOf>
</owl:Class>
```

If a concept has more than one super-concept, it is represented as a subclass of the intersection of its super-concepts, for example '*accident_radiation'* is defined using the intersection of '*accidents*' and '*accident_travail'*(occupational accident):

```
<owl:Class rdf:ID="accident_radiation">
  <rdfs:subClassOf>
    <owl:Class>
  <owl:intersectionOf rdf:parseType="Collection">
      <owl:Class rdf:about="#accident_travail" />
      <owl:Class rdf:about="#accidents" />
    </owl:intersectionOf>
    </owl:Class>
  </rdfs:subClassOf>
</owl:Class>
```

- ***OWL properties***

The qualifiers are represented as OWL properties, hierarchically organized. Each qualifier from the CISMeF qualifiers table, issued from the MeSH text files, is automatically transformed into a corresponding OWL property with a defined domain *"domain_qu_"*, but without any range. For example, the CISMEF qualifier '*contre-indications*' is transformed into:

```
<owl:ObjectProperty
rdf:ID="qu_contre_indications">
<rdfs:domain
rdf:resource="#domain_qu_contre_indications" />
<rdfs:subPropertyOf>
<intersectionOf rdf:parseType="Collection">
<owl:ObjectProperty  rdf:about="#qu_pharmacologie"
/>
<owl:ObjectProperty
rdf:about="#qu_usage_therapeutique" />
</intersectionOf>
</rdfs:subPropertyOf>
</owl:ObjectProperty>
```

- ***The "part-of" property***

The keywords that belong to the trees *Anatomy, Biological Sciences* and *Geographic Locations* are organized hierarchically according to the *part-of* relationship. They are processed separately. The OWL property *partOf* is defined as:

```
<owl:ObjectProperty rdf:ID="partOf">
</owl:ObjectProperty>
```

In the CISMeF (MeSH) terminology, the keyword *"abdomen"* is placed under the keyword "*region corps*" (body region) in the *Anatomy* tree. As this hierarchical relation corresponds in fact to a "*partOf"* relationship the concept *"abdomen"* is defined as:

```
<owl:Class rdf:ID="abdomen">
  <rdfs:subClassOf>
    <owl:Restriction>
        <owl:onProperty rdf:resource="#partOf" />
         <owl:someValuesFrom
         rdf:resource="#region_corps" />
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

- ***Domains of properties***

Since a qualifier can be applied to several hierarchies of keywords, the domain of a property associated to a qualifier is represented by the union of the related qualified concepts. In CISMeF, this information is stored as a string in the item *Restriction* and the hierarchies roots are delimited by a comma, and was inserted manually by the medical librarian into the database. For example, "C01-C05, D, G" indicates that the considered qualifier can be applied to the keywords C01 to C05, D01 to D27, G01 to G14. For each restriction (84) such strings have been automatically processed so as to determine all the related keywords. For example the domain of the property *"qu_contre_indications"* has been defined as:

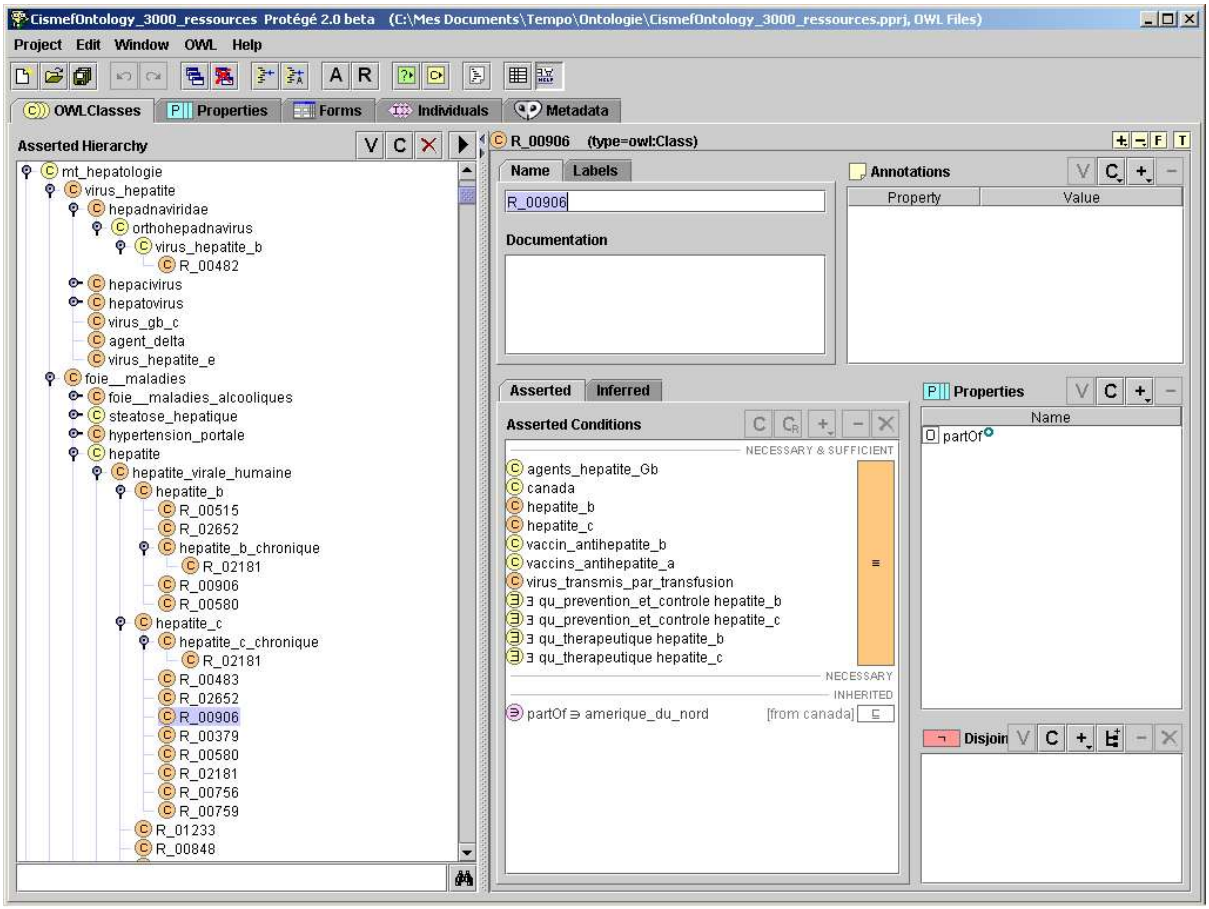


Figure 2. OWL ontology import into Protégé.

```
<owl:Class rdf:ID="domain_qu_complications">
<owl:unionOf rdf:parseType="Collection">
<owl:Class rdf:about="#anesthesie_analgesie" />
<owl:Class  rdf:about="#intervention_chirurgicale"
/>
<owl:Class
rdf:about="#produits_chimiques_inorganiques" />
        …
</owl:unionOf>
</owl:Class>
```

Representing the resources descriptions in OWL

The concepts related to the resources (n=13,198) have also been defined. For each resource, a new concept of the ontology has been created. For example the resource number 112, which is concerned by a diagnosis of some hepatitis and a viral vaccine, is indexed by *'hepatite/diagnostic'* (hepatitis/diagnosis) and *'vaccin antiviral'* (antiviral vaccines), therefore its description field of the metadata is represented as an instance of the defined concept R_112 = ∃ diagnostic.hepatite ∩ vaccin_antiviral.

```
<owl:Class rdf:ID="R_112">
<owl:intersectionOf rdf:parseType="Collection">
<owl:Class rdf:about="#vaccin_antiviral" />
<owl:Restriction>
<owl:onProperty rdf:resource="#qu_diagnostic" />
<owl:someValuesFrom rdf:resource="#hepatite" />
</owl:Restriction>
</owl:intersectionOf>
</owl:Class>
```

CHECKING AND CLASSIFYING THE IMPORTED OWL 'ONTOLOGY'

Protégé OWL import. The size of the TBox is very large: 23,239 concepts (9,765 keywords; 65 specialties; 127 resource types; 84 domains; 13,198 concepts related to the resources) and 85 relations (84 qualifiers plus the relation *partOf*). It was not possible to import the resulting OWL file into the Protégé 2000 editor[7] thanks to its OWL (plug-in build 119) because the virtual Java machine had no sufficient memory, due to the file size (20.75 MB). Thus it was necessary to reduce the number of the concepts related to the resources to 3,000. The file loading has then been successfully processed in ~ 30 min (Fig.3). The ontology sub-language has been checked to be OWL-DL.

Figure 3 shows the concept R_00906, which represents a resource indexed with the concepts *agents hepatite Gb, Canada, hepatite b, hepatite c, vaccin anti-hepatite b, vaccins anti-hepatite a, virus transmis par transfusion, hepatite b/prevenrion et controle, hepatite c/ prevention et controle, hepatite b/therapeutique, hepatite c/therapeutique.* It which inherits the property partOf from its definition, as the concept *canada* is part of the concept *amerique_du_nord* (America, North).

Consistency checking. The consistency checking of all the terminology, augmented by the subset of 3,000 concepts describing resources, has approximately taken three hours (with Protégé 2.0 beta and the OWL plug-in build 119) using Racer[8]. No inconsistent class has been found. A little surprising, this may be explained by several reasons:

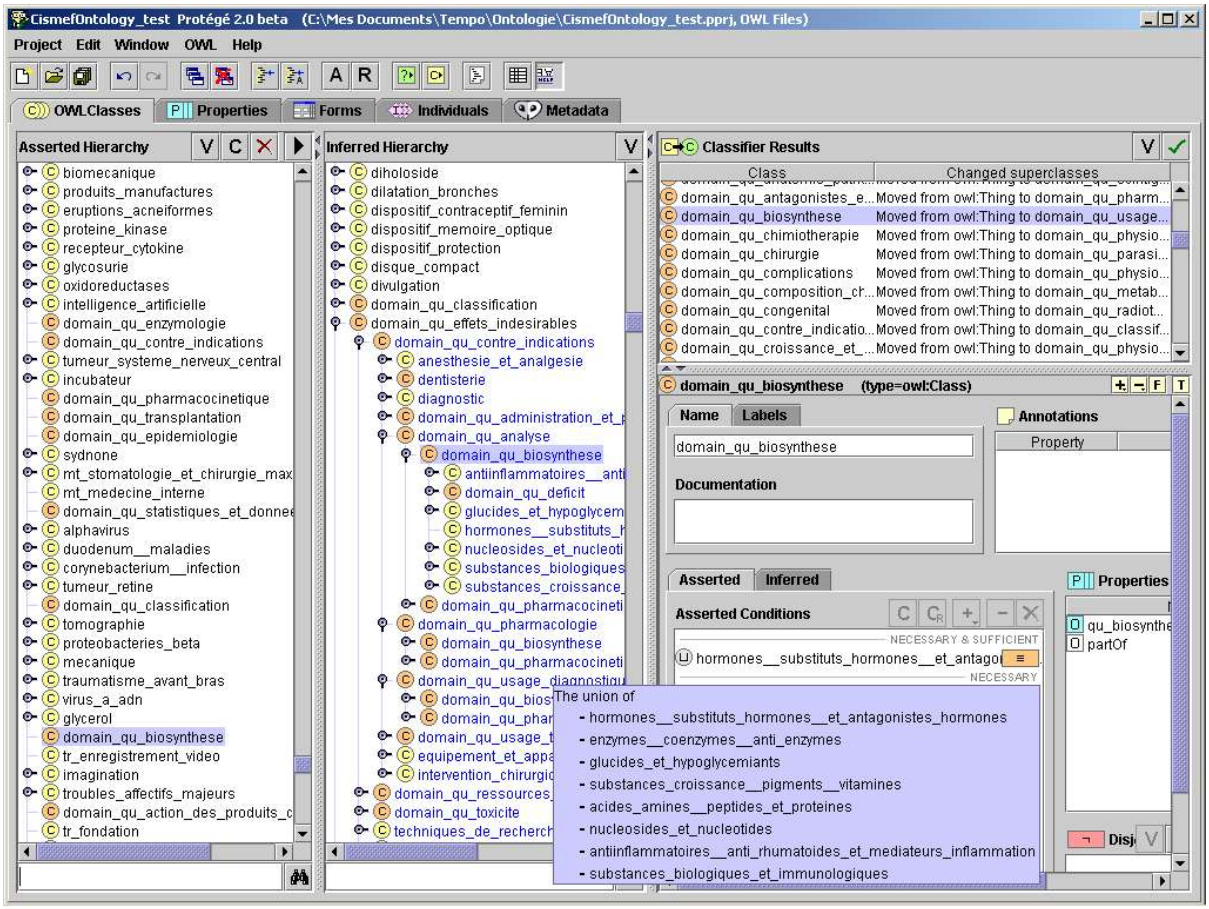


Figure 3. New concepts classification (domains and resources' concepts).

- the pre-processing import of the MeSH into a structured database
- the distinction between the different notions (specialties, keywords, qualifiers and resource types)
- the use of the intersection operator for a class (object property) having several super-classes (super-properties)
- the classes, except those for resources and domains, have no description
- classes that describe the resources are OWL defined concepts, based on the CISMeF manual indexation, checked by the librarian team.

Classification. The classification was also very long (checking first whether the ontology is consistent). A new hierarchy has been inferred and all the domains and the concepts used for the resources have been classified according to their description. Fig 2 and Fig 3 show how the domains, initially defined as subclasses of Thing, and also many resources concepts, have been moved from Thing to another place. For example, the class '*domain_qu_biosynthese'* representing the domain of the property *qu_biosynthese* is subsumed by the class *domain_qu_analyse* describing the domain of the property *qu_analyse* (Fig.3) :

Because of it definition, the domain *domain_qu_biosynthese* has been moved from `owl:Thing` to *domain_qu_analyse:*

```
<owl:Class rdf:ID="domain_qu_biosynthese">
<owl:unionOf rdf:parseType="Collection">
<owl:Class
rdf:about="#hormones_substituts_hormones" />
<owl:Class
rdf:about="#enzymes__coenzymes__anti_enzymes" />
<owl:Class
rdf:about="#glucides_et_hypoglycemiants" />
<owl:Class
rdf:about="#acides_amines__peptides_et_proteines"
/>
<owl:Class
rdf:about="#nucleosides_et_nucleotides" />
<owl:Class
rdf:about="#substances_biologiques_immunologiques
" />
</owl:unionOf>
</owl:Class>

<owl:Class rdf:ID="domain_qu_analyse">
<owl:unionOf rdf:parseType="Collection">
<owl:Class
rdf:about="#produits_chimiques_inorganiques" />
<owl:Class
rdf:about="#composes_chimiques_organiques" />
<owl:Class
rdf:about="#composes_heterocycliques" />
<owl:Class
rdf:about="#hydrocarbures_polycycliques" />
<owl:Class
rdf:about="#hormones__substituts_hormones" />
<owl:Class
rdf:about="#agents_regulateurs_reproduction" />
<owl:Class
rdf:about="#enzymes__coenzymes__anti_enzymes" />
<owl:Class
rdf:about="#glucides_et_hypoglycemiants" />
<owl:Class rdf:about="#lipides_et_hypolipemiants"
/>
<owl:Class
rdf:about="#acides_amines__peptides_et_proteines"
/>
<owl:Class
rdf:about="#nucleosides_et_nucleotides" />
<owl:Class
rdf:about="#agents_systeme_nerveux_central" />
<owl:Class
rdf:about="#agents_systeme_nerveux_peripherique"
/>
<owl:Class
rdf:about="#agents_cardiovasculaires" />
<owl:Class rdf:about="#antiinfectieux" />
<owl:Class
rdf:about="#antineoplasiques_et_immunodepresseurs
" />
<owl:Class
rdf:about="#produits_dermatologiques" />
<owl:Class
rdf:about="#substances_biologiques_immunologiques
" />
<owl:Class
rdf:about="#materiaux_biomedicaux_et_dentaires" /
>
<owl:Class
rdf:about="#drogues_et_agents_divers" />
<owl:Class
rdf:about="#actions_chimiques_et_utilisations" />
</owl:unionOf>
</owl:Class>
```

CONCLUSION AND FUTURE WORK

Like the Gene Ontology migration[23], the MeSH formalization is a several steps process. This paper has presented the first steps achieved to transform the MeSH thesaurus into OWL-DL. The main contributions are its modeling principles, such as the distinction between *is-a* and *part-of* hierarchies, between concepts denoting different notions, the elicitation of properties domains etc., which support the automatic process. These first steps aiming at being automatic, are mainly based on syntactic transformations, achieved from the existing MeSH hierarchical organization. For the moment, this one has only been partly enhanced, but we are aware that a more "semantic" step, based on a careful investigation, is still needed and further improvements are planed. For example, particular links in the *Anatomy* hierarchy should be fixed, and defined as "is-a" relations instead of "part of": the MeSH sub-trees A11 (cells), A12 (fluids and secretions) and A15 (hemic and immune systems) are *is-a* hierarchies, and "*blood cell"* [A11.118] *is-a* "*cell"* [A11]. Other problems come from the MeSH 'is-a' hierarchies, that are not really well principled. For example *diagnosis_error* is defined in the MeSH, thus in consequence also in our OWL ontology, as a specialization of *diagnosis* and *medical_ error*, although an error *****is not***** a diagnosis. Instead, the concept *diagnosis_error* should be defined as a *medical_error "*about" a *diagnosis*, thus represented in OWL by *medical_error* $\cap \forall$*about.diagnosis,* instead of their conjunction. A possibility to improve it and obtain such descriptions, is to use the UMLS Semantic Network relations, for instance like *is_complicated_by*, *is_treated_by* etc. for the diseases hierarchy. In addition, other properties, such as classical metadata (title, authors, format

etc…), may be added to the concepts that describe the resources. The next steps of this project will be to enhance the OWL representation, to define all the individuals (resources), to use the retrieval reasoning service for query processing. Such a formal ontology issued from the MeSH, is promising and may be exploited in many applications, based on the MeSH thesaurus, mainly bibliographic databases such as Medline, and health gateways [19-22].

References

[1] Darmoni, SJ., Thirion, B., Leroy, JP. et al. (2001). A Search Tool based on 'Encapsulated' MeSH Thesaurus to Retrieve Quality Health Resources on the Internet. *Medical Informatics & the Internet in Medicine*, 26(3):165-178.

[2] Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34-43.

[3] Nelson, SJ., Johnson, WD., Humphreys, BL. (2001) Relationships in Medical Subject Headings. In Bean and Green (Eds), 171-184.

[4] Soualmia, LF., Barry, C., Darmoni, SJ. (2003). Knowledge-Based Query Expansion over a Medical Terminology Oriented Ontology. Dojat, Keravnou, Barahona (Eds.), *LNAI # 2780*, Springer-Verlag, p.209-213.

[5] Golbreich, C. (2003) Towards a Sophisticated Multimedia Documents Search Engine. *Bulletin AFIA,* n°55, 40-54.

[6] Horrocks, I., Patel-Schneider, PF., van Harmelen, F. (2003) From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*. 1(1):7-26, 2003

[7] Noy, NF., Sintek, M., Decker, S., et al. (2001) Creating Semantic Web Contents with Protege-2000. *IEEE Intelligent Systems* 16(2):60-71.

[8] Haarslev, V., Möller, R. (2001) Description of the RACER System and its Applications. In *International Workshop in Description Logics 2001 (DL2001),* Stanford.

[9] Sowa, JF. (2000) Ontology, Metadata and Semiotics. B.Ganter, G.W.Mineau (Eds), Conceptual Structures: Logical, Linguistic, and Computational Issues, *LNAI #1867*, 55-81.

[10] Baker, T. (2000) A Grammar of Dublin Core. *Digital-Library Magazine,* vol 6 n°10.

[11] Mayer, MA., Darmoni, SJ., Fiene, M., et al. (2003). MedCIRCLE Modeling on the Semantic Web. Surjan, Engelbrecht, McNair (Eds) *Stud. Health Technol. Inf.* 95:667-672.

[12] Schulz, S. Hahn, U.(2001) Medical Knowledge Re-engineering – converting major portions of the UMLS into a terminological knowledge base. *IJMI*, 64(2-3):207-221.

[13] Horrocks, I., Rector, A. (1997) Experience Building a Large, Re-usable Medical Ontology using a Description Logic with Transitivity and Concept Inclusions. *Workshop on Ontological Engineering AAA Spring Symposium.*

[14] Cornet, R., Abu-Hanna, A. (2002) Usability of Expressive Description Logics – A Case Study in UMLS. *AMIA 2002,* 180-184.

[15] Kashyap, V., Borgida, A. (2003) Representing the UMLS Semantic Network using OWL. *ISWC 2003.*

[16] Darmoni, SJ., Jarousse, E., Zweigenbaum, P., et al. VuMeF: Extending the French Involvement in the UMLS Metathesaurus, *AMIA 2003,* 824.

[17] Baader, F, Calvanese, D., McGuinness,D., et al (2003) The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press.

[18] Golbeck, J., Fragoso, G., Hartel, F., et al. (2003) The National Cancer Institute's Thésaurus and Ontology. *Journal of Web Semantics*.

[19] Hersh, WR., Brown, KE., Donohoe, LC., et al. (1996) CliniWeb: managing clinical information on the World Wide Web. *JAMIA,* 3(4):273-80.

[20] Norman, F. (1998) Organising Medical Networks' information. *Med. Inf.* 23:43-51.

[21] Boyer, C., Baujard O., Baujard, V., et al. (1997) Health On the Net automated database of Health and medical information. *IJMI* 47(1-2):27-9.

[22] Deacon, P., Smith, JB., Tow, S. (2001) Using metadata to create navigation paths in the HealthInsite Internet gateway. *Health Info Libr J.* 18 (1) piii: 20-9.

[23] Wroe, C.J., Stevens R.., Goble C.A., Ashburner M.. A Methodology To Migrate The Gene Ontology To A Description Logic Environment Using DAML+OIL. *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB)*, Hawaii. January 2003.

[24] Lindberg DAB, Humphreys BL, McCray AT, The Unified Medical Language System. Meth Inform Med, 1993, 32(4): 281-91

Examining SNOMED from the Perspective of Formal Ontological Principles: Some Preliminary Analysis and Observations

Kent A. Spackman, MD PhD[a], Guillermo Reynoso, MD MBA[b]

[a]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA.*

[b]*InTerm Medical Terminology Research Center, Foundation for Healthcare Informatics Research,Buenos Aires, Argentina.*

Abstract

The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) is a terminological resource designed to support electronic applications in health and medicine. Its design has evolved over a period of more than thirty years, and continues to evolve. Recently several authors working on formal ontological theory have observed that applying certain principles and constraints to terminology construction may result in a more consistent and useful terminology. In this paper we report on a preliminary analysis of SNOMED CT by two of its developers, from the perspective of a few such formal ontological principles, giving examples of prior design decisions that appear to be supported by these principles as well as examples of prior design decisions that may be at variance with them. We believe that design changes suggested by formal ontological principles have great potential for improving consistency. Empirical evidence of usefulness should accompany theoretically-inspired moves towards more fine-tuned representations of reality.

Introduction

The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) continues to evolve with a goal of being both theoretically well-founded and clinically useful. Recently, several authors have observed that applying formal ontological principles and constraints to terminology construction may result in a more consistent and useful terminology. In this paper we report on a preliminary analysis of SNOMED CT from the perspective of a few such formal ontological principles, giving examples of prior design decisions that appear to be supported by these principles as well as examples of prior design decisions that may be at variance with them. We believe that design changes suggested by formal ontological principles have great potential for improving consistency. We also agree with Welty and Guarino[1] that these changes can result in creation of additional concepts with the same or nearly the same *term*, and some users may tend to view this as duplicative and redundant. It is not possible to be sure *a priori* that all such design changes will improve the terminology's value to its users, and we believe, at least for SNOMED, that empirical evidence of usefulness should be sought to accompany theoretically-inspired moves towards more fine-tuned representations of reality.

The ability of SNOMED CT to scale as a global terminology to be used in heterogeneous scenarios depends on several key factors. A significant one is a documented concept model that enables users to use formalized methods for the development of local extensions or for effectively contributing feedback for collaborative refinement of the terminology. While the concept model underlying the development of SNOMED CT by merging SNOMED RT and CTV3 followed a set of design and modeling principles described elsewhere, those principles may need to be reconsidered in terms of recent advances and experience in the application of formal ontological analysis methodologies that facilitate the explication of the modeler assumptions and ontological decisions.

SNOMED Background

SNOMED Clinical Terms is the latest in a long series of works of terminology developed and distributed by the College of American Pathologists (CAP) for the purpose of encoding, storing, and retrieving information on disease and health. Beginning with the Systematized Nomenclature of Pathology (SNOP) in 1965, and continuing through expansion to the Systematized Nomenclature of Medicine (SNOMED) in 1976 and subsequent major editions

in 1979 and 1993, the CAP focused on making a practical and comprehensive terminology that could be used by manual coders as well as by computerized information systems. Beginning in the mid-1990's, the CAP embarked on a radical re-engineering of SNOMED with the understanding that manual coding would become a thing of the past, and that substantial changes were required to support increasingly sophisticated electronic systems in healthcare and public health. As a consequence of this re-engineering and substantial re-work, in cooperation with the Kaiser Permanente "Convergent Medical Terminology" (CMT) project, CAP published the SNOMED Reference Terminology (RT) in 2000[2]. An even larger transformation (more than doubling in size, expansion of the concept model and other features) occurred as a result of merging SNOMED RT with the UK National Health Service's (NHS) Clinical Terms version 3 (CTV3), resulting in the first release of SNOMED Clinical Terms (CT) in January of 2002. Since that time there have been an additional four releases, one every six months. In 2003, the US Government licensed SNOMED CT and the National Committee on Vital and Health Statistics (NCVHS) recommended it as the general terminology for patient medical record information in the US. In the UK, SNOMED CT is a draft national standard and a key element of the NHS National Program for IT. Thus SNOMED is not a theoretical academic exercise, but is being developed with serious expectations and demands for practical usability.

Purpose of SNOMED Clinical Terms

SNOMED Clinical Terms is a terminological resource designed to be implemented in software applications to represent clinically relevant information reliably and reproducibly. Through the use of this information, SNOMED CT enabled applications can support effective delivery of high quality healthcare to individual people and populations. SNOMED CT is an international, multilingual terminological resource that can also represent concepts and terms unique to particular organizations or localities.[3]

Guiding Principles of SNOMED Development and Maintenance

Ever since the 1960's, the College of American Pathologists has regarded coding and classification systems as a vital interest. In 2003, working together with their colleagues from the UK NHS, they reiterated their commitment and outlined several basic principles upon which ongoing work by CAP is premised.[4] Prominent among these principles are commitments to 1) clinical integrity and quality, 2) usefulness for support of patient care, patient safety, audit, research, analysis, and planning, 3) scientific validation, 4) sustainability, with direct input from stakeholders, 5) widespread adoption, 6) protection of legacy data, and 7) accommodation of local needs. These are all laudable and necessary commitments, but in reality there are many constraints on any organization's ability to approach perfection in all of these areas, and there are natural tensions between these principles that require pragmatic and ongoing tradeoffs and judgments. This balancing process is analogous to attempting to find a suitable path towards the optimum in a large but constrained search space. There are natural tensions between sustainability, requiring a significant ongoing commitment of resources, versus widespread adoption, requiring minimal barriers and therefore free access. Government support is the preferred means of resolving this tension. There are also natural tensions between clinical integrity/quality/validity, requiring a significant degree of complexity with ongoing changes (enhancements, it is hoped), versus widespread use with protection of legacy data, requiring simplicity, face validity, and careful attention to backwards compatibility. It is in the context of this tension that analyses based on formal ontological principles must be placed, since one cycle's new formalisms, full of promise to "clean up" our problems, may become the next cycle's follies. SNOMED is demonstrably in this for the long haul, so changes will require due deliberation.

Evolutionary Design

Clinical terminology is difficult, and it is unreasonable to expect it ever to be perfect.[5] Rather than an excuse for ignoring problems in the terminology, this is a recognition that the design must adapt and change in order to continue to serve the needs for which it is intended. Campbell's influential work provided the basic evolutionary design principles upon which SNOMED development is still based.[6] There are six main points:

1. Evolution without pre-ordained design
2. Accumulation of design
3. Heterogeneity
4. Participatory consensus-based approach
5. Semantics-based concurrency control
6. Configuration management

As SNOMED development has continued, these broad principles have been operationalized using three fundamental criteria, abbreviated as "URU". The initials stand for understandability,

reproducibility, and usefulness. The first criterion, understandability, makes reference to whether a concept (or other design feature of the terminology) can be fully and unambiguously comprehended by users of the terminology. Understandability is tested by checking to see whether users believe they can tell whether the concept is relevant or not relevant to a given patient or situation. It is clear from this subjective test of understandability that two individuals may believe they understand what is meant, but their understanding may differ significantly. This leads to the need for the second criterion: Reproducibility indicates whether multiple users apply the concept to the same situations. Tests of reproducibility generally depend on independent modeling or coding followed by comparison. Finally, usefulness refers to the level of helpfulness and appropriateness conveyed in a concept or feature. A challenge for clinical terminologies is the need to provide explanation to naive users in order to make a sophisticated and complicated terminology accessible and useful.

The description logic definitions used to classify SNOMED CT support conjunction, existential restrictions, role hierarchies and the SNOMED CT notion of role groups, which can be represented using existential role restrictions in any description logic (DL) language.[7] This set of concept constructors is a small subset of DL features compared with the expressivity of $\mathcal{ALC}$, $\mathcal{SHIQ}$, and others. Future significant changes in the concept model might depend on the support of concept constructors like disjunction, negation and transitive properties. Classification tests have shown that the supertype and subtype relationships inferred by any correct and complete classifier will match those obtained and distributed in the SNOMED tables.

Basic Definitions and Ontological Principles

Here we briefly review some of the definitions and principles that have been proposed for subjecting terminologies to formal analysis. Guarino and Welty have proposed a set of principles collectively known as the OntoClean methodology.[8,9,10] This methodology appears to be gaining acceptance as a guidance and evaluation framework. Fundamental to the method is the idea of a *property*, a term roughly corresponding to what is ordinarily called a *concept* in taxonomies and description-logic based terminologies like SNOMED. From an ontological perspective, SNOMED's concepts such as *disorder*, *substance* and *organism* might be called *properties*. To quote Guarino and Welty, "In this paper we show how a formal ontology of unary properties (corresponding to concepts in taxonomies) can help using the subsumption relation in a disciplined way."[8] In this view, we distinguish properties like *organism* from the real-world bearers of those properties (actual organisms).

OntoClean provides formal definitions of meta-properties, which are a group of special properties characterizing other properties. These meta-properties (see examples below) help in the explication of the intended meaning of concepts from a formal ontological point of view. The ability to derive constraints on subsumption from the value assigned to these meta-properties provides assistance in the evaluation of modeling decisions.

Meta-properties

Meta-properties define characteristics of properties by saying what is or is not necessarily true of the instances of those properties. Here we restate the definitions of four of the meta-properties and provide examples attempting to convey an intuitive understanding of what is meant; readers should refer to the primary sources in the references for formal definitions.[8,9,10]

Rigidity: Guarino and colleagues define rigidity as a property that necessarily holds for all its instances in any instant of time and in every possible world. For example, *dog* is a rigid property because all instances of dog must always be dogs; they cannot be a dog at one time and not a dog at another. On the other hand, *pet* is called anti-rigid, meaning that instances of pet are not necessarily pets, since they could cease being a pet when, for example, they no longer have an owner. This is assuming that what we mean by *pet* makes it dependent on being owned, so a pet dog that no longer has an owner is a stray, not a pet, but it must remain a dog.

Identity: This meta-property aims to characterize what is unique for an entity that allows it to be identified, or re-identified, in different times and places. A property is said to carry an identity criterion if all its instances can be re-identified by a criterion that judges sameness. For example, the property *organism* is said to carry an identity criterion, since any instance of organism can be identified as being the same across time, based on biological criteria. On the other hand, the property *asymmetric* would be said not to have an identity criterion, since it is not possible to define criteria to

determine whether two instances of asymmetry are necessarily the same.

Unity: To hold the *unity* meta-property, every instance of a property must be an intrinsic whole. The determination of wholeness can depend on topological wholeness, or, alternatively, on a morphological, functional, or other relation. The relation that determines that a property carries a unity condition is called its equivalence relation. A property is said to have *anti-unity* if all its instances are not intrinsic wholes. For example, the property *water* has anti-unity, because there is no sense in which one can specify a relation that determines that instances of water are whole. In contrast, the property *lake* (which consists of water but is not an instance or subtype of water) can have a topological relation that defines its whole (based on the boundaries of the lake bed and the surface of the lake), and can therefore carry the unity meta-property.

Dependence: The meta-property *dependence* implies that all the instances of a given property require the existence of some instance of another entity that is not part of the former. For example, the property *mother* requires the existence of a child (at some point in time), and therefore is dependent. In contrast, the property *female* is independent.

Continuants and Occurrents

In addition to the meta-properties defined in the OntoClean methodology, we believe the distinction between continuants and occurrents, as defined, for example, by Smith and colleagues, provides potentially valuable insights for structuring clinical terminology resources like SNOMED CT.[11,12,13,14] A continuant is an entity that has no temporal parts, and therefore can be understood to exist in a slice of time. Objects, persons, substances, and locations are all in this category. On the other hand, occurrents have temporal duration. Procedures, processes and movements fit into this category.

Taxonomic constraints

The value of distinguishing OntoClean's meta-properties, and the fundamental properties of continuant and occurrent, is that these provide perspectives that enable us to eliminate inconsistencies in terminology hierarchies based on subsumption constraints. In other words, the is-a relation should behave in a consistent manner, and these constraints help us to identify possible inconsistencies and eliminate them.

The following seven constraints represent merely a subset of all possible constraints that might be generated by formal ontological analysis. However, these appear to be important and potentially very useful. If we let "x $\not\subset$ y" mean that properties (concepts) having meta-property x should not be subsumed by (should not have an "is-a" relationship to) any property having meta-property y, then the first five in this list are restatements of Guarino's constraints,[8] and the last two are restatements of, for example, constraints expressed by Fielding[11] and Smith[13].

1. Unity $\not\subset$ Anti-unity
2. Non-unity $\not\subset$ Unity
3. Rigidity $\not\subset$ Anti-rigidity
4. Non-identity $\not\subset$ Identity
5. Independent $\not\subset$ Dependent
6. Continuant $\not\subset$ Occurrent
7. Occurrent $\not\subset$ Continuant

Examples of Taxonomic Constraints

Example concepts: aspirin (product), aspirin (substance)
Constraint: Unity $\not\subset$ Anti-unity

Aspirin (acetylsalicylic acid, ASA) is used to name an ingredient and also to name a class of prepared product that contains ASA. It is an exemplar of a systematic decision in SNOMED to separate ingredient substances from the products of which they are made, even though they have the same name. Ingredient substances would be properties with anti-unity, but the products of which they are made would be properties with unity. Thus formal ontological principles confirm our decision to separate them, but there is evidence that not everyone agrees with the decision. In particular, the editors of the US National Library of Medicine's Unified Medical Language System (UMLS) Metathesarus, which incorporates SNOMED CT into its structure, have decided to leave these two SNOMED codes (aspirin product, aspirin substance) linked to the same concept unique identifier (CUI), and likewise for all other product/substance pairs in SNOMED. (B. Humphreys, personal communication). In other words, they are representing a concept "aspirin" that does not differentiate between the drug product itself and the stuff of which it is made. We also considered this approach because it initially appeared simpler (one concept instead of two), but eventually rejected it because of the difficulties it creates in correctly representing subsumption hierarchies of drug

ingredients and drug products. We recognized that we have to accept, and explain to our users, that there will be two concepts carrying the simple name *aspirin* and they will have to choose between them because they are truly different. We provide the (substance) and (product) phrases in the fully specified name to help users to see the difference.

This example nicely illustrates the kinds of tensions that sometimes arise between the requirements of formal rigor and the (apparent) requirements of commonsense thinking and simplicity.

Parenthetically, this mismatch between SNOMED and UMLS once again confirms the interpretation of the UMLS CUI, proposed by Campbell et al,[15] that claims that it must be viewed as representing extensional meaning since it will not always match the intensional meaning of its source vocabularies.

Example concepts: infectious agent, bacterium
Constraints: Rigid ⊄ Anti-rigid, Independent ⊄ Dependent

The property *infectious agent* would be called a *role* in OntoClean. It is an anti-rigid property (in some possible world, all instances can possibly be non-infectious) and dependent on an infectious relationship between the agent and an infected (or perhaps infectable) organism. The property *bacterium*, on the other hand, would be called a *type*. It is a rigid property, carrying identity, and independent. *Infectious agent* currently subsumes *bacterium* in SNOMED. The taxonomic constraints suggest this is inconsistent and should be changed. Once again, there is tension between simplicity and usability on the one hand and formal rigor on the other, since practical use calls for a simple categorization of infectious agents, and the simplest solution appears to be an is-a relationship from bacteria, fungi, parasites, viruses and prions to infectious agent. However, we agree that this role vs. type distinction provides a useful criterion to untangle the taxonomic primitive backbone which, as noted by Welty and Guarino[1] should consist only of rigid properties, although strict adherence to this idealized structure may not always be possible.

This distinction also helps to prioritize the incorporation of new attributes into the SNOMED CT concept model. Adding new attributes results in a more faithful representation of meaning and avoids inconsistencies. As another example of the use of attributes to eliminate incorrect is-a's, consider the relationship of *insulin*, *hormone*, and *antidiabetic agent*, as in Alan Rector's tutorial:[16]

Insulin ⊆ (Antidiabetic agent)
Antidiabetic agent ⊆ (protein ∩
∃hasFunction.HormonalAction)

Could instead be modeled as:
Insulin ⊆ (protein ∩
∃hasFunction.AntidiabeticAction)
Hormone = (substance ∩
∃hasFunction.HormonalAction)

Certain semantic categories like *physical object*, *social context*, *substance* and *organism* may benefit from this kind of analysis focused on the type/role distinction in the upper level hierarchy.

Other semantic categories like *finding*, *disorder* and *procedure* may realize less benefit from this methodology, and therefore the case for applying it should be based on future research.

Example concepts: morphologic abnormality, pathological process.
Constraint: Continuant ⊄ Occurrent

Early work on SNOMED RT involved significant discussion and consensus-building resulting from dual independent modeling (dissection, definition) of concepts, followed by examination of differences, as a means of seeking reproducibility. An early disagreement arose about *acute inflammation*. From a clinical examination perspective this term described the combination of redness, pain, swelling and heat of an inflammatory process. From a histological perspective this term described the existence of an infiltrate of acute inflammatory cells. Although described using the same words, the two concepts are very different. If we assume one meaning refers to an acute inflammatory process, an occurrent, and the other meaning refers to an acute inflammatory infiltrate, a continuant, then it is clear we need two different codes and that neither can subsume the other. In fact, they should be in totally different hierarchies: the structure in the morphologic abnormality hierarchy, and the process in the pathological process hierarchy. It is instructive to realize (and useful to apply as a general rule) that the process – structure distinction provided by the words *infiltration* and *infiltrate* may not be provided by the words used commonly for other situations. In this example, we speak of *inflammation* but not "inflammate"; instead we use the same word for both meanings. This re-emphasized for us a universally known but frequently forgotten lesson that simply using the same words, even technically detailed words, is no guarantee of meaning the same thing.

Example concepts: morphologic abnormality, disorder
Constraint: Continuant ⊄ Occurrent

Early in the Kaiser CMT project, there was a discussion about whether SNOMED III morphology (M) codes and SNOMED III disease (D) codes meant the same thing. There was consensus that there is a clear difference between the structural "snapshot" of a disease that is observed by a pathologist in a tissue section, and the temporally extended disease that is experienced by the patient. This decision accounts for the apparent duplication of codes that have very similar names, one for morphologic abnormality and the other for disorder. It seems fairly easy to explain that we have both "neoplasm, benign (morphologic abnormality)" and "benign neoplastic disease (disorder)". But it is more difficult, to some, to follow the same logic to "Burkitt lymphoma (morphologic abnormality)" and "Burkitt's lymphoma (disorder)". Commonsense thinking seems to lead people to believe that we should have only one code for Burkitt's lymphoma.

Singular versus Plural Naming

Formal ontologies carefully distinguish between single whole entities and groups of wholes. Typical thesaurus construction tends to use plurals to describe more general classes, in order to signal to users that the code represents a category. SNOMED III (1993) used plurals in this way for "headers", which were published in uppercase and carried a data field that set them apart (Eclass=00). CTV3 routinely used plurals for higher-level categories and singular tense at lower levels. The transition to SNOMED CT was accompanied by a systematic effort, unfortunately still incomplete, to convert these names to singular tense unless the intended meaning actually implied multiples. For example, "infiltrations (procedure)" is the general procedure subsuming procedures such as "intradermal infiltration of steroid (procedure)". "Infiltration (procedure)" should be its name. Although this decision helps SNOMED align better with an ontological rigor, some users have told us they would like us to present the hierarchies using plurals for the upper level categories, because they feel it would look better. This may result from a mental habit of using hierarchies to name a set of things, and then name the things in the set.

Beyond the necessary attention to singular-plural naming, there is additional attention required to differentiate wholes from collections.

Anatomy and Mereology

Although beyond the scope of this paper, we also want to mention the anatomy model that SNOMED has used to support classification of findings, disorders and procedures. The use of structure-entire-part (SEP) triplets to represent anatomy was inspired by the challenges of combining SNOMED RT and CTV3. More theoretically-oriented justification of the SEP model were independently developed by Schulz et al.[17] This pragmatically-oriented model of body structures has been sufficient for the great majority of concept definitions. While the SEP triplet implementation present in SNOMED CT has been a significant improvement to the SNOMED RT model, certain aspects like the relationship between microscopic - macroscopic structures and part – region modeling would require further analysis. The mereology foundations have a significant impact in how semantic categories are structured by the classifier, and therefore merit further research.

Some Observations and Discussion

In a large terminology such as SNOMED (over 350,000 concepts), there are bound to be errors and inconsistencies from the perspective of formal ontology. In order to properly understand their source and determine what to do about them, it is necessary to know whether they are attributable to design decisions and therefore intentional, or are unintentional errors that, with ongoing maintenance, are being eliminated without the need for further design work. Attempting to determine these questions by examining the terminology alone without asking the developers about the current state of the design, is like reading tea leaves: it is highly subjective, unreliable, and prone to misinterpretation.

OntoClean appears to bring good organizing principles to the modelling of several of SNOMED's semantic categories such as physical object, social context, organism, and substance. In general, untangling these hierarchies is very desirable. The applicability of some of the formal ontology principles in providing consistent guidance in the very large areas like clinical finding and procedure is not as clear, and appears to require further elaboration of ontological foundations as well as further study of the impact of newly proposed distinctions on the structure of the terminology. Yet these represent the most numerous and important of SNOMED's content. Since ontological analysis is

labor-intensive, implementing its constraints in certain areas of SNOMED or adding modelling criteria should be based on proven benefits and/or pilot experiences. Meeting the needs of users need not be in conflict with detailed faithful representation of reality. However, in our experience, users often want things that turn out to be incompatible:

1. They want it useful and able to satisfy their functional requirements
2. They want it the way they like it and think it should be.
3. They want it correct and theoretically sound at the same time.

Physicians are often criticized for unnecessarily using arcane and complicated language to describe clinical situations in a way inaccessible to the average patient or their family. Although physicians do sometimes purposefully obfuscate, and rarely justifiably do so, the usual reason for their choice of words derives from habit and a desire to be more precise than is possible with layman's terms. If formal ontologists are the would-be healers of terminology systems, they face an analogous criticism, and an analogous dilemma. New ideas often call for special language, but formal ontologists face a challenge at least as great as that of physicians in communicating their ideas in an accessible way to those who might make use of them. For ordinary clinicians, jargon like "endurant" and "rigid property" can mislead and obfuscate. Because of the need to have ordinary medical practitioners involved in the development and use of clinical terminology, restricting the process of clinical terminology development to a narrow group of ontological practitioners formally trained in philosophy is not a sensible way forward, and therefore the ideas of formal ontology must be communicated in a clear and understandable way. This remains an ongoing challenge.

Formal ontological principles can help to make distinctions understandable and reproducible. Some distinctions will not be useful for electronic health records or decision support, and we need to guard against a tendency towards arcane and complicated distinctions that are inaccessible to all but the most sophisticated users. However, SNOMED is actively embracing and exploiting methodologies that are shown to improve its quality and usefulness.

Acknowledgments

This work is supported in part by a grant from the College of American Pathologists.

References

[1] Welty C and Guarino N. Support for ontological analysis of taxonomic relationships. *J. Data and Knowledge Engineering*. 2001;39(1):51-74.

[2] Spackman KA, Campbell KE, Cote RA. SNOMED RT: A reference terminology for health care. *Proceedings AMIA Annual Symposium*. 1997; :640-4.

[3] College of American Pathologists. SNOMED Clinical Terms Requirements Document. 2001; http://www.snomed.org/about/documents/SNOMEDCT_Objective_V05.pdf

[4] College of American Pathologists. Commitment to Health Care. 2003;www.snomed.org/about/index.html

[5] Rector A. Clinical terminology: Why is it so hard? *Meth. Inf. Med*., 1999; 38:147-157.

[6] Campbell KE. Distributed development of a logic-based controlled medical Terminology. PhD Dissertation, Stanford University, June 1997.

[7] Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog motivated by concept modeling in SNOMED. *Proceedings AMIA Annual Symposium*. 2002; :712-716.

[8] Guarino N, Welty C, A formal ontology of properties, in R. Dieng (Ed.), *Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management*, Springer-Verlag LNCS, Berlin, Germany, 2000

[9] Guarino N and Welty C. Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*. 2002;45(2):61-65.

[10] Guarino N and Welty C. Ontological analysis of taxonomic relations. In A. Lander and V. Storey (eds.) *Proceedings of ER-2000: The International Conference on Conceptual Modeling*. 2000; Springer Verlag LNCS Vol. 1920.

[11] Fielding JM, Simon J, Smith B. Formal ontology for biomedical knowledge systems integration. http://ontology.buffalo.edu/medo/FOBKSI.pdf.

[12] Schulze-Kremer S, Smith B, Kumar A. Revising the UMLS semantic network. http://ontology.buffalo.edu/medo/UMLS_SN.pdf.

[13] Smith B, SNAP and SPAN. Actor2002 Action-oriented approaches in geographic information science. http://www.spatial.maine.edu/~actor2002/participants/smith-pres.pdf

[14] Smith B, Basic Formal Ontology and Medical Ontology, http://ontology.buffalo.edu/bfo/ BFO.htm.

[15] Campbell KE, Oliver DE, Spackman KA, and Shortliffe EH, Representing thoughts, words, and things in the UMLS. *J Am Med Inform Assoc*. 1998; 5(5): 421–431.

[16] Rector A, A tutorial on creating biomedical ontologies in DAML+OIL using OilEd http://www.cs.man.ac.uk/mig/ontology-tutorial/oiled-biomedical-ontology-tutorial.zip

[17] Schulz S, Romacker M, Hahn U. Part-whole reasoning in medical ontologies revisited - Introducing SEP triplets into classification-based description logics. *Proceedings AMIA Annual Symposium*. 1998; :830-4.

Using C-OWL for the Alignment and Merging of Medical Ontologies

Heiner Stuckenschmidt[1], Frank van Harmelen[1]
Paolo Bouquet[2,3], Fausto Giunchiglia[2,3], Luciano Serafini[3]

[1] Vrije Universiteit Amsterdam
[2] ITC-IRST, Trento
[3] DIT - University of Trento

January 28, 2004

Abstract

A number of sophisticated medical ontologies have been created over the past years. With their development the need for supporting the alignment of different ontologies is gaining importance. We proposed C-OWL, an extension of the Web Ontology Language OWL that supports alignment mappings between different, possibly incompatible ontologies on a semantic level. In this paper we report experiences from using C-OWL for the alignment of medical ontologies. We briefly review key concepts of the C-OWL semantics, explain the setting of the case study including some examples from the alignment and discuss the possibility of reasoning about the mapping based on the C-OWL semantics We conclude by arguing that C-OWL provides an adequate framework for aligning complex ontologies in the medical domain.

Keywords: Biomedical Knowledge representation, validation and maintenance; Knowledge Representation Languages; Terminology Integration

1 Introduction

The need for terminology integration has been widely recognized in the medical area leading to a number of efforts for defining standardized terminologies. It is, however, also acknowledged by the literature, that the creation of a single universal terminology for the medical domain is neither possible nor beneficial, because different tasks and viewpoints require different, often incompatible conceptual choices [Gangemi et al., 1998]. As a result a number of communities of practice have been evolved that commit to one of the proposed standards. This situation demands for a weak notion of integration, also referred to as alignment in order to be able to exchange information between the different communities.

In [Bouquet et al., 2003] we argued that the current design of the web ontology and its semantics is not suitable for situations where different view on the same domain have to be aligned in a loose way. We proposed an extension of the OWL semantics that allows the specification of semantic relations between different OWL models. The resulting notion of contextualized ontologies can provide such an alignment by allowing the co-existence of different, even in mutually inconsistent models that are connected by semantic mappings. The nature of the proposed semantic mappings satisfies the requirements of the medical domain, because they do not require any changes to the connected ontologies and do not create logical inconsistency even if the models are incompatible.

This paper is organized as follows. We first briefly review the central definitions of the extended OWL semantics. In particular, we introduce the notion of local domains and mappings between them as well as their formal interpretation. In section 3 we describe the setting of a case study we conducted in using OWL to define and reason about alignments of medical ontologies and present some examples from the alignment. The use of C-OWL for reasoning about alignments is discussed in section 4. We conclude with a summary of our experiences and a discussion of the role of C-OWL for terminology integration in the medical domain.

2 Contextual semantics for OWL

The main idea of the proposed contextual semantics for OWL is split to up the global interpretation of different OWL ontologies into a set of local interpretations for each ontology. In order to make the alignment of ontologies with contradicting definitions possible, the notion of a hole is introduced which makes every statement in an ontology satisfiable. As a consequence statements are allowed to hold in one ontology but not in another one[1].

Definition 1 (OWL interpretation with local domains) *An* OWL interpretation with local domains *for a set of OWL ontologies* $\{\langle i, O_i\rangle\}_{i\in I}$*, is a family* $\mathcal{I} = \{\mathcal{I}_i\}_{i\in I}$*, where each* $\mathcal{I}_i = \left\langle \Delta^{\mathcal{I}_i}, (.)^{\mathcal{I}_i}\right\rangle$*, called the* local interpretation *of* O_i*, is either an interpretation of* O_i *into* $\Delta^{\mathcal{I}_i}$*, or a hole.*

The definition above completely separates the interpretations of different ontologies. As our aim is, however, to represent and reason about alignment between different ontologies, we have to introduce a way of connecting their domains. C-OWL does this by means of so-called bridge rules that define the semantic relations between concepts in different ontologies. C-OWL defines the following kinds of bridge rules stating that a concept from an ontology O_i is more general, more specific, equivalent, disjoint or overlapping with a concept from another ontology O_j:

$$i{:}x \xrightarrow{\sqsubseteq} j{:}y, \quad i{:}x \xrightarrow{\sqsupseteq} j{:}y, \quad i{:}x \xrightarrow{\equiv} j{:}y, \quad i{:}x \xrightarrow{\perp} j{:}y, \quad i{:}x \xrightarrow{*} j{:}y,$$

A mapping between two ontologies is a set of bridge rules between them. A context space is a pair composed of a set of OWL ontologies $\{\langle i, O_i\rangle\}_{i\in I}$ and a family $\{M_{ij}\}_{i,j\in I}$ of mappings from i to j, for each pair $i, j \in I$. To give the semantics of context mappings the definition of an OWL interpretation with local domains is extended with the notion of *domain relation*. A domain relation $r_{ij} \subseteq \Delta^{\mathcal{I}_i} \times \Delta^{\mathcal{I}_j}$ states, for each element in $\Delta^{\mathcal{I}_i}$ to which element in $\Delta^{\mathcal{I}_j}$ it corresponds to. The semantics for bridge rules from i to j can then be given with respect to r_{ij}. The interpretation for a context space is composed of an OWL interpretation with holes and local domains and the an interpretation domain relation from i to j, which is a subset of $\Delta^{\mathcal{I}_i} \times \Delta^{\mathcal{I}_j}$. As suggested above, the definition of bridge rules introduces semantic relationships between concepts in different ontologies thereby constraining

[1]For technical details about interpretations with holes see [Bouquet et al., 2003]

the global interpretation. As the way bridge rules are interpreted is important with respect to the possibilities for reasoning about alignments we give the formal definition of satisfiability of bridge rules.

Definition 2 (Satisfiability of bridge rules[2]) *Let $\mathfrak{I}$ be the global interpretation of a context space, then*

1. $\mathfrak{I} \models i:x \xrightarrow{\sqsubseteq} j:y$ *if* $r_{ij}(x^{\mathcal{I}_i}) \subseteq y^{\mathcal{I}_j}$;
2. $\mathfrak{I} \models i:x \xrightarrow{\sqsupseteq} j:y$ *if* $r_{ij}(x^{\mathcal{I}_i}) \supseteq y^{\mathcal{I}_j}$;
3. $\mathfrak{I} \models i:x \xrightarrow{\equiv} j:y$ *if* $r_{ij}(x^{\mathcal{I}_i}) = y^{\mathcal{I}_j}$;
4. $\mathfrak{I} \models i:x \xrightarrow{\perp} j:y$ *if* $r_{ij}(x^{\mathcal{I}_i}) \cap y^{\mathcal{I}_j} = \emptyset$;
5. $\mathfrak{I} \models i:x \xrightarrow{*} j:y$ $r_{ij}(x^{\mathcal{I}_i}) \cap y^{\mathcal{I}_j} \neq \emptyset$;

An interpretation for a context space is a model for it if all the bridge rules are satisfied.

3 Aligning Medical Ontologies: An Experiment in Using C-OWL

In the medical area a lot of work has been done on the definition and standardization of terminologies [3]. The result of these efforts is a large number of medical terminologies and classifications. The complexity of the terminologies used in medicine and the strong need for quality control has also lead to the development of ontologies that feature complex concept definition (compare [Golbreich et al., 2003] for a discussion of the required expressiveness). Some of these ontologies are available in OWL and can be seen as the first OWL applications that have a use in real life applications. C-OWL and especially its formal semantics provides us with several possibilities concerning the alignment of the medical ontologies mentioned above.

[3] see e.g. http://www.medinf.mu-luebeck.de/ ingenerf/terminology/Index.html for a collection of standards

3.1 Alignment Scenario

In our Case study, we used available representations of the the following medical ontologies:

Galen The Motivation for the GALEN project [Rector and Nowlan, 1993] is the difficulty in exchanging clinical data between different persons and organizations due to the heterogeneity of the terminology used. As a result of the project, the GALEN Coding Reference model has been developed. This reference model is an ontology that covers general medical terms, relations between those terms as well as complex concepts that are defined using basic terms and relations. We used an OWL version of the GALEN model that contains about 3100 classes and about 400 relations.

Tambis The aim of the Tambis [Baker et al., 1999] (Transparent Access to Bioinformatics Information Sources) is to provide an infrastructure that allows researchers in Bioinformatics to access multiple sources of biomedical resources in a single interface. In order to achieve this functionality, the project has developed the Tambis Ontology, which is an explicit representation of biomedical terminology. The complete version of Tambis contains about 1800 terms. The DAML+OIL version we used in the case study actually contains a subset of the complete ontology. It contains about 450 concepts and 120 Relations.

UMLS The Unified Medical Language System UMLS [Nelson and Powell, 2002] is an attempt to integrate different medical terminologies and to provide a unified terminology that can be used across multiple medical information sources. Examples of medical terminologies that ave been integrated in UMLS are MeSH and SNOWMED. In our case study, we used the UMLS semantic network. The corresponding model that is available as OWL file contains 134 semantic types organized in a hierarchy as well as 54 relations between them with associated domain and range restrictions.

We assume that the goal is to establish a connection between the Tambis and the GALEN ontology in such a way that the two models with their different focus supplement each other. An option for aligning Tambis and GALEN is an indirect alignment based on a third, more general model

of the domain. In this setting the two models are made comparable by aligning each one with the third, more general model and using the semantic relations in this third model together with the mapping to determine the relation between classes in the two ontologies.

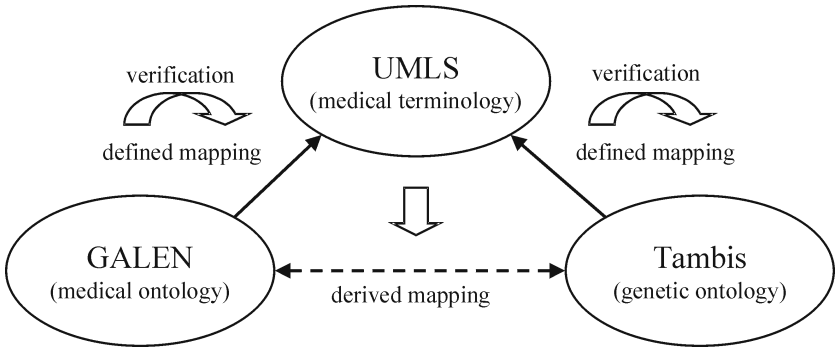


Figure 1: Indirect Alignment of Tambis and GALEN using UMLS

The UMLS semantic network is such a general model. Being the result of an integration of different medical terminologies(compare [Bodenreider, 2004]), we can assume that the network is general enough to cover the content of Tambis, GALEN and also other prospective ontologies that we might want to align. In order to explore the use of C-OWL for the alignment of medical ontologies, we conducted a small case study in aligning the ontologies mentioned above using the UMLS semantic network as a central terminology. We investigated the upper parts of the ontologies and identified areas with a sufficient overlap. Such an overlap between all three models exists with respect to the following three areas:

Processes: Different physiological, biological and chemical processes related to the functioning of the human body and to the treatment of malfunctions.

Substances: Substances involved in physiological processes including chemical, biological and physical substances.

Structures: Objects and object assemblies that form the human body or parts of it. Further, structures used in the treatment of diseases.

We analyzed the three models with respect to these three topics. Based on the comparison of the three models, we defined mappings between Tambis and GALEN and the UMLS terminology. These mappings consist of sets of bridge rules each connecting single concepts or concept expressions. In the following, we present some alignment examples from the case study. In particular we describe some of the alignment of GALEN and UMLS with respect to substances A more detailed description of the case study can be found in [Stuckenschmidt, 2004].

3.2 Examples from the Alignment

GALEN contains the notion of a generalized substance which is a notion of substance that subsumes substances in a physical sense and energy making it more general than the notion of substance in UMLS

$$\text{GeneralisedSubstance} \stackrel{\sqsupseteq}{\longleftrightarrow} \text{Substance}$$

The actual notion of substance as defined in GALEN is not as we might expect equivalent to the notion of substance in UMLS, because it also contains some notions that are found under anatomical structures in UMLS. We can, however, state that the GALEN notion of substance is more specific than the union of substances and anatomical structures in UMLS.

$$\text{Substance} \stackrel{\sqsubseteq}{\longleftrightarrow} \text{Substance} \sqcup \text{Anatomical\_Structure}$$

The next GALEN concept that also occurs in UMLS but has a slightly different meaning is the notion of body substance. The difference is illustrated in the fact that it also covers the notion of tissue which is found under anatomical structures in UMLS. We conclude that the notion of body substance in GALEN in a broader one than in UMLS.

$$\text{BodySubstance} \stackrel{\sqsupseteq}{\longleftrightarrow} \text{Body\_Substance}$$

The other main class of substances mentioned in GALEN are chemical substances. Looking at the things contained under this notion, we conclude that it is equivalent to the notion of chemical in UMLS.

$$\text{ChemicalSubstance} \stackrel{\equiv}{\longleftrightarrow} \text{Chemical}$$

We can also find the correspondences to the distinction between elementary and complex chemicals made by GALEN in UMLS. Elementary chemicals are a special case of the UMLS concept of elements ion or isotope.

$$\text{ElemetaryChemical} \overset{\sqsubseteq}{\longleftrightarrow} \text{Element\_Ion\_or\_Isotop}$$

Complex chemicals contain all kinds of chemical substances sometimes viewed structurally, sometimes functionally. Therefore, we cannot related this concept to one of these views taken by UMLS. We also notice that there are notions of complex chemicals in GALEN that do not occur under chemicals in UMLS - e.g. Drugs that related to the concept of clinical drug classified under manufactured objects.

$$\text{Drug} \overset{\equiv}{\longleftrightarrow} \text{Clinical\_Drug}$$

Further, the UMLS views on chemicals also contain elementary chemicals. Consequently, we can only define the notion of complex chemical to be compatible with the union of the two views in UMLS

$$\text{ComplexChemical} \overset{*}{\longleftrightarrow} \text{Chemical\_Viewed\_Structurally} \sqcup \text{Chemical\_Viewed\_Functional}$$

On the level of more concrete chemical notions we find a number of correspondences mentioned in the following. Named hormones are equivalent to hormones in UMLS

$$\text{NAMEDHormone} \overset{\equiv}{\longleftrightarrow} \text{Hormone}$$

Proteins are more specific than amino acids, peptides or proteins.

$$\text{Protein} \overset{\sqsubseteq}{\longleftrightarrow} \text{Amino\_Acid\_Peptide\_or\_Protein}$$

The notions of lipid and of carbohydrate are the same in the two models

$$\text{Lipid} \overset{\equiv}{\longleftrightarrow} \text{Lipid}$$

$$\text{Carbohydrate} \overset{\equiv}{\longleftrightarrow} \text{Carbohydrate}$$

There is an overlap between the notion of acid in GALEN and the concepts amino acid, peptide or protein and Nucleic acid , nucleosid or protein in UMLS.

$$\text{Acid} \overset{*}{\longleftrightarrow} \text{Amino\_Acid\_Peptide\_or\_Protein} \sqcup \text{Nucleic\_Acid\_Nucleosid\_or\_Protein}$$

Finally metals can be defined to be a special case of inorganic chemicals.

$$\text{Metal} \overset{\sqsubseteq}{\longleftrightarrow} \text{Inorganic\_Chemical}$$

In summary, we were able to find a lot of correspondences on the level of groups of chemicals. While the models disagreed on the higher level structuring of substances, they shared a lot of more concrete concepts. As a consequence, we found a number of equivalence and subsumption relationships between substances at a lower level while at the more general level, we often had to use weak relations or link to very general concepts.

4 Reasoning about Alignments

In the experiment, we defined mappings in a ad-hoc rather than a systematic fashion. Such an ad hoc approach for defining mappings bears the risk of inconsistency and in completeness. We cannot prevent the creation of inconsistent or incomplete mappings, but the semantics of C-OWL can be used to verify and extend a defined mapping in order to detect inconsistencies and implied mappings. In the following we give examples of the use of the C-OWL semantics to verify and extend the mappings between the substance information in the different medical ontologies.

4.1 Verification of Mappings

A mapping can become inconsistent if two classes who are known to overlap, e.g. because they are subclasses of each other, link to disjoint concepts in another model. An example of this situation can be found in the substance related part of the alignment between Tambis and UMLS. Figure 2 shows the situation. On the right hand side the extensions of the UMLS concept chemical substances and some of its subclasses are sketched. UMLS distinguishes between chemical from a structural and a functional view. In the case where these two views are defined to be disjoint (one can either take a structural or a functional view but not both) we get an inconsistency with the mappings defined for the Tambis

ontology, because the mappings claims that the image of the concept chemical is exactly the extension of the structural view. At the same time, we claim that the image of enzyme which is a subclass of chemical is exactly the extension of the UMLS concept Enzyme which is classified under the functional view on chemicals in UMLS and therefore disjoint from the structural view. This however is now possible in the C-OWL semantics as the image of enzyme is a subset of the image of chemical by definition.

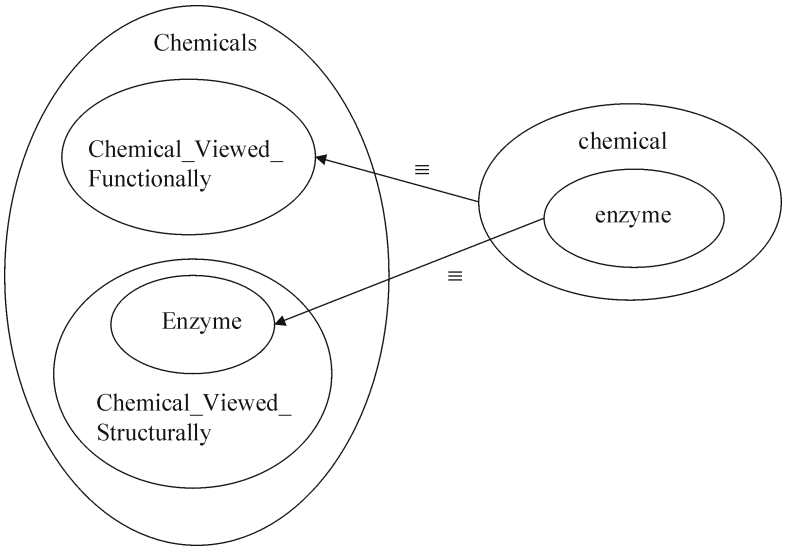


Figure 2: An Inconsistent Mapping

This ability to detect inconsistencies depends on the existence of appropriate disjointness statements in the ontology the mappings point to. Alternatively, the use of disjointness mappings can provide the same effect. If we want to make clear that chemicals in Tambis are not classified according to the functional view (which we just found to be not entirely true) we can also add a corresponding mapping stating that the image of chemicals is disjoint from the extension of the functional view on chemicals. The definition of this mapping will have the same effect leading to an inconsistency as described above.

4.2 Derivation of Semantic Relations

Besides the possibility to detect inconsistencies in the mappings, we can also infer additional bridge rules between the same models based on existing ones thereby making the complete mapping implied by the defined rules explicit. We illustrate this possibility by discussing possible implications of an equivalence mapping. Figure 3 illustrates parts of the alignment of substance related alignment of UMLS and GALEN. In particular, it shows the rule stating an equivalence between the GALEN class chemical and the UMLS class chemical substance which is part of the alignment. The definitions in UMLS state that chemical substances are less general than the class generalized substance, more general than complex chemicals and disjoint from processes. As the existing bridge rule states that the image of chemical is exactly the extension of chemical substance in UMLS, these relations also hold between this image and the other UMLS classes mentioned. The relations can be explicated by adding corresponding bridge rules stating that the image of chemicals is more general than complex chemicals, less general that generalized substance and disjoint from processes.

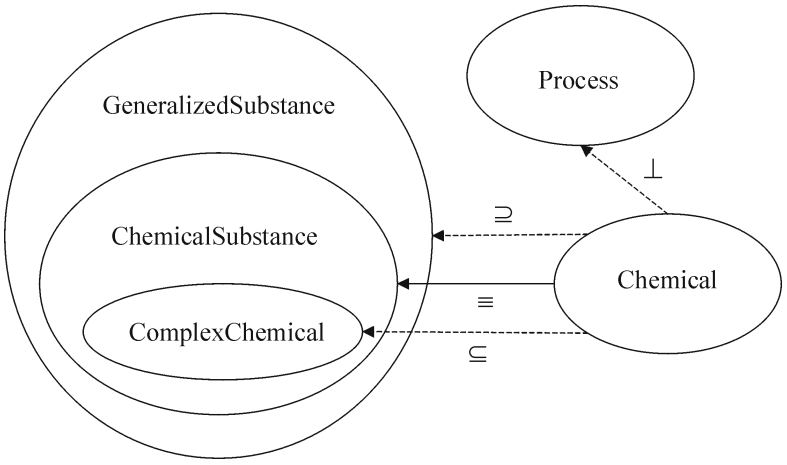


Figure 3: Derivation of additional Mappings

Similar inferences can be made based on bridge rules indicating specialization and generalization relations. If we replace the equivalence in figure 3 by a rule stating that chemicals is more specific than chemical substances, we

are still able to infer the relations to generalized substances and to processes. Just the one to complex chemicals will be lost, because the image of chemicals might only overlap or be disjoint from the extension of the respective concept. Conversely, replacing the equivalence by bridge rule stating that chemicals is more general than chemical substances would have preserved the conclusion that chemicals is more general than complex chemicals. Finally, stating that chemicals is disjoint from chemical substances would have implied that it is also disjoint from complex chemicals.

4.3 Merging Local Models

Another thing we would like to do based on the alignments is to compare the local models (Tambis and GALEN) with each other and derive semantic correspondences between classes in these models as well. It turns out that we cannot really drive mappings between the two local models from their mappings to UMLS, because referring to different interpretation domains, we cannot compare the constraints imposed by these mappings. This situation changes, however, when we assume that the local models are to be merged. In this case, their interpretation domain becomes the same and we can use the constraints to derive semantic correspondences between concepts in the two models from the existing mappings.

Figure 4 shows two examples of derived relations between concepts from GALEN and Tambis. The figure shows two concepts from each, UMLS (upper part), Tambis (lower left part) and GALEN (lower right part). We assume that we have fixed the inconsistency detected in the mapping from Tambis to UMLS by removing the bridge rule relating chemical substances to the structural view on chemicals and replacing it by an equivalence between chemical substance and chemicals in general. As the GALEN concept chemical is also defined to be equivalent to Chemical, we can derive that these two concepts are equivalent in the merged ontology. Further, we defined the notion of substance in Tambis to be more specific than the same notion in UMLS which is again defined to be more specific than generalized substance in GALEN. From these mappings, we can derive that the Tambis notion of substance is more specific than Generalized substance and add a corresponding axiom to the merged ontology.

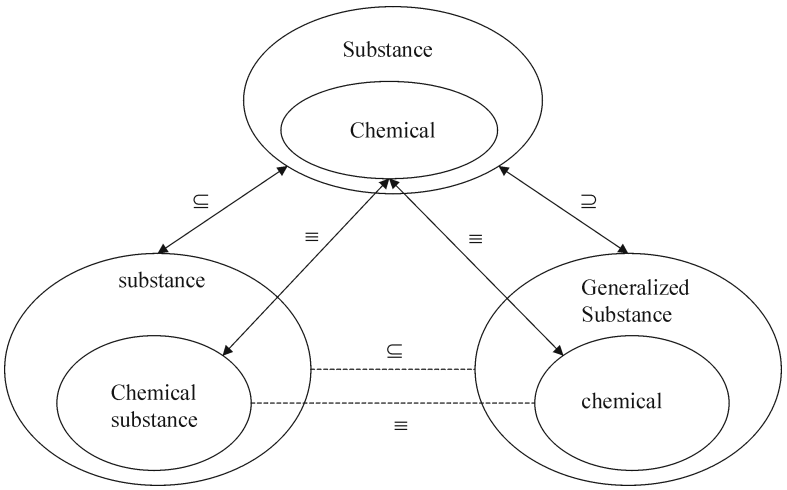


Figure 4: Derivation of semantic relations in the merged model

5 Discussion

We conclude that C-OWL provides a suitable formalism for supporting the alignment of complex terminologies like the ones we face in the medical area. While allowing the co-existence of different views, C-OWL still provides powerful reasoning support for the verification and derivation of mappings and even supports the process of merging terminologies based on existing mappings. These possibilities are essential for support knowledge engineers in the task of specifying mappings which currently mainly is a manual task. C-OWL is designed in such a way that no changes to existing OWL ontologies are required. Alignment mappings can be specified independently just referring to existing ontologies. This makes C-OWL directly applicable to existing ontologies like the ones mentioned in this paper. We are currently developing an RDF-based syntax for mapping definitions in C-OWL. The next steps of the developments of C-OWL is the develop of tools that support the creation, visualization and the reasoning about alignments.

References

[Baker et al., 1999] Baker, P., Goble, C., Bechhofer, S., Paton, N., Stevens, R., and Brass, A. (1999). An ontology for bioinformatics applications. *Bioinformatics*, 15(6):510–520.

[Bodenreider, 2004] Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32.

[Bouquet et al., 2003] Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., and Stuckenschmidt, H. (2003). C-OWL: Contextualizing ontologies. In Sekara, K. and Mylopoulis, J., editors, *Proceedings of the Second International Semantic Web Conference*, number 2870 in Lecture Notes in Computer Science, pages 164–179. Springer Verlag.

[Gangemi et al., 1998] Gangemi, A., Pisanelli, D., and Steve, G. (1998). Ontology integration: Experiences with medical terminologies. In Guarino, N., editor, *Proceedings of Conference on formal ontologies in information systems (FOIS'98)*. IOS Press.

[Golbreich et al., 2003] Golbreich, C., Dameron, O., Gibaud, B., and Burgun, A. (2003). Web ontology language requirements w.r.t expressiveness of taxonomy and axioms in medicine. In *The SemanticWeb - ISWC 2003*, volume 2870 of *Lecture Notes in Computer Science*, pages 180–194. Springer Verlag.

[Nelson and Powell, 2002] Nelson, S. and Powell, Tammy andn Humphreys, B. (2002). The unified medical language system (umls) project. In Kent, A. and Hall, C., editors, *Encyclopedia of Library and Information Science.*, pages 369–378. Marcel Dekker, Inc.

[Rector and Nowlan, 1993] Rector, A. and Nowlan, W. (1993). The galen project. *Computer Methods and Programs in Biomedicine*, 45:75–78.

[Stuckenschmidt, 2004] Stuckenschmidt, H. (2004). Using C-OWL for the alignment of medical ontologies a case study. Technical report, Vrije Universiteit Amsterdam.

Lessons Learned from Aligning two Representations of Anatomy

Songmao Zhang[1], Ph.D., Peter Mork[2,3,4], M.S., Olivier Bodenreider[1], M.D., Ph.D.

[1]U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland
[2]Microsoft Research, Redmond, WA, USA
[3]Computer Science & Engineering, University of Washington, Seattle, WA, USA
[4]Biomedical & Health Informatics, University of Washington, Seattle, WA, USA

{szhang|olivier}@nlm.nih.gov, pmork@cs.washington.edu

The objective of this study is to analyze the comparison, through their results, of two distinct methods applied to aligning two representations of anatomy. The same versions of FMA and GALEN were aligned by each method. 2199 concept matches were obtained by both methods. For matches identified by one method only (337 and 336 respectively), we analyzed the reasons that caused the other method to fail. Alignment 1 could be improved by addressing partial lexical matches and identifying matches based solely on structural similarity. Alignment 2 may be improved by taking into account synonyms in FMA and identifying semantic mismatches. However, both methods identify only a fraction of all possible matches and new approaches need to be explored in order to handle more complex matches.

INTRODUCTION

Anatomy is central to the biomedical domain. While macroscopic anatomy is required for the representation of diseases and procedures, subcellular anatomy has become increasingly important for molecular biology. Not only is a sound representation of anatomy fundamental to biomedicine, but the various representations of anatomy currently available also need to be aligned in order to ensure interoperability. This need inspired two groups of researchers to take up the challenge of aligning two sizeable representations of anatomy: the Foundational Model of Anatomy (FMA) and the GALEN common reference model.

The first effort in aligning these two systems occurred at the US National Library of Medicine (NLM). In parallel, but unrelated to it, another alignment was performed at Microsoft Research. Both approaches use a combination of lexical and structural techniques. In addition, the first approach takes advantage of domain knowledge, while the second approach is domain-independent and thus can be applied to other domains.

The contribution of this study is a comparison and analysis of the results of the two alignments in an effort to determine the strengths and weaknesses of each. This analysis illustrates how each approach can be improved based on the results of the other.

MATERIALS

FMA and GALEN

The Foundational Model of Anatomy[1] (FMA) [July 2, 2002 version] is an evolving ontology that has been under development at the University of Washington since 1994 [1, 2]. Its objective is to conceptualize the physical objects and spaces that constitute the human body. The underlying data model for FMA is a frame-based structure implemented with Protégé-2000 [3]. 58,957 concepts cover the entire range of macroscopic, microscopic and subcellular canonical anatomy. Concept names in FMA are pre-coordinated, and, in addition to preferred terms (one per concept), 28,499 synonyms are provided (up to 6 per concept).

The Generalized Architecture for Languages, Encyclopedias and Nomenclatures in medicine[2] (GALEN) [v. 4] has been developed as a European Union AIM project led by the University of Manchester since 1991 [4, 5]. The GALEN common reference model is a clinical terminology represented using GRAIL [6], a formal language based on description logics. GALEN contains 23,428 concepts and intends to represent the biomedical domain, of which canonical anatomy is only one part. Concept names in GALEN are post-coordinated, and only one name is provided for each non-anonymous concept. There are 2,960 anonymous concepts.

Both FMA and GALEN are modeled by *is-a* relationship. Additionally, FMA uses two kinds of partitive relationships and GALEN 26. The hierarchy of associative relationships is also more extensive in GALEN (514) than in FMA (54). In addition to inter-concept relationships, there are 85 slots in FMA describing atomic properties of concepts, whose types are Boolean, Integer, Symbol, and Instance.

[1] http://fma.biostr.washington.edu/
[2] http://www.opengalen.org/

The UMLS

An additional resource used in the alignment is the Unified Medical Language System® (UMLS®)[3] developed by NLM. The UMLS Metathesaurus® is organized by concept or meaning. A concept is defined as a cluster of terms representing the same meaning (synonyms). The 14th edition (2003AA) of the Metathesaurus contains over 1.75 million unique English terms drawn from more than sixty families of medical vocabularies, and organized in some 875,000 concepts. In the Metathesaurus, each concept is categorized by at least one semantic type from the UMLS Semantic Network. A subset of these semantic types is used to define the domain of anatomy. Also part of the UMLS distribution is the SPECIALIST Lexicon, a large syntactic lexicon of both general and medical English.

METHODS

Alignment 1

Alignment 1 first compares the concepts between FMA and GALEN in two steps: lexical alignment and structural alignment [7]. Then, based on the matching concepts identified, Alignment 1 compares the associative relationships across systems [8].

The **lexical alignment** identifies shared concepts across systems lexically through exact match and after normalization. Concepts exhibiting similarity at the lexical level across systems are called anchors, as they are going to be used as reference concepts in the structural alignment and associative relationship comparison. Additional anchors are identified through UMLS synonymy. Two concepts across systems are considered anchors if their names are synonymous in the UMLS Metathesaurus (i.e., if they name the same concept) and if the corresponding concept is in the anatomy domain (i.e., has a semantic type related to *Anatomy*). For FMA, both preferred concept names and synonyms were used in the lexical alignment process. For GALEN, only non-anonymous concept names were used.

The **structural alignment** first consists of acquiring the semantic relations explicitly represented within systems. Inter-concept relationships are generally represented by semantic relations <$concept_1$, *relationship*, $concept_2$>, where *relationship* links $concept_1$ to $concept_2$. For the purpose of aligning the two ontologies, we considered as only one *part-of* relationship the various subtypes of partitive relationships present in FMA (e.g., *part of, general part of*) and in GALEN (e.g., *isStructuralComponentOf*, *IsDivisionOf*). Only hierarchical relationships were considered at this step, i.e., *is-a*, *inverse-is-a*, *part-of*, and *has-part*. Implicit semantic relations are then extracted from concept names and various combinations of hierarchical relations. Augmentation and inference are the two main techniques used to acquire implicit knowledge from FMA and GALEN.

Augmentation attempts to represent with relations knowledge that is otherwise embedded in the concept names. Augmentation based on reified *part-of* relationships consists of creating a relation <*P*, *part-of*, *W*> between concepts *P* (the part) and *W* (the whole) from a relation <*P*, *is-a*, *Part of W*>, where the concept *Part of W* reifies, i.e., embeds in its name, the *part-of* relationships to *W*. For example, <*Neck of Femur*, *part-of*, *Joint*> was added from the relation <*Neck of Femur*, *is-a*, *Component of Joint*>, where the concept *Component of Joint* reifies a specialized *part-of* relationship. Examples of augmentation based on other linguistic phenomena include <*Prostate gland*, *is-a*, *Gland*> (from the concept name *Prostate gland*) and <*Extensor Muscle of Leg*, *part-of*, *Leg*> (from the concept name *Extensor Muscle of Leg*).

Inference generates additional semantic relations by applying inference rules to the existing relations. These inference rules, specific to this alignment, represent limited reasoning along the *part-of* hierarchy, generating a partitive relation between a specialized part and the whole or between a part and a more generic whole. For example, <*First tarsometatarsal joint*, *part-of*, *Foot*> was inferred based on the relations <*First tarsometatarsal joint*, *is-a*, *Joint of foot*> and <*Joint of foot*, *part-of*, *Foot*>.

With these explicit and implicit semantic relations, the structural alignment identifies structural similarity and conflicts among anchors across systems. Structural similarity, used as **positive structural evidence**, is defined by the presence of common hierarchical relations among anchors across systems, e.g., <c_1, *part-of*, c_2> in one system and <c_1', *part-of*, c_2'> in another where {c_1, c_1'} and {c_2, c_2'} are anchors across systems. Conflicts, on the other hand, are used as **negative structural evidence**. The first type of conflict is defined by the existence of opposite hierarchical relationships between the same anchors across systems, e.g., <c_1, *part-of*, c_2> in one system and <c_1', *has-part*, c_2'> in another. The second type of conflict is based on the disjointedness of top-level categories across systems. For example, *Nail* in FMA is a kind of *Skin appendage* which is an *Anatomical structure*, while *Nail* in GALEN is a *Surgical Fixation Device* which is an *Inert Solid Structure*. *Anatomical structure* and *Inert Solid Structure* being disjoint top-level categories, the two concepts of *Nail*

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

across systems are semantically distinct, although they share the exact same name.

Based on the anchors (except those receiving negative structural evidence), **associative relationships** are compared across systems. The most frequent matches indicate a correspondence between an associative relationship in one system and one relationship (hierarchical or associative) or combination thereof in the other. For example, from *Heart -contained in→ Middle mediastinum -part-of→ Mediastinum* in FMA and *Heart -boundsSpace→ Mediastinum* in GALEN, the relationship match *{FMA: contained in - part-of, GALEN: boundsSpace}* can be extracted.

Alignment 2

The second alignment also includes a lexical phase and a structural phase, followed by a hierarchical match phase [9]. For each phase, generic schema matching algorithms were adapted to 1) cope with the number of concepts present and 2) handle the more expressive modeling environments (Protégé-2000 and GRAIL). Summarizing from [9], the second alignment proceeds as follows.

The **lexical phase** identifies concepts whose names are similar. Each concept name from FMA and GALEN is first mapped to the UMLS Metathesaurus after normalization and reduced to a set of UMLS concept identifiers. Each concept identifier is further annotated with part-of-speech information identified using the SPECIALIST Lexicon. The similarity between two concepts from FMA and GALEN depends on the ratio of shared UMLS concepts to the total number of UMLS concept mapped to. Part-of-speech information is further used to distinguish between roots (nouns and verbs) and modifiers (adjectives and adverbs) [10].

For example, *Valve In Heart* from GALEN is first normalized to *heart valve* and mapped to two UMLS concepts. *Cardiac Valve* from FMA is normalized to *cardiac valve* and mapped to three UMLS concepts, two of which being shared with the mappings of *Valve In Heart*. Based on this, the similarity between *Valve In Heart* and *Cardiac Valve* was assigned a score of .8 (where 0 indicates no similarity and 1.0 indicates a perfect match).

The **structural phase** attempts to identify concepts (and relationships) that are used similarly in both systems. The first step is to reify every relation present in FMA or GALEN, thereby creating new, artificial concepts. For example, one such concept is created from the relation <*Cardiac Valve*, *part-of*, *Heart*>. Similarity scores can then be assigned to matches among these artificial concepts, corresponding to relation matches. The similarity of two relations in a match is estimated to be the average similarity of the concepts and relationships involved in the relations. This process makes it possible to identify the similarity of relations, not only concepts. For example, this is how we identified that both FMA and GALEN assert that cardiac valves are part of the heart.

Moreover, the similarity between relations can be *back-propagated* to improve the similarity of the corresponding concepts and relationships. Whenever two concepts (or relationships) are mentioned in similar relations, the similarity between those concepts is increased. This back-propagation detects similarity of use, especially between relationships. For example, the similarity between *isBranchOf* and *branch of* increases from .28 to .98 using back-propagation.

The final **hierarchical phase** attempts to identify concepts with similar descendants. Similarity scores across leaf concepts were established during the previous phases, but few higher-level correspondences were identified. In this final phase, the similarity between two concepts is increased if there are many descendants that match. In theory, similarity is pushed up the inheritance hierarchy from the leaves, but [11] notes that few matches were found in this manner.

Comparing Alignment 1 and 2

Alignment 1 identified a set of concept matches across systems with an indication of the presence of structural evidence and relationship matches with their frequency. A concept match is supported by Alignment 1 if it receives positive structural evidence; not supported otherwise.

Alignment 2 identified a set of matches for both concepts and relationships, each match being qualified by similarity score. A match is supported by Alignment 2 if its similarity score is higher than or equal to a prespecified threshold; not supported otherwise. The threshold selected in this study is .83, determined heuristically by examining the validity of a subset of matches.

We compared the concept matches obtained by Alignment 1 and 2 by classifying them into four categories: 1) matches supported by both alignments, 2) matches supported by Alignment 1 but not supported or identified by Alignment 2, 3) matches supported by Alignment 2 but not supported or identified by Alignment 1, and 4) matches ignored by both alignments. We then used a similar approach to compare the relationship matches obtained by the two alignments.

RESULTS

The matches obtained in Alignment 1 and 2 are first presented separately. Then, we analyze the results of their comparison. These results are summarized in Table 1 (concept matches).

			Alignment 2		
			Identified		Not identified
			Similarity ≥ .83	Similarity < .83	
Alignment 1	Identified	Positive evidence	2,199	42	295
		No evidence	168	*3*	*29*
		Negative evidence	36	*0*	*4*
	Not identified		132	*1,074*	

Table 1 – Concept matches in Alignment 1 and 2

Matches in Alignment 1

2,410 pairs of matching concepts across systems were identified by lexical alignment between FMA and GALEN. Through UMLS synonyms, 366 additional pairs of matching concepts were found across systems, resulting in totally **2,776 concept matches** in Alignment 1.

By structural alignment, 2,536 (91.4%) of the 2,776 matches received positive evidence, 40 (1.4%) negative evidence and 200 (7.2%) no evidence. *Cardiac valve* (synonym: *Valve of heart*) in FMA and *Valve In Heart* in GALEN exemplify a match with positive evidence as they share hierarchical links to some of the other anchors across systems, e.g., *Heart* (*part-of*), *Tricuspid valve* (*inverse-is-a*) and *Mitral valve* (*inverse-is-a*). *Pectoral girdle* (synonym: *Shoulder girdle*) in FMA and *Shoulder Girdle* in GALEN, although matching lexically, were identified to be a mismatch from the conflicting relationships these concepts have across systems, i.e., <*Pectoral girdle, has-part, Shoulder*> in FMA and <*Shoulder Girdle, part-of, Shoulder*> in GALEN. Finally, although linked to anchors including *Cardiovascular System* (*part-of*) and *Body Part* (*is-a*) in GALEN, *Carotid Body* does not have any hierarchical links to other anchors in FMA, and therefore receives no structural evidence.

The alignment of associative relationships resulted in **182 relationship matches**. Matches with high frequency include *{FMA: branch of, GALEN: isBranchOf}* and *{FMA: tributary of, GALEN: isBranchOf}*.

In summary, a total of 2,958 matches (2,776 for concepts and 182 for relationships) were identified between FMA and GALEN by Alignment 1.

Matches in Alignment 2

A total of 3,780 matches were identified by Alignment 2, 3,503 of them in the lexical phase, 64 in the structural phase, and 213 in the hierarchical phase. 2,583 (68.3%) of the 3,780 matches were assigned similarity scores above the threshold of .83. As a matter of fact, 2,539 of these matches have the similarity score of 1.0 (e.g., {*FMA: Pancreas, GALEN: Pancreas}*). 1,197 (31.7%) of the 3,780 matches have a similarity score lower than .83 and were ignored (e.g., *{FMA: Upper lobe of lung, GALEN: Lobe of Left Lung}* has a similarity of .5).

Among the 3,780 matches, there are **3,654 concept matches** and **22 relationship matches** (e.g., *{FMA: part-of, GALEN: IsDivisionOf}* has a similarity of 1.0). The remaining 104 matches associate things other than two concepts or two relationships. In 102 cases, a concept in one system matches a relationship in the other (e.g., *{FMA: insertion, GALEN: Insertion Point*}). Finally, two FMA Boolean-typed slots match GALEN relationships (e.g., *has dimension* in FMA and *hasDimension* in GALEN).

Concept matches supported by both alignments

2,776 concept matches were identified by Alignment 1 and 3,654 by Alignment 2. Among them, 2,199 both received positive structural evidence and had a similarity score above the threshold of .83, as shown in the upper left part of Table 1. These matches are supported by both alignments. For example, the match *{FMA: Cardiac valve, GALEN: Valve In Heart}*, presented earlier, received positive evidence in Alignment 1, and its similarity score is .88 in Alignment 2.

Concept matches supported by Alignment 1 only

As shown in the upper right part of Table 1, 42 concept matches received similarity scores lower than the threshold by Alignment 2, and 295 were not identified by Alignment 2. However, these 337 matches were supported by positive structural evidence of Alignment 1.

- 167 are FMA synonyms matching GALEN concept names in Alignment 1. Alignment 2 failed to identify or to select these matches in the lexical phase because it did not use synonyms in FMA. For example, *Prostate* in FMA was matched to *Prostate Gland* in GALEN by Alignment 1 because the former has a synonym *Prostate gland* in FMA. The positive structural evidence for this match includes their sharing *is-a* link to *Gland*

and *has-part* link to *Lobe of prostate* across systems.

- 158 were obtained through UMLS synonyms in Alignment 1. One such match is *{FMA: First Tarsometatarsal joint, GALEN: First Tarso Metatarsal Joint}*. This match received positive structural evidence from the shared hierarchical links to other anchors such as *Foot* (*part-of*) and *Joint of foot*[4] (*is-a*) across systems. It was not obtained by Alignment 2 because the two alignments used slightly different matching criteria for mapping to UMLS concepts.
- 12 are FMA preferred concept names matching GALEN concept names in Alignment 1, e.g., *{FMA: Immunoglobulin M, GALEN: Immunoglobulin M}*, which shared hierarchical links to anchors such as *Immunoglobulin* (*is-a*) and *Protein* (*is-a*) across systems. The reasons why these matches were not obtained by Alignment 2 were investigated and found to be essentially unimportant.

Concept matches supported by Alignment 2 only

The lower left part of Table 1 shows the concept matches with similarity scores above the threshold by Alignment 2 but not supported or identified by Alignment 1.

- 168 received no structural evidence by Alignment 1, e.g., {*FMA: Carotid body, GALEN: Carotid Body}*, presented earlier. Although its similarity score is 1.0 by Alignment 2, this match was not supported by Alignment 1 because no structural evidence could be found (in this case, because of a lack of relations being represented in FMA for this concept).
- 36 received negative structural evidence by Alignment 1. Both *{FMA: Nail, GALEN: Nail}* and *{FMA: Pectoral girdle, GALEN: Shoulder Girdle}*, with negative evidence in Alignment 1 as presented earlier, received the similarity score of 1.0 by Alignment 2. These 36 matches were inappropriately supported by Alignment 2 because, unlike Alignment 1, this method does not attempt to identify semantic mismatches.
- 132 were only identified by Alignment 2.
  - 78 could have been obtained by Alignment 1 through UMLS synonymy. They were filtered out by Alignment 1 because they caused two different concepts in one system to be synonymous. In the UMLS Metathesaurus, the terms *Prostate*, *Prostate gland* and *Prostatic gland* are synonymous. In FMA, *Prostate* refers to the organ while *Prostatic gland* is subdivision of the organ. Being different concepts in FMA, their matching to the same UMLS synonym was rejected. Therefore, Alignment 1 did not get the match *{FMA: Prostatic gland, GALEN: Prostate Gland}* while Alignment 2 did.
  - 18 were rejected by Alignment 1 through the Semantic Network filter for Anatomy, e.g., *{FMA: Flatulence, GALEN: Flatus} (similarity = 1.0)*. Neither *Flatulence* nor *Flatus* is related to Anatomy in UMLS and this match was rejected by Alignment 1 for this reason.
  - 36 were not identified by Alignment 1 because at least one of the concept names did not match any UMLS synonyms. For example, Alignment 1 missed *{FMA: Colic flexure, GALEN: Colonic Flexure} (similarity = 1.0)* through UMLS because *Colonic Flexure* in GALEN does not match any UMLS synonyms. Some of these matches of Alignment 2 were determined to be valid by a domain expert.

Concept matches ignored by both alignments

The lower right part of Table 1 shows the concept matches ignored by both alignments. These matches are either not identified by one alignment and not supported by the other or identified but not supported by either alignment.

- 1,074 were only identified by Alignment 2 but their similarity scores are lower than the threshold. 72 are FMA concepts matching GALEN anonymous concepts, purposely ignored by Alignment 1. 1,002 are FMA concepts matching GALEN non-anonymous concepts. Most of these matches correspond to partial matches, not addressed by Alignment 1 (e.g., *{FMA: Ligament of knee joint, GALEN: Ligament of Knee}*, with a similarity score of .35).
- 32 received no structural evidence by Alignment 1, while 3 of them had similarity scores lower than the threshold and 29 were not identified by Alignment 2.
- 4 received negative structural evidence by Alignment 1 and were not identified by Alignment 2.

Relationship matches

182 relationship matches were identified in Alignment 1. Alignment 2 identified 22 matches, of which 17 were supported by a similarity score above .83. Seven relationship matches were identified by both alignments (e.g., *{FMA: nerve supply, GALEN: isServedBy}*). Seven were supported by Alignment 2 only (e.g., *{FMA: lymphatic drainage, GALEN: is-*

[4] The anchor is named *Foot Joint* in GALEN.

ServedBy}). Alignment 1, relying on the concepts already aligned, failed to identify these matches, because these relationships occurred among concepts that have not been aligned. Finally, in three cases, the match identified by Alignment 2 corresponded to a match created manually in Alignment 1 between the subtypes of *part-of* relationships (e.g., *{FMA: part-of, GALEN: IsDivisionOf}*).

DISCUSSION

Improving the alignments

In fact, the philosophy behind each method is different. Alignment 1 takes advantage of domain knowledge. It requires lexical matches to be supported by structural matches, at the cost of inaccurately rejecting some valid matches. Therefore, it favors precision over recall. On the other hand, Alignment 2 relies on generic algorithms and, by imposing no penalty for lack of structural matches, favors recall over precision. Theoretically, the two approaches could be combined. In practice, however, despite their differences, their results are surprisingly close and any improvement would only be marginal at best.

Nevertheless, each approach can be improved based on the results of the other. Alignment 1 would benefit from addressing partial lexical alignment and identifying matches based solely on structural similarity. Alignment 2 could be improved by taking into account synonyms in FMA and identifying semantic mismatches.

Of particular interest are the 875 relation matches obtained by Alignment 2 in the structural phase for the purpose of increasing the similarity scores of the corresponding concepts and relationships. In addition to increasing the chances of identifying concept matches, these relation matches could be used for themselves. For example, the match by *{FMA: <Lung, contained in, Thoracic cavity>, GALEN: <Lung, isSpecificallyNonPartitivelyContainedIn, Pleural Membrane>}* whose similarity score is .33, captured the difference the two ontologies have in representing the knowledge about equivalent concepts.

Validating the alignments

The validation of the results of the alignment has been an issue for both groups. Anatomy is a vast domain and, in addition to domain knowledge, the experts are also required to have some knowledge of the two systems under investigation. No group has achieved a comprehensive evaluation of its results. One interest of disposing of two alignments is that there is the possibility of a cross-validation. In fact, while the matches of Alignment 1 can certainly validate those of Alignment 2, the contrary is not necessarily true. In Alignment 1, a lexical match is required to be supported by some structural evidence. Conversely, in Alignment 2, lexical matches get the highest score possible and structural evidence, if any, is only used to increase the score of partial lexical matches. However, matches from Alignment 2 supported by structural evidence could be used to validate the results of Alignment 1. Unfortunately, the similarity score used in Alignment 2 to indicate the quality of the match does not strictly reflect the presence of structural evidence.

Challenges

Neither alignment identified enough matches. A total of 3,982 concept matches were identified by the two alignments together, only accounting for about 7% of FMA concepts and 17% of GALEN concepts. All concept matches identified by the two alignments are one-to-one matches. However, there are more complex cases where a single concept in one ontology may match a group of concepts in the other [11]. Groups of concepts may also match across ontologies. For example, along the *is-a* hierarchy of FMA, *Lobe of lung* is first modeled by upper/lower positions (i.e., *Upper lobe of lung* and *Lower lobe of lung*). These concepts are further subdivided by laterality (including *Upper lobe of left lung* and *Upper lobe of right lung*). On the other hand in GALEN, *Lobe of Lung* is first modeled by laterality (i.e., *Lobe of Left Lung* and *Lobe of Right Lung*) and further subdivided by upper/lower positions (i.e., *Upper Lobe of Left Lung* and *Lower Lobe of Left Lung*). These modeling differences revealed that *Lobe of Left Lung* in GALEN, rather than to one single concept in FMA, should be matched to two concepts: *Upper Lobe of left lung* and *Lower lobe of left lung*. New alignment techniques need to be explored to handle such complex cases.

Acknowledgements

This work was done while Songmao Zhang was a visiting scholar at the Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Department of Health and Human Services. Support for Peter Mork's work was provided in part by NLM training grant T15LM07442.

Thanks for their support and encouragement to Cornelius Rosse for FMA, Alan Rector for GALEN, and their collaborators. Thanks also to Phil Bernstein for his constructive comments and to Microsoft Research for making available to us their results (of Alignment 2).

References

1. Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform* 2003;36(6):478-500.
2. Noy NF, Musen MA, Mejino JLV, Rosse C. Pushing the envelope: challenges in a frame-based representation of human anatomy. *Data & Knowledge Engineering* 2004;48(3):335-359.
3. Musen M, Crubézy M, Fergerson R, Noy NF, Tu S, Vendetti J. Protégé-2000. In. Stanford, CA: Stanford Medical Informatics.
4. Rector AL, Bechhofer S, Goble CA, Horrocks I, Nowlan WA, Solomon WD. The GRAIL concept modelling language for medical terminology. *Artif Intell Med* 1997;9(2):139-71.
5. Rogers J, Rector A. GALEN's model of parts and wholes: experience and comparisons. *Proc AMIA Symp* 2000:714-8.
6. Zanstra PE, van der Haring EJ, Flier F, Rogers JE, Solomon WD. Using the GRAIL language for Classification Management. In: Fifteenth International Congress of the European Federation for Medical Informatics; 1997; Thessaloniki, Greece; 1997. p. 897–901.
7. Zhang S, Bodenreider O. Aligning representations of anatomy using lexical and structural methods. *Proc AMIA Symp* 2003:753-757.
8. Zhang S, Bodenreider O. Comparing Associative Relationships among Equivalent Concepts across Ontologies. *Medinfo* 2004:(to appear).
9. Mork P, Bernstein PA. Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. In: 20th International Conference on Data Engineering; 2004 March 30–April 2; Boston, MA: IEEE; 2004.
10. Madhavan J, Bernstein PA, Rahm E. Generic Schema Matching Using Cupid. In: Apers PMG, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass RT, editors. Proceedings of 27th International Conference on Very Large Data Bases; 2001 Sept 11–14; Roma, Italy: Morgan Kaufmann; 2001. p. 49–58.
11. Mork P, Pottinger RA, Bernstein PA. Challenges in Precisely Aligning Models of Human Anatomy Using Generic Schema Matching. In: Proceedings of MedInfo; 2004 September 7–11; San Francsico, CA: IMIA; 2004.