Representing the UMLS® Semantic Network Using OWL
(Or "What's in a Semantic Web Link?")

Vipul Kashyap[1] and Alex Borgida[2]

[1] LHCNBC, National Library of Medicine,
8600 Rockville Pike, Bethesda, MD 20894
[2] Department of Computer Science, Rutgers University,
New Brunswick, NJ 08903

Abstract. The Semantic Network, a component of the Unified Medical Language System® (UMLS), describes core biomedical knowledge consisting of semantic types and relationships. It is a well established, semi-formal ontology in widespread use for over a decade. We expected to "publish" this ontology on the Semantic Web, using OWL, with relatively little effort. However, we ran into a number of problems concerning alternative interpretations of the SN notation and the inability to express some of the interpretations in OWL. We detail these problems, as a cautionary tale to others planning to publish pre-existing ontologies on the Semantic Web, as a list of issues to consider when describing formally concepts in any ontology, and as a collection of criteria for evaluating alternative representations, which could form part of a methodology of ontology development.

1 Introduction

The Unified Medical Language System® (UMLS®) project was initiated in 1986 by the U.S. National Library of Medicine (NLM). Its goal is to help health professionals and researchers use biomedical information from *a variety of different sources* [1]. The UMLS consists of (i) biomedical concepts and associated strings, comprising the Metathesaurus (*MT*), (ii) a Semantic Network (*SN*) [2], and (iii) a collection of lexical tools (including SPECIALIST lexicon). Both data and tools are integrated in the UMLS Knowledge Source Server[1] and used in a large variety of applications (e.g. PubMed[2], ClinicalTrials.gov[3]). The *MT* provides a common structure for integrating more than 95 source biomedical vocabularies, organized by "concept" (cluster of terms representing the same meaning). The *SN* is a structured description of core biomedical knowledge consisting of semantic types and relationships, used to categorize *MT* concepts, with the *SN* being viewed by some as a semi-formal ontology. It (along with the *MT*) has been in use for more than a decade in the context of information retrieval applications. We expected to "publish" the *SN* on the Semantic Web by expressing it in OWL with relative ease, since there have been lots

[1] http://umlsks.nlm.nih.gov
[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[3] http://www.clinicaltrials.gov

D. Fensel et al. (Eds.): ISWC 2003, LNCS 2870, pp. 1–16, 2003.

of papers that discuss the representation of medical terminologies using OWL style notations, called description logics (*DLs*) (e.g., [11][17][26][27]). Besides, there are numerous papers on the UMLS, including ones specifically about the semantics of the *SN* [2].

We ran into a number of difficulties in this undertaking. Some obstacles were due to ambiguities in the semantics of the *SN* notation or the under-specification of the notation (e.g., what can be inferred from the absence of edges?). Other problems were due to the inability to express the *SN* as OWL axioms which would provide the desired inferences, and the difficulty of making choices between multiple possible representations. We detail these problems: (i) as a cautionary tale to others wanting to publish ontologies on the semantic web, (ii) as a list of issues/alternatives to be considered in the process, and (iii) explore criteria for choosing among the above alternatives. We discuss next the motivation for expressing the *SN* and *MT* using OWL.

Motivation: Formal Representations of Biomedical Knowledge

Biomedical vocabularies and ontologies have always played a critical role in the context of healthcare information. For example, clinical and hospital information systems have used terms from a variety of biomedical vocabularies to specify codes for healthcare transactions and other pieces of information. eGov initiatives such as consolidated health informatics[4] (CHI) and government regulations such as HIPAA[5] have standardized on biomedical vocabularies included in the UMLS, for example, SNOMED, ICD-9-CM. Vocabularies such as the Medical Subject heading (MeSH), a component of the UMLS, have also been developed to help better specify queries for full text retrieval and for annotation of research articles in PubMed. Therefore the main motivations for a *formal* representation of biomedical knowledge are: (a) creation and maintenance of *consistent* biomedical terminology; (b) enabling translations of concepts across multiple autonomous vocabularies; and (c) improved specification of queries for information retrieval. An instance of the latter is the annotation of MEDLINE documents using descriptors built with concepts from the MeSH vocabulary [20]. For example, the semantics of the keyword "mumps" can be qualified by the subheading "complication", which can be conjoined with the main heading "pancreatitis" qualified by "etiology", to produce the MeSH descriptor `(Mumps/CO AND Pancreatitis/ET)`. This semi-formal descriptor can be used to improve text retrieval by use as a label or as part of a query. It can also be expressed more precisely using a Description Logic concept like $\exists$`complication.Mumps` $\sqcap$ $\exists$`etiology.Pancreatitis`, thus allowing for inferences during query answering. The above applications require functionality enabled by the use of OWL and its associated DL reasoner:

- Recognizing inconsistent (empty) concepts/relationships, and faulty subclass/ sub-property relationships (for creation and consistency maintenance).
- Recognizing concept equivalence (for creation/merging of terminologies, and matching of search queries and document annotations).
- Determining the position of a concept expression in a given hierarchy (to enable vocabulary merging into a directed acyclic graph (DAG) structure).

[4] http://www.jrfii.com/chi/

[5] http://www.cms.hhs.gov/hipaa/

- Determining the closest parents and children of a concept in the DAG (for concept translations e.g., [6]).
- Subsumption checking to tighten estimates of semantic distance between concepts, and to limit navigation of the DAG (for concept translations [6], and determination of relevant articles and result ranking for IR.)

We discuss next the "Vanilla" *SN* and its naïve OWL representation. Section 3 then presents various possible interpretations of "links" in the *SN* (prompting the sub-title of the paper as a twist on the famous paper by W. Woods [12]), and the resulting multiple representations. Section 4 discusses possible criteria that might be used to choose between these multiple representations. Conclusions and ongoing/future work are presented in Section 5.

2 The ("Vanilla") Semantic Network and OWL

We start by relating the simple, uncontroversial aspects of the *SN* to the OWL ontology language. The *SN* is a typical semantic network (see Figure 1) with nodes (called "semantic types") and links ("semantic relationships"). The types are organized into two high level hyponym/is-a hierarchies rooted at **`Entity`** and **`Event`**. Intuitively, but not formally, types are organized either by their inherent properties (e.g., a `Mammal` is a `Vertebrate` with constant body temperature) or by certain attributed characteristics (e.g., `ProfessionalGroup` is a set of individuals classified by their vocation). As illustrated in Figure 1, a `MentalBehavioralDysfunction` is a `DiseaseOrSyndrome`, which in turn is a `PathologicFunction`. The relationships used in the SN are also organized in a (shallow) is-a hierarchy, with 5 roots: (a) **`physically_related_to:`** e.g., `part_of`; (b) **`spatially_related_to:`** e.g., `surrounds`, (c) **`temporally_related_to:`** e.g., `precedes`, (d) **`functionally_related_to:`** e.g., `performs`, (e) **`conceptually_related_to:`** e.g., `measures`, `property_of`. For example, the relationship root **`functionally_related_to`** has several is-a children, including `affects` which in turn has many children, including `manages` and `treats`. As is, the relationships in the semantic network are binary.

In order to represent the above on the Semantic Web, RDF Schema (RDFS) would seem to be sufficient. However, RDFS cannot deal with some more advanced aspects of the *SN* to be presented below, and is not equipped to provide the kinds of inferences we had asked for earlier, thus leading us to consider OWL [22]. OWL, based on DAML+OIL [4], is intended to describe the terminology of a domain in terms of *classes/concepts* describing sets of individuals, and *properties/roles* relating these. An ontology consists of a set of *axioms* that assert characteristics of these classes and properties. OWL DL is a kind of *DL* – a logic with clear, formal semantics, (usually corresponding to a subset of First Order Predicate Calculus,) with desirable computational properties (e.g. decidable decision procedures). As in all *DL*, classes can be names (URIs) or *composite expressions*, and a variety of *constructors* are provided for building class expressions. The expressive power of the language is determined by the class (and property) constructors provided, and by the kinds of axioms allowed. Table 1 summarizes these for the underlying OWL. The connection between the *DL* notation and OWL's RDF syntax is shown by the translation of the disjunctive *DL* concept `Bacterium` ∪ `Virus`:

```
<owl:Class>
     <owl:unionOf rdf:parseType="Collection">
         <owl:Class rdf:about="#Bacterium"/>
         <owl:Class rdf:about="#Virus"/>
     </owl:unionOf>
</owl:Class>
```

In a *DL* representation of the UMLS Semantic Network, it is natural to associate *SN* semantic types with *DL* primitive concepts. So, the node `Organism` corresponds to *DL* concept `Organism`, which would be represented in OWL as the class `<owl:Class rdf:ID="Organism"/>`.

An *SN* relationship, such as `process_of`, corresponds to a *DL* primitive role, `process_of`, which would be translated to OWL object properties. In the simplest case, one could associate with a relationship the source and destination of the edge as the "domain" and "range" specification:

```
<owl:ObjectProperty rdf:ID="process_of">
                <rdfs:domain rdf:resource="#BiologicFunction">
                <rdfs:range rdf:resource="#Organism">
</owl:ObjectProperty>
```

However, as we shall see in the next section, this translation could be controversial.

Axioms originate from inheritance hierarchies of the various types and relationships. Thus the type/relationship hierarchy in the *SN* can be represented as a collection of subclass/subproperty axioms such as:

`Fungus` $\subseteq$ `Organism` *(sub-types using <owl:subClassOf>)*
`Virus` $\subseteq$ `Organism`
`part_of` $\subseteq$ `physically_related_to` *(sub-relationships using <owl:subPropertyOf>)*
`contains` $\subseteq$ `physically_related_to`

Some relationships in the *SN* have inverses, which is specified through axioms involving the inverseOf role constructor

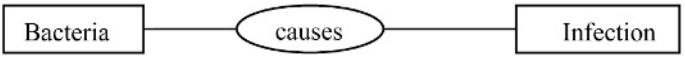
`part_of` $\equiv$ `has_part`$^-$ *(asymmetric property)*
`adjacent_to` $\equiv$ `adjacent_to`$^-$ *(symmetric property)*

3 Semantics of a "Link" in the UMLS Semantic Network

SN types, relationships and their hierarchies, as well as inverses have clear corresponding OWL/DL constructs. However, there are serious difficulties in accurately capturing the semantics of the *SN*, due both to the under-specified meaning of the notion of "link" between two semantic types, and the somewhat unusual inferences/constraints that are associated with them in *SN*. We explore next various possible interpretations of "link", proposing OWL axioms for each, identifying when necessary new *DL* constructs needed. We then evaluate each in light of additional special "inferences" required in the *SN*, such as the notions "domain and range inheritance", "inheritance blocking" and "polymorphic relationships".

3.1 Multiple Interpretations of a "Link"

Consider the following simple diagram, with link labeled "causes" connecting two nodes, Bacteria and Infection:



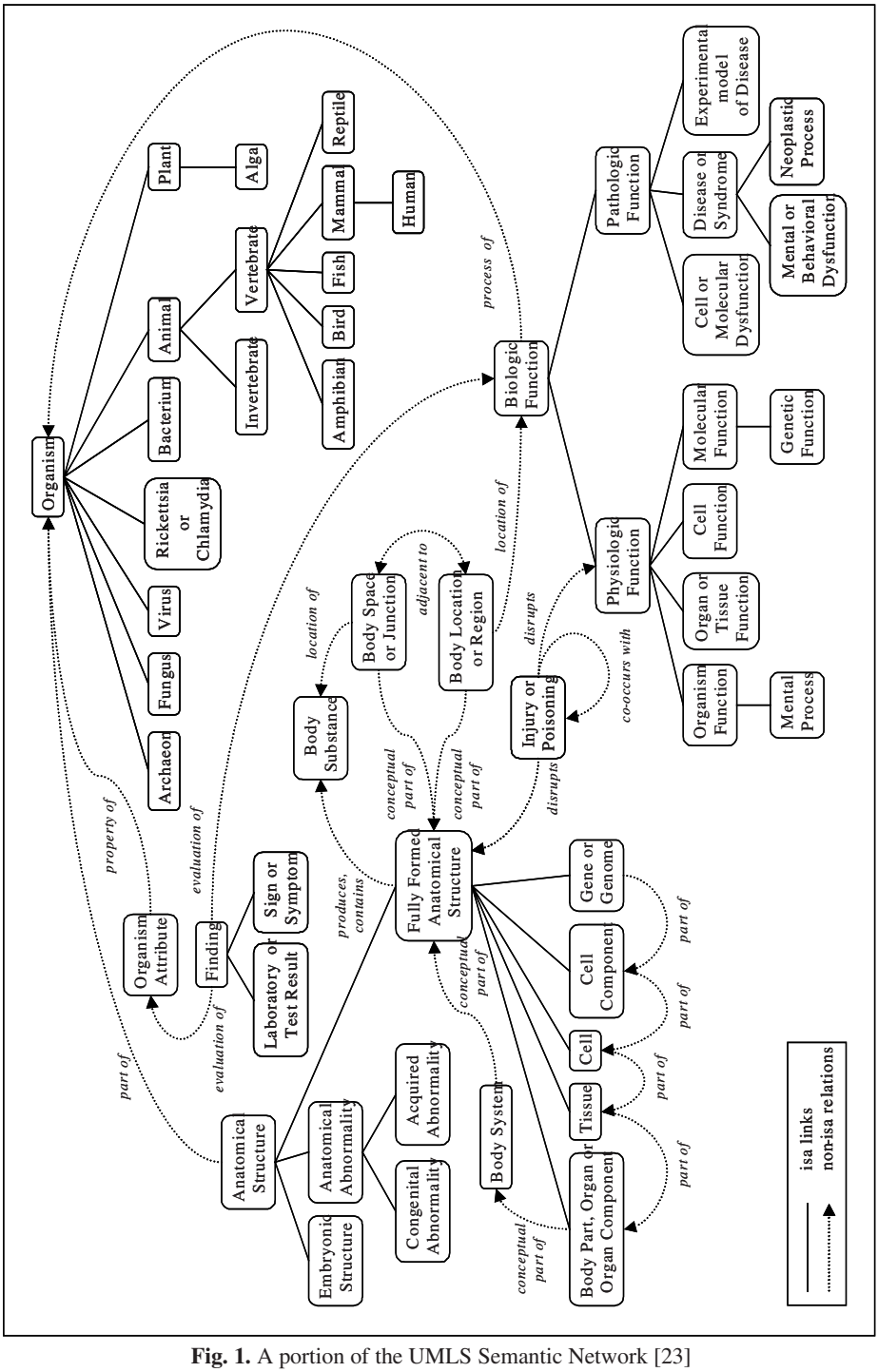


Fig. 1. A portion of the UMLS Semantic Network [23]

Table 1. OWL/RDF constructors and axioms

Constructor/Axiom	DL Syntax	Example
intersectionOf	$C_1 \cap \ldots \cap C_n$	Bacterium ∩ Animal
unionOf	$C_1 \cup \ldots \cup C_n$	Bacterium ∪ Virus
complementOf	$\neg C$	¬Plant
oneOf	$\{ x_1, \ldots, x_n \}$	{aspirin, tylenol}
allValuesFrom	$\forall P.C$	∀partOf.Cell
someValuesFrom	$\exists P.C$	∃processOf.Organism
hasValue	$\exists P.\{x\}$	∃treatedBy{aspirin}
top concept	$\top$	ENTITY
bottom concept	$\bot$	NOTHING
subClassOf	$C_1 \subseteq C_2$	Human ⊆ Animal ∩ Biped
sameClassAs	$C_1 \equiv C_2$	Man ≡ Human ∩ Male
subPropertyOf	$P_1 \subseteq P_2$	part_of ⊆ physically_related_to
samePropertyAs	$P_1 \equiv P_2$	has_temperature ≡ has_fever
disjointWith	$C_1 \subseteq \neg C_2$	Vertebrate ⊆ ¬Invertebrate
sameIndividualAs	$\{x_1\} \equiv \{x_2\}$	{aspirin}≡{acetyl_salicylic_acid}
differentIndFrom	$\{x_1\} \subseteq \neg\{x_2\}$	{aspirin} ⊆ ¬{tylenol}
inverseOf	$P_1 \equiv P_2^{-}$	has_evaluation ≡ evaluation_of^{-}
transitiveProperty	$P^{+} \subseteq P$	part_of^{+} ⊆ part_of
functionalProperty	$\top \subseteq \leq 1\ P$	⊤ ⊆ ≤ 1 has_genetic_profile
inverseFunctionalPropty	$\top \subseteq \leq 1\ P^{-}$	⊤ ⊆ ≤ 1 is_genetic_profile_of^{-}
domain	$\exists P.\top \subseteq C$	∃evaluation_of.⊤ ⊆ Finding
range	$\top \subseteq \forall P.C$	⊤ ⊆ ∀evaluation_of.DiagnosticTest

Rector [9] identifies 5 possible interpretations of the above, corresponding to the English statements:

"**All** bacteria cause {**each/only/some**} infection(s)"

"**Some** bacteria cause {**all/some**} infections".

Since semantic web ontology languages emphasize describing relationships in terms of domains and ranges, let us also consider some statements using these notions. We start with defining two operators δ and ρ as follows:

$$\delta(\mathbf{P}) = \{x \mid (\exists y)P(x, y)\} \text{ and } \rho(\mathbf{P}) = \{y \mid (\exists x)P(x, y)\}$$

These operators define two sets for presentation purposes and under some interpretations they might correspond to the domain/range interpretations associated with RDFS, and DAML+OIL/OWL. These operators suggest three more interpretations.

"The set of Bacteria {**equals / is contained in/contains**} δ(causes)."

"The set of Infections {**equals/ is contained in /contains**} ρ(causes)."

Consider now representing each of the above 8 cases using *DLs*. *DL* descriptions can be used to represent **δ(P)** and **ρ(P)** as follows:

$$\delta(\mathbf{P}) \equiv \exists \text{causes}.\top, \quad \rho(\mathbf{P}) \equiv \exists \text{causes}^{-}.\top$$

we have the following axioms for the last three cases above:

- **"δ/ρ equals":**

  axioms: ∃causes.⊤ ≡ Bacteria, ∃causes^{-}.⊤ ≡ Infection

- **"δ/ρ subsumed":**
  axioms: $\exists causes.\top \subseteq Bacteria$, $\exists causes^{-}.\top \subseteq Infection$
  It may be noted that this corresponds to the domain/range interpretations specified in the RDFS, DAML+OIL/OWL context [24] [25].
- **"δ/ρ subsumes":**
  axioms: $Bacteria \subseteq \exists causes.\top$, $Infection \subseteq \exists causes^{-}.\top$

For the 5 possible statements discussed earlier, we have:

- **"All/some"** (*"All bacteria cause some infection")* :
  axiom: $Bacteria \subseteq \exists causes.Infection$
- **"All/only"** *("All bacteria cause only infections.")*:
  axiom: $Bacteria \subseteq \forall causes.Infection$
- "**All/each**" (*"All bacteria cause each infection.")* This interpretation corresponds to the FOL formula:
  $(\forall x)(Bacteria(x) \wedge (\forall y)(Infection(y) \rightarrow causes(x,y)))$
  $\equiv (\forall x)(Bacteria(x) \wedge (\forall y)(\neg causes(x,y) \rightarrow \neg Infection(y)))$
  This can be represented as a subsumption axiom using the role complement operator in *DLs*:
  axiom: $Bacteria \subseteq \forall \neg causes.\neg Infection$
  or using the special concept constructor $\forall \mathbf{C.r}$ ("*objects related by r to all objects in C*"), which has been investigated by Lutz and Sattler [13]:
  $Bacteria \subseteq \forall Infection.causes$
  In either case, we go beyond the limits of OWL.
- **"Some/some"** *("Some bacteria cause some infections.")* This interpretation can be represented in a number of different ways, though none using axioms of the kinds described in **Table 1**. The alternatives include:
  (i) "There is *some* state of the world where a bacterium causes an infection"
  axiom: $Bacteria \not\subset (\leq 0\ causes\ Infection)$ or
  axiom: "$Bacteria \cap \exists causes.Infection$ **is consistent"**
  (ii) "A bacterium causes an infection in every possible state of the world"
  axiom: "the concept $(Bacteria \cap \exists causes.Infection)$ **is never empty"**
  It is the latter which corresponds to the desired logical formula $(\exists x)(\exists y)(Bacteria(x) \wedge Infection(y) \wedge causes(x,y))$; it can be expressed using a new kind of axiom, concerning the cardinality of *concepts*[6], which was introduced by Baader et al [8]:
  axiom: $\geq 1\ (Bacteria \cap \exists causes.Infection)$
- **"Some/any"** *("Some bacteria cause all infections.")* This requires a combination of the two previous techniques
  axiom: $\geq 1\ (Bacteria \cap \forall \neg causes.\neg Infection)$

A summary of the above interpretations and corresponding encodings may be viewed in Table 2, at the end of this section. We consider next three aspects of the *SN*, some corresponding to inferences and some to constraints, and evaluate the above listed encodings with them in mind.

---

[6] $\geq 1\ C)$ is encoded in OWL by asserting the following axioms: $\top \subseteq \exists P.\{b\}$ and $\{b\} \subseteq \exists P^{-}.C$, where b is a new atomic individual and P is a new role.

3.2 δ and ρ Inheritance

The "is-a" link gives rise to "inheritance" relationships, a hallmark of semantic networks. For example, the type `BiologicFunction` has a relationship `process_of` to the type `Organism` in the semantic network (Figure 1) – to be written henceforth as `process_of(BiologicFunction,Organism)`. By inheritance, the descendants of `BiologicFunction` such as `PhysiologicFunction`, `MentalProcess`, etc. are all understood to have the `process_of` relationship to `Organism`. Surprisingly, *SN* also assumes inheritance on the "range" of the relationship, i.e., `process_of(BiologicFunction,Animal)`, `process_of(PhysiologicFunction, Animal)` also hold. An encoding of *SN* will be said to support δ-inheritance if, given (an encoding of) `P(A,B)`, and a concept `C` such that $C \subseteq A$, (the encoding of) `P(C,B)` is entailed; and ρ-inheritance is supported if for a `D` such that $D \subseteq B$, `P(A,D)` is entailed. Consider now whether/how the encodings discussed in Section 3.1 support δ-inheritance and ρ-inheritance.

- **"δ/ρ equals":** This encoding doesn't support δ-inheritance or ρ-inheritance: from `P(A,B)` we have $\delta(P) \equiv A$, and if `P(C,B)` were to be true, then $\delta(P) \equiv C$, which means that `A` must have been equal `C`.
- **"δ/ρ subsumed":** This encoding also doesn't support δ-inheritance or ρ-inheritance, since $\{C \subseteq A,\ \delta(P) \subseteq A\}$ doesn't entail $\delta(P) \subseteq C$.
- **"δ/ρ subsumes":** Interestingly , this encoding supports both δ-inheritance and ρ-inheritance, because $C \subseteq A \subseteq \delta(P)$ entails $C \subseteq \delta(P)$, and $D \subseteq B \subseteq \rho(P)$ entails $D \subseteq \rho(P)$, so that the representation of `P(C,D)` holds.
- **"All/some":** The encoding of `P(A,B)` as $A \subseteq \exists P.B$ supports δ-inheritance, but not ρ-inheritance since from $C \subseteq A$ and $D \subseteq B$ we get $C \subseteq \exists P.B$ but not $C \subseteq \exists P.D$
- **"All/only":** The encoding $A \subseteq \forall P.B$ behaves like the previous one, supporting δ-inheritance, but not ρ-inheritance.
- **"All/each":** The encoding $A \subseteq \forall \neg P.\neg B$ supports δ-inheritance as in the previous cases. It also supports ρ-inheritance since $D \subseteq B$ entails $\forall \neg P.\neg B \subseteq \forall \neg P.\neg D$, so that $A \subseteq \forall \neg P.\neg D$ holds.
- **"Some/some":** The encoding $(\geq 1\ (A \cap \exists P.B))$ doesn't support δ-inheritance or ρ-inheritance because the addition of $C \subseteq A$ and $D \subseteq B$ doesn't entail either $(\geq 1\ (C \cap \exists P.B))$ or $(\geq 1\ (A \cap \exists P.D))$.
- **"Some/all":** The encoding $(\geq 1\ (A \cap \forall \neg P.\neg B))$ doesn't support δ-inheritance, but supports ρ-inheritance: Since $D \subseteq B$ entails $\forall \neg P.\neg B \subseteq \forall \neg P.\neg D$, we can infer $(A \cap \forall \neg P.\neg B) \subseteq (A \cap \forall \neg P.\neg D)$, and hence $(\geq 1\ (A \cap \forall \neg P.\neg D))$ holds if $(\geq 1\ (A \cap \forall \neg P.\neg B))$ holds.

In general, if an encoding does not support some form of inheritance, we will need to explicitly assert axioms corresponding to the missed inheritance inferences.

3.3 Inheritance Blocking

In some cases there will be a conflict between the placement of types in the *SN* and the links to be inherited. For example, `process_of(MentalProcess,Plant)` is

inherited from process_of(BiologicFunction,Organism) though this is an undesirable inference, since plants are not sentient beings. For this purpose, the *SN* provides a mechanism to explicitly "block" inheritance.[7] In general, whenever a mechanism does *not* support a form of inheritance, it can deal with blocking by simply not adding explicitly the axioms. However, when inheritance is a logical consequence of the axioms, preventing the relationship from holding would normally lead to logical inconsistency. Rather than rely on "default/non-monotonic reasoning", which however is quite complex, we will instead adopt the approach of *modifying* axioms whenever exceptions are encountered. This approach will be made easier in our case by the fact that the *SN* does not support multiple inheritance. Let us look then at ways to block inheritance in those cases where it does occur:

- **"δ/ρ subsumes":** Let $C_1 \subseteq A$ and $D_1 \subseteq B$, and suppose that although $P(A,B)$ holds, we don't want the property P to be inherited to C_1 and D_1. We could specify: $A \cap \neg C_1 \subseteq \delta(P)$ and $B \cap \neg D_1 \subseteq \rho(P)$. However, suppose $C_2 \subseteq A$ and $D_2 \subseteq B$, and we also want to block $P(C_2,D_2)$; then asserting $A \cap \neg(C_1 \cup C_2) \subseteq \delta(P)$ and $B \cap \neg(D_1 \cup D_2) \subseteq \rho(P)$ has the unintended effect of blocking the links $P(C_1,D_2)$ and $P(C_2,D_1)$. To **compensate** for this, one could explicitly add axioms specifying $P(C_1,D_2)$ and $P(C_2,D_1)$.
- **"All/each":** Recall that the encoding $A \subseteq \forall\neg P.\neg B$ supports both δ-inheritance and ρ-inheritance. Suppose we are given $C_1, C_2 \subseteq A$ and $D_1, D_2 \subseteq B$, and we want to block $P(C_1,D_1)$ and $P(C_2,D_2)$. We can start by asserting $A \cap \neg(C_1 \cup C_2) \subseteq \forall\neg P.(\neg(B \cap \neg(D_1 \cup D_2))$. But once again we need to add ***compensating axioms*** $C_2 \subseteq \forall P.\ D_1$ and $C_1 \subseteq \forall P.\ D_2$ to represent the links $P(C_1,D_2)$ and $P(C_2,D_1)$, which could no longer be deduced.
- **"All/only":** To block the δ-inheritance of $P(C_1,B)$ when $P(A,B)$ and $C_1 \subseteq A$, replace the axiom $A \subseteq \forall P.B$ by $A \cap \neg(C_1 \cup \ldots) \subseteq \forall P.B$, so that exceptions are explicitly noted.
- **"All/some"** is similar to "**all/only**".
- **"Some/all":** To block ρ-inheritance of $P(A,D)$ when $P(A,B)$ and $D_1 \subseteq B$, use the axiom $\geq 1\ (A \cap \forall\neg P.(\neg B \cup D_1))$ instead of $\geq 1\ (A \cap \forall\neg P.\neg B)$.
- **"Some/some":** Although this representation does not support either δ-inheritance and ρ-inheritance, and hence has no problem with blocking, it does have an interesting property: $\{(\geq 1\ (A \cap \exists P.B)),\ A' \supseteq A\ ,\ B' \supseteq B\}$ entails $(\geq 1\ (A' \cap \exists P.B))$ and $(\geq 1\ (A \cap \exists P.B'))$, which suggests that if $P(A,B)$ is asserted then $P(A',B')$ can be deduced for all *super-classes* A' of A and B' of B. This undesirable "upwards inheritance" can be blocked by specifying axioms such as $(\leq 0\ (A' \cap \exists P.B))$ and $(\leq 0\ (A \cap \exists P.B'))$.

3.4 Polymorphic Relationships

Polymorphic relationships are ones whose arguments, i.e., domain and range values can be instances of multiple classes. One (benign) source of such polymorphism is

[7] The *SN* also allows blocking to be applied to ***all*** children, without explicitly having to list all possible pairs. So from P(A, B), P(C, D) is blocked for *any* descendants C of A and D of B. We do not examine this feature here for lack of space.

inheritance – it is called "subtype polymorphism" in Programming Languages. However, in UMLS SN, the same relationship can be stated to connect pairs of types that are not is-a related. For example, in Figure 1, we have, among others

```
location_of(BodySpaceorJunction,BodySubstance)
location_of(BodyLocationOrRegion,BiologicFunction)
contained_in(BodySubstance,EmbryonicStructure)
contained_in(BodySubstance,FullyFormedAnatomicalStructure)
```

Such examples, with edges $P(A_1, B_1)$ and $P(A_2, B_2)$, exhibit what might be called "ad-hoc polymorphism/overloading" in Object Oriented languages. One could interpret this to be equivalent to the introduction of two new relationships, P_1 and P_2, adding the axiom $P \equiv P_1 \cup P_2$, and modeling $P(A_1, B_1)$ and $P(A_2, B_2)$ with $P_1(A_1, B_1)$ and $P_2(A_2, B_2)$. Unfortunately, OWL does not support property union; and even if it did, the above encoding is ***non-incremental*** in the sense that we must detect cases of overloading, and *remove* earlier axioms, replacing them with new ones. Such non-locality seems unavoidable for blocking, which is inherently non-monotonic, but is otherwise undesirable since it makes it difficult to maintain sets of axioms.

Some intuitions related to polymorphism which may be useful to evaluate alternatives are: (i) When $A_1 \cap A_2$ and $B_1 \cap B_2$ are empty, as in the first pair above, the constraints should not affect each other. (ii) When $A_1 \equiv A_2 \equiv A$, as in the second example pair above, the encoding should not require R to be the empty relation if $B1 \cap B2 \equiv \emptyset$. Consider again the encodings from Section 3.1.

- **"δ/ρ subsumed":** From $\{\delta(P) \subseteq A_1,\ \delta(P) \subseteq A_2,\ \rho(P) \subseteq B_1,\ \rho(P) \subseteq B_2\}$ we get $\{\delta(P) \subseteq A_1 \cap A_2,\ \rho(P) \subseteq B_1 \cap B_2\}$, which means that case (i) above is miss-handled. It may be noted that this case corresponds to the RDFS, DAML+OIL/OWL interpretation of multiple ranges and domains [24][25].
- **"δ/ρ subsumes":** While the above intuitions are satisfied, $\{A_1 \subseteq \delta(P), A_2 \subseteq \delta(P),\ B_1 \subseteq \rho(P),\ B_2 \subseteq \rho(P)\}$ entails $\{A_1 \cup A_2 \subseteq \delta(P),\ B_1 \cup B_2 \subseteq \rho(P)\}$, so it seems we get $P(A_1 \cup A_2,\ B_1 \cup B_2)$ from $P(A_1,B_1)$ and $P(A_2,B_2)$. This is overly permissive, since it gives rise to *unintended* models, when $(x,y) \in P,\ x \in A_1,\ y \in B_2$ or $x \in A_2,\ y \in B_1$.

The previous two cases are ones where using sub-properties is appropriate.

- **"All/only":** The encoding $\{A_1 \subseteq \forall P.B_1,\ A_2 \subseteq \forall P.\ B_2\}$ supports polymorphism properly in the case when the A's are disjoint, but in case (ii) above we get $A \subseteq \forall P.(B_1 \cap B_2) \equiv \forall P.\bot \equiv (\leq 0\ P)$, which does not allow A to be related to anything via P, thus contradicting our intuitions for (ii). Therefore, we must replace the original axioms $\{A_1 \subseteq \forall P.B_1,\ A_2 \subseteq \forall P.B_2\}$ by a new set $\{(A_1 \cap \neg A_2) \subseteq \forall P.B_1,\ (\neg A_1 \cap A_2) \subseteq \forall P.B_2,\ (A_1 \cap A_2) \subseteq \forall P.(B_1 \cup B_2)\}$, another case of non-incrementality.
- **All/each, all/some, some/each, some/some**: All these encodings support polymorphism following an analysis similar to the previous case.

3.5 Summary

The various encoding schemes and their suitability for the three special aspects of *SN*, viz. domain and range inheritance, inheritance blocking and polymorphic relationships are summarized in Table 2.

Table 2. Interpretation, axioms and support for *SN* requirements

Interpretation	Encoding	δ/ρ Inheritance	Inheritance Blocking	Polymorphic Relations
δ/ρ equals	δ(P) ≡ A ρ(P) ≡ B	No/No	N/A	No
δ/ρ subsumed	δ(P) ⊆ A ρ(P) ⊆ B	No/No	N/A	Missed model
δ/ρ subsumes	A ⊆ δ (P) B ⊆ ρ(P)	Yes/Yes	Exceptions + compensation	Unintended model
all / some	A ⊆ ∃P.B	Yes/No	Exception in axiom	ok
all / only	A ⊆ ∀P.B	Yes/No	Exception in axiom	Modification
some / some	≥1(A ∩ ∃P.B)	No/No	N/A	ok
some / all	≥1(A∩∀¬P.¬B)	No/Yes	Exception in axiom	ok
all / each	A ⊆ ∀¬P.¬B	Yes/Yes	Exceptions + compensation	ok

4 First Steps towards a Representation Choice Methodology

In the previous section we considered a list of alternative encodings of the *SN* into *DL*. We now propose an (incomplete) set of questions that could guide ontology developers in making choices among alternative representations in a formal ontology language such as OWL; these might form the basis of a methodological framework. The questions fall into several categories:

(a) Does the encoding support the "inferences" of the original notation?
(b) Does the encoding support inferences needed by expected applications?
(c) Does the encoding provide a reasonable intuitive model of the domain?
(d) Is the encoding supported by the formal ontology language and its reasoner?

Let us examine the alternatives from Section 3 in this regard as a way of illustrating and adding details to the list given below.

(a) Support for Inferences of the Notation

After identifying a number of possible representations for the node+link notation of *SN*, in Section 3, we looked to see which provided a logically consistent mechanism for performing inferences *explicitly, though informally, sanctioned by the notation*. Surprisingly, it appears that for *SN* the "**all/each"** encoding most closely captures these intuitions, with "**δ/ρ subsumes"** as the next best encoding. Several other aspects need to be considered, when dealing with graphical notations:

Does an encoding entail unintended inferences?

The "some/some" statement has the effect of "upwardly" inheriting a link to all the superclasses of the nodes associated with the link. This requires the ontology developer to identify such situations and assert axioms to prevent them.

Can/should something be inferred from the absence of a link?

It is not clear in *SN* whether links should be read as type constraints in programming languages: what is not explicitly permitted is forbidden. If so, the encoding A ⊆ ∀P.B

doesn't prohibit instances of ¬A being related to B. To prevent this, we would have to add the axiom (¬A ⊆ ≤ 0 P).

Should relationships be inferred to be asymmetric by default?
This is a special case of the previous general "default negation". According to [2], in the *SN*, links are "*usually asymmetric – MedicalDevice prevents PathologicFunction but not vice versa*"; moreover one can specify it: adjacent_to ≡ adjacent_to⁻. Axioms to this effect can be added automatically during translation, although asymmetricity again requires non-standard axioms: ¬(P ≡ P-).

Are the is-a children of nodes disjoint?
In the *SN*, there are some examples where this is not the case. Horrocks and Rector [17] give good arguments that the proper way to model ontologies is to start with trees of disjoint primitive concepts, and *define* everything else in terms of them.

(b) Support for Intended Application
If it is important to detect inconsistency in an ontology without overloading [11] e.g., {hasGender(FATHER,{male}), hasGender(FATHER,{female})}, where the relationship hasGender relates a concept FATHER to concepts represented as enumerations, e.g. {male}, {female}. An encoding of the form A ⊆ ∀P.B will not infer inconsistency, unless one also adds A ⊆ ∃P.T. Alternately, if the application uses only a limited set of inferences (e.g., because the form of the questions asked is limited), then one may not need to represent difficult kinds of axioms (e.g., properties being asymmetric). Such criteria can be criticized on the grounds that an ontology is supposed to be "application neutral", although there is always some arbitrary decision to be made about what is included and what is not.

(c) Reasonableness of the domain model
A strong case can be made that one should start first with an idea of how the world should be *modeled* – what are individuals, properties, etc, in the domain of discourse, and how concepts related to them, tied to the denotational semantics of the formal notation. The standard interpretation of a concept is a set of individual objects in the world, connected by properties (DL denotational semantics). Thus, causes (Bacteria, Infection) constrains the way in which any particular individual bacterium can be related by "causes" to a case of infection. So the questions are:

What are the intuitive encodings?
The "**all/some**" encoding Bacteria ⊆ ∃cause.Infection seems to be the representation of choice in several sophisticated medical ontologies developed with DL-knowledgeable collaborators [10][11], although in any state of the world some bacteria may not cause any infection. On the other hand, the "**all/only**" encoding Bacteria ⊆ ∀cause.Infection, is most frequently used in object-centered representations (e.g., [16]), though it runs into problems with polymorphism, since some bacteria might cause rashes and infections. Either way, interpretations such as "*all bacteria cause all infections",* or "*there is a bacterium that causes some infection*" seem quite odd, and were in fact rejected out of hand in [9].

Is an alternative interpretation possible?
The above seems rather discouraging for the **"all/each"** encoding, even though it satisfies all the *SN* requirements. However, suppose we prefix the relationship names by *"can/could_have/could_be"*: *"a bacterium could be the cause of any case of infection*". The resulting reading is much more reasonable, and may explain the inheritance and polymorphism properties of the *SN*, especially when noting that most relationships in the *SN* (unlike `fatherOf`, say) can be read this way.

The **"some/some"** reading – "some bacterium causes some infection", also seems to be unnatural, but interestingly McCray and Nelson explicitly endorse it: "**a relation is only established if at least some members of a set map onto at least one member of another set linked by the relationship**"[2]. The explanation for this may be that in the medical informatics community, concepts such as `INFECTION` are viewed by many as having kinds of infections, rather than specific cases of those infections, as instances. This is hinted at by the UMLS terminology of "labeling" concepts in the MT by the semantic types in the SN, rather than saying that the MT concepts are subclasses of semantic types, which form a high-level ontology. We note that this approach has been successful in the context of medical research literature search/retrieval, and propose this as an interesting topic of future research in the OWL/RDF context.

(d) Representation and inference in ontology language
Can the desired encoding be expressed in the ontology language of choice?
Given that the Semantic Web appears to be settling on a common ontology language (OWL) it is clearly important to encode the axioms in the constructs of this language. In this respect the representations based on role operators (negation, disjunction) and concept cardinality appear to be unsatisfactory. (But see below.)

Can the interpretation be represented using less "expensive" terms?
In addition to addressing the previous issue, reformulating an encoding using different constructors/axioms could provide significant computational benefits in view of the wealth of information about the computational complexity of various collections of DL concept and role constructors (e.g., see [7]) .

Is there some (better) way to capture the desired encoding knowing the technology used to implement the reasoner for the ontology language?
We have seen that one encoding of the some/some interpretation uses a language extension proposed by Baader et al. [8], who used esoteric techniques (based on "Hintikka trees" and automata), to reason about theses. There is an alternate encoding using nominals (individuals), and we saw that for the case of $(\geq 1\ C)$, there is even a solution in the basic ALC DL. Similarly, the axiom $\top \subseteq \forall P.C$ can be reasoned withmuch more efficiently by some tableaux reasoners than the logically equivalent $\exists P^{-}.\top \subseteq C$

Are there acceptable approximations to concepts/axioms?
Another approach to deal with the limited expressiveness of the ontology language, or the complexity of reasoning, is by representing concepts and/or axioms in an *approximate* way; this requires understanding what kinds of questions applications

need to have answered, in order to evaluate the price of information loss. Consider the axiom $P \equiv P_1 \cup P_2$, used in property polymorphisms. To avoid property union, not available in OWL, we can assert $P_1 \subseteq P$ and $P_2 \subseteq P$. Approximation of disjunction using hierarchies has been considered in [18], and there has been considerable research on approximation in *DL*s [19].

5 Conclusions and Future Work

We have reviewed in this paper some of our experiences with representing in OWL, a well established semi-formal ontology, the UMLS Semantic Network, which has been in use for over a decade in the context of medical informatics. Whereas the representation of the "vanilla" *SN* was straightforward, we encountered obstacles representing the semantics of "links" in the *SN*, especially in the context of requirements such as δ/ρ inheritance, inheritance blocking and polymorphism. This led us to investigate the possible interpretations and encodings of a "link" in the *SN*. We did not use role transitivity and number restrictions, but did use class disjunctions, role hierarchies, axioms, inverses, all and some property restrictions. The encodings were evaluated based on their support to *SN* requirements above. The various issues enumerated in this context should be considered by ontology and content developers while formally describing concepts in an ontology. Among the criteria we have identified are (i) support for inferences desired, (ii) intuitiveness of the resulting denotational semantic model, (iii) representation and effective reasoning in the ontology language. The latter involves formal worst case complexity results about the cost of reasoning, direct exploitation of the reasoner technology underlying the ontology language, and the possibility of approximate representation. We hope that the parts of a methodology provided above will be helpful to ontology developers that have embarked on the task of expressing their ontologies using OWL. We observe that although some of our problems could have been avoided if *SN* would have had a formal semantics. However, even starting with a language with equivalent translations to OWL, is not enough, since questions related to expressibility and intended modeling semantics, among others, still remain.

At the NLM, we are exploring various research issues related to the Semantic Web [3], both in the context of enhancing existing biomedical applications and for enabling new applications. Some ongoing investigations are: (a) The Semantic Vocabulary Interoperation project [21], which aims to provide tools and techniques to translate a term in a source biomedical vocabulary (e.g., ICD-9-CM) to a target biomedical vocabulary (e.g., SNOMED) by using the knowledge present in the *SN* and *MT*; (b) Potential improvement for searching and retrieving text and citation information by annotation of biomedical content using semantic web markup languages such as RDF and OWL; and (c) Enhancement and consistency maintenance of biomedical vocabularies based on reasoning with OWL descriptions as proposed in [5].

Acknowledgements. We would like to acknowledge helpful discussions with Alexa McCray, Olivier Bodenreider and Sergio Tessari. Support for this work was provided by the Oak Ridge Institute for Science and Education (ORISE).

References

[1] Lindberg D, Humphreys B, McCray A. "The Unified Medical Language System." Methods Inf Med 1993:32(4):281–91.

[2] McCray A, Nelson S. "The representation of meaning in the UMLS." Methods Inf Med 1995:34(1–2):193–201

[3] Berners-Lee T, Hendler J, Lassila O. "The Semantic Web." Scientific American, May 2001. http://www.sciam.com/2001/0501issue/0501berners-lee.html

[4] Horrocks I. "DAML+OIL: A Description Logic for the Semantic Web." IEEE Bulletin for the Technical Committee on Data Engineering, 2002.

[5] Stevens R., Goble R., Horrocks I. and Bechhofer S. "Building a Bioinformatics Ontology using OIL." IEEE Information Technology in Biomedicine, special issue on Bioinformatics, Vol 6(2):135–141, June 2002.

[6] Mena E, Kashyap V, Illarramendi A and Sheth A. "Imprecise answers in a Distributed Environment: Estimation of Information Loss for Multiple Ontology-based Query Processing." Int. J. of Cooperative Information Systems (IJCIS), 9(4), December 2000.

[7] "The Description Logic Handbook: Theory, Implementation and Applications." F. Baader, D. Calvanese, D. McGuiness , D. Nardi and P F Patel-Schneider (editors), Cambridge University Press, 2003.

[8] Baader F, Buchheit F, and Hollunder B. "Cardinality Restrictions on Concepts." Artificial Intelligence, 88(1–2):195–213, 1996.

[9] Rector A L, Rogers J and colleagues. "Introduction to Ontologies." Tutorial presented at the AMIA Annual Symposium 2002.

[10] Rector A L, Bechhofer S K, Goble C A, Horrocks I, Nowlan W A, Solomon W D. "The GRAIL Concept Modelling Language for Medical Terminology." Artificial Intelligence in Medicine, Volume 9, 1997

[11] Gangemi A, Pisanelli D M and Steve G. "An overview of the ONIONS project: Applying ontologies to the integration of medical terminologies." Data and Knowledge Engineering, 31(2):183–220, 1999.

[12] Woods W A. "What's in a link: Foundations for Semantic Networks." In R. J. Brachman and H J Levesque (editors), Readings in Knowledge Representation, 218–241. Morgan Kaufman Publishers, 1985

[13] Lutz C and Sattler U. "Mary likes all Cats." Proc. 2000 Int. Workshop on Description Logics (DL 2000).

[14] http://phd1.cs.yale.edu:8080/umls/UMLSinDAML/NET/SRSTR.daml

[15] Horrocks, I. "The FaCT system." In Proc. Int. Conf. Tableaux '98, Springer-Verlag LNCS 1397, pp. 307–312, May 1998.

[16] Calvanese D, Lenzerini M and Nardi D. "Unifying Class-Based Representation Formalisms." Journal of Artificial Intelligence Research (JAIR) 11: 199–240 (1999)

[17] Horrocks I., and Rector A. "Experience building a large, re-usable medical ontology using a description logic with transitivity and concept inclusions." Proc. Workshop on Ontological Engineering, AAAI Spring Symposium (AAAI'97).

[18] Borgida, A., and Etherington, D.W. "Hierarchical Knowledge Bases and Efficient Disjunctive Reasoning." Proc. KR'89, pp. 33–43

[19] Brandt S., Küsters R., and Turhan A.-Y. "Approximating ALCN-Concept Descriptions." In Proc. of the 2002 Int. Workshop on Description Logics (DL 2002).

[20] Medical Subject Headings – Home Page, http://www.nlm.nih.gov/mesh/meshhome.html

[21] The Semantic Vocabulary Interoperation Project, http://cgsb2.nlm.nih.gov/~kashyap/projects/SVIP.html

[22] OWL Web Ontology Language Guide, http://www.w3.org/TR/2003/WD-owl-guide-20030331/

[23] McCray, A. T. and Bodenreider O. "A Conceptual Framework for the Biomedical Domain." In R. Green, C. A. Bean and S. H. Myaeng, (editors), The Semantics of Relationships: An Interdisciplinary Perspective, Kluwer Academic Publishers, 2002
[24] http://www.w3.org/TR/rdf-mt/#rdfs_interp
[25] http://www.daml.org/2001/03/model-theoretic-semantics
[26] Schulz, S. and Hahn, U. "Medical Knowledge Re-engineering – converting major portions of the UMLS into a terminological knowledge base". International Journal of Medical Informatics 64(2–3):207–221 (2001).
[27] Cornet, R. and Abu-Henna, A. "Evaluation of a frame-based Ontology. A formalization-oriented Approach". In Proceedings of MIE 2002.