Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 38 (2005) 114–129
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Modeling a description logic vocabulary for cancer research

Frank W. Hartel[a], Sherri de Coronado[a,*], Robert Dionne[b], Gilberto Fragoso[a], Jennifer Golbeck[c]

[a] National Cancer Institute Center for Bioinformatics, 6116 Executive Blvd, Suite 403, Bethesda, MD 20892-8335, United States
[b] Apelon Inc. Ridgefield, CT, United States
[c] Department of Computer Science, University of Maryland, United States

Received 19 June 2004
Available online 11 November 2004

Abstract

The National Cancer Institute has developed the NCI Thesaurus, a biomedical vocabulary for cancer research, covering terminology across a wide range of cancer research domains. A major design goal of the NCI Thesaurus is to facilitate translational research. We describe: the features of Ontylog, a description logic used to build NCI Thesaurus; our methodology for enhancing the terminology through collaboration between ontologists and domain experts, and for addressing certain real world challenges arising in modeling the Thesaurus; and finally, we describe the conversion of NCI Thesaurus from Ontylog into Web Ontology Language Lite. Ontylog has proven well suited for constructing big biomedical vocabularies. We have capitalized on the Ontylog constructs Kind and Role in the collaboration process described in this paper to facilitate communication between ontologists and domain experts. The artifacts and processes developed by NCI for collaboration may be useful in other biomedical terminology development efforts.
Published by Elsevier Inc.

Keywords: Biomedical vocabulary; Ontology development; Cancer research; OWL; Description logic; T-Box model

1. Introduction

The NCI Thesaurus is a vocabulary designed to meet the needs of the cancer research community for consistent, unambiguous codes and definitions for basic and clinical concepts used in cancer research, and the semantic links among concepts that enable traversal of relationships. The NCI Thesaurus enables retrieval of information across a wide range of domains used in cancer research, facilitating translational research, the process of migrating basic research into clinical research and practice. It was built using Ontylog, a description logic (DL) developed explicitly for building large complex terminologies. Several other major terminologies also are developed in Ontylog, including SNOMED/RT, SNOMED/CT, and NDF/RT.

Use Cases from a variety of potential and current end users are driving the evolution of the NCI Thesaurus. This paper focuses on collaboration with end users, and more specifically, the methodology developed for interacting and collaborating with end users to ascertain their vocabulary needs. Other challenges relating to the development and maintenance of large vocabularies are not addressed here, among them, the challenge of distributed editing and the need to enable domain experts to edit vocabulary without understanding the underlying description logic.

Following Nardi and Brachman [1], we define the T-Box as that portion of the knowledge space that contains intensional knowledge, i.e., general knowledge, in the form of a terminology (p. 12). This paper discusses the tools we use to convey the semantics of

* Corresponding author. Fax: +1 925 377 5961.
E-mail address: decorons@mail.nih.gov (S. de Coronado).

1532-0464/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.jbi.2004.09.001

the T-Box to the domain expert users, so that they in turn can tell us what changes they need to the vocabulary. Our collaboration methodology is based on the use of a pseudo T-Box model that displays the semantic relationships in the Thesaurus, but does not contain all of the instantiations of these relationships.

To lay the groundwork for understanding the collaboration methodology, we first describe the context for cancer research and the NCI Enterprise Vocabulary Services, and the features of Ontylog DL. We then discuss the methodology for developing and maintaining the Thesaurus for science research via a collaborative process among the ontologists and domain experts. Lastly, we discuss the conversion of the Ontylog version of Thesaurus into Web Ontology Language (OWL) Lite, to provide a mechanism for wider dissemination.

2. Background

2.1. NCI enterprise vocabulary services project

The National Cancer Institute (NCI), a component of the US Department of Health and Human Services, National Institutes of Health, is the largest US source of funding for basic and clinical research into the causes, prevention, and treatment of cancer. It has a major goal to lessen the time between the occurrence of a basic research insight and incorporation of the insight into clinical prevention and treatment. The NCI Enterprise Vocabulary Services (EVS) Project (described below) and in particular, the NCI Thesaurus [2], is intended to facilitate translational research.

The NCI EVS is a set of vocabulary services intended to facilitate the integration of knowledge stored in the many disparate data sources created in cancer research. The two major products are the NCI Thesaurus, a stand-alone description logic vocabulary containing NCI terminology, and the NCI Metathesaurus, which is derived from the National Library of Medicine (NLM) Unified Medical Language System (UMLS) Metathesaurus. The EVS Project is a collaboration between the NCI Office of Communication (NCI OC) and the NCI Center for Bioinformatics (NCICB).

2.1.1. NCI Thesaurus

The NCI Thesaurus provides a unique combination of features not found elsewhere. It includes broad coverage of the cancer domain, including cancer related diseases, findings and abnormalities; anatomy; agents, drugs and chemicals; genes and gene products and so on. In certain areas, like cancer diseases and combination chemotherapies, it provides the most granular and consistent terminology available. NCI Thesaurus contains about 6000 cancer disease concepts, with careful attention to synonymy, organized both by organ site and morphology, and reviewed by members of the College of American Pathologists. It also contains almost 3000 chemotherapy regimen concepts, which are found nowhere else. Unlike many other biomedical vocabularies, NCI Thesaurus combines terminology from numerous cancer research related domains and provides a way to integrate or link these kinds of information together through semantic relationships. For example, cancer diseases currently are being modeled with role relationships that link diseases to molecular abnormalities. The NCI Thesaurus currently contains over 34,000 concepts, structured into 20 taxonomic trees. It is a living vocabulary. Some areas are more fully modeled with semantic relationships or properties (attributes) than others at the present time, and user needs dictate where we put our resources. We also collaborate with other Federal agencies where possible, to exchange and reuse vocabulary, and participate in standards organizations like HL7. Importantly, the NCI Thesaurus also provides concept history tables [3] to record changes in the vocabulary over time as the science changes.

The NCI Thesaurus is integrated with the caCORE bioinformatics infrastructure [4], including the caDSR (Cancer Data Standards Repository), which is used in developing clinical trial protocols, and in storing the metadata about those protocols, and caBIO, a set of biomedical objects, middleware and programming interfaces. It is used by the caBIO "portal" applications like the Cancer Models Database and Cancer Image Database. It is also used with the Cancer.gov website. Although we built NCI Thesaurus primarily for NCI purposes, it is increasingly being used outside of NCI. The caBIG (Cancer Bioinformatics Grid project), which includes collaboration among a large group of cancer centers, is interested in using it, or parts of it, separately and/or as part of the caCORE infrastructure. However, because it is publicly available and requires no registration, we are not sure how widespread its use is. NCI Thesaurus is published under an open content license in a number of formats including OWL [5]. The current and prior versions of Thesaurus, and the open source license, are available for download at ftp://ftp1.nci.nih.gov/pub/cacore/EVS/. An API provides direct access for applications, and phone support via a helpdesk is available.

2.1.2. NCI Metathesaurus

NCI Metathesaurus is based on the UMLS Metathesaurus, a concept based system that maps terms that mean essentially the same thing in numerous biomedical vocabularies to each other and assigns a unique concept identifier [6]. NCI removes certain vocabularies not relevant for cancer researchers and adds other vocabularies important to cancer research. NCI Metathesaurus provides rich synonymy, easy browsing, lexical search capability, links to related terms and the ability to find the

relevant term in one vocabulary, given the term in another vocabulary.

2.1.3. Access to NCI Thesaurus and Metathesaurus

NCI Thesaurus and several other terminologies of import to cancer researchers are available on the NCI Terminology Service, and access to the NCI Metathesaurus is available on NCI Metathesaurus Service. Access via a public domain API is provided as part of the caCORE (http://ncicb.nci.nih.gov/core). Web access to the NCI Terminology Service is available at http://nciterms.nci.nih.gov/, and to the NCI Metathesaurus Service at http://ncimeta.nci.nih.gov.

2.2. Intensional knowledge representation in NCI Thesaurus

Ontologists distinguish between intensional knowledge, i.e., general knowledge about an issue, problem or field of study, and extensional knowledge, i.e., application of intensional knowledge to individuals. Ontologists working in description logic use a graphical representation to distinguish extensional and intensional knowledge. A box containing an acyclic graph or some other abstraction represents axioms, classes and relations which compose the intensional knowledge. A second box containing a diagram of some sort is used to represent the assertions that can be derived from the intensional knowledge about individuals that belong to the classes. By convention, the box containing the representation of intensional knowledge is called the T-Box, and the box containing the assertions about individuals is called the A-Box. (The T-Box contains the abstract *terminology* model: the A-Box contains the *assertions* that can be inferred about concrete individuals belonging to the classes using the information in the T-Box). In description logic and artificial intelligence, the T-Box is the ontology, while the T-Box and A-Box together compose a knowledge base against which reasoning can be performed.

Design of the T-Box is ubiquitous in development of complex ontologies, with considerable published work in this area [7], but there has been little discussion in the literature of practical methods for dealing with real-world challenges that ontologists must overcome as they build, refine and maintain T-Box models. In cancer research many fields of knowledge interact, so development and validation of the T-Box model is especially challenging. Not only must the ontologists understand the formalisms of the DL, but also they must find effective ways to explain the model, including its formal and operational properties, to subject matter experts from many end user communities.

By the conventions of DL, the T-Box describes the classes and the relationships that are available for making inferences about instances contained in the A-Box. The NCI Thesaurus contains no A-Box. The concepts of interest in a biomedical ontology are largely abstract classes, while individual instances of biomedical concepts are found in data repositories that are external to the Thesaurus. Inference about these instances is possible; the NCI Center for Bioinformatics provides a technology stack, the caCORE [4], which enables certain aspects of such reasoning. From the viewpoint of the NCI Thesaurus user, it is fair to say that the T-Box constrains the *questions* that the user can obtain answers to using the ontology.

Ontologies or vocabularies in rapidly changing domains such as cancer research will be useful to their intended audience only to the extent that ontologists and practitioners from the disciplines who will use the ontology can collaborate in its development and refinement. NCI has invested heavily in recruiting and training domain experts from disciplines such as pathology, oncology nursing, pharmacology, molecular biology, public health and epidemiology as ontologists proficient in using software tools to build and maintain our terminology products [8]. NCI has therefore a rare resource: ontologists who are also domain experts and can communicate with their scientific and clinical colleagues using the conventions of thought common to their fields of study. This resource has proven powerful when defining and representing in the Thesaurus the overlap and divergence in meaning and nuance between similar terms used in the various cancer research disciplines. However our ontologists still have had difficulty dealing with domain expert users over the issue of semantic associations between concepts.

The problem stems from the ontologists' structural viewpoint, i.e., ontologists ask "What constraints do I need to model?" while the users have a functional viewpoint, asking "What questions can I answer using the ontology?" We believe that this divergence of viewpoints impedes ontologist–user collaboration. In the case of the Thesaurus, the problem is exacerbated by the large number of scientific disciplines that need to use the Thesaurus, and whose needs and preferences have to be learned and engineered into the T-Box. Our methods for beginning to bridge this gap are described later in the paper.

2.3. Ontylog

The NCI Thesaurus was built using Ontylog DL.[1] Ontylog, like most DLs, can be given a model or set-theoretic semantics in which concepts describe sets of individuals from some universe, hence the name description logic. DLs, subsets of first order logic and originally called term subsumption systems, evolved out of early

[1] Ontylog is a publicly available description logic. It has been implemented by Apelon in their proprietary Terminology Development Environment (TDE).

work on frames. [9,10]. Early systems dealt with classes and instances and slots in the classes that could point to other classes. DLs are more formal in that concept descriptions are terms in the logic. Various practical systems have been built using DLs. [11–15].

Classification in DLs is a form of inference and normally there is a trade-off between expressiveness of the logics and space and/or time complexity of the inferencing. Recently, DLs have seen growing interest due to their importance to the OWL standard in the W3C committee work on the Semantic Web [16]. At NCI interest in DL predates OWL, however. Many of the features in Ontylog have been provided to address the complexity of modeling in the cancer science domain as well in clinical medicine and pathology (SNOMED).

Ontylog is a descendent of K-REP, a declarative knowledge representation language, also a DL [17]. The fundamental construct in Ontylog is the concept. Concepts can be defined compositionally in terms of other concepts, thereby inheriting information. Concepts are arranged into a directed acyclic graph through a computational process called classification. This is a pair wise operation; two concepts are compared to see if one concept subsumes the other. This subsumption process computes the is_a relation and it is this relation that is reflected in the graph. It is often helpful to think of this graph as a taxonomy where concepts can have multiple parents, the reason being that there is a natural sense in which information is increasing as one moves from parents to children in the graph.[2]

Another major feature of concepts in almost all description logics is the notion of primitive versus defined. This difference can be quite subtle and is best explained by appealing to the first order semantics. Assume that a concept is mapped to a set in the semantics so when we talk about an instance belonging to a concept we mean set membership. A primitive concept is considered to be incomplete as a definition. The conditions on the concept (i.e., its collection of role constructs) are considered to be necessary but not sufficient. Logically an implication exists in one direction. By this is meant that for instances known to be in the set modeling the concept, the conditions necessarily hold, but for any instance in the universe, its satisfaction of the condition is not *sufficient* for it to be a member of the set modeling that concept. (In other words the "only if" implication does not hold.) In other words the classifier will never compute that a given primitive concept subsumes another concept because the conditions are not sufficient for set membership. As an example consider *Non-Hodgkin's Lymphoma*. Suppose it was defined as a *Lymphoma* whose cellular origin is *B-Cell*, or *T-Cell* or *NK-Cell*. Furthermore suppose it was designated as a primitive concept. This implies that for any instance of *Non-Hodgkin's Lymphoma* it is necessarily true that its cellular origin is *B-Cell*, *T-Cell* or *NK-Cell*. However, the converse does not hold. An instance of *Lymphoma* whose cellular origin is *B-Cell*, or *T-Cell* or *NK-Cell* is not sufficient to make it an instance of *Hodgkin's Lymphoma*. This is what it means to be primitive versus defined. For a defined concept the implication exists in the other direction as well, i.e., the "only if" direction. Whether or not a definition is complete and should be designated as a defined concept is for the modeler to decide, as it affects the semantics of classification.

Besides is_a, there are other binary relations between concepts, called roles. Concepts are given meaning compositionally as conjunctions of other concepts and role expressions or phrases. Role phrases make logical assertions about the relations between concepts. When a concept is defined in terms of other concepts it inherits information from those concepts, specifically their roles. As a simple example consider the following concepts from the NCI Thesaurus.

Lymphoma
  Diffuse Large B-Cell Lymphoma
    Immunoblastic Lymphoma (inferred subconcept)
Neoplastic B-Cell
  Neoplastic Large B-Cell
    Neoplastic B Immunoblast (stated subconcept)

Consider also the role Disease_Has_Abnormal_Cell, which is a relation from these cancers to different abnormal or neoplastic cell types. Some of these concepts can be defined in terms of others, for instance *Diffuse Large B-Cell Lymphoma* is_a *Lymphoma* with abnormal cells of type *Neoplastic Large B-Cell*. *Immunoblastic Lymphoma* is_a *Lymphoma* whose abnormal cell type is *Neoplastic B Immunoblast*. Notice that if it is known or previously asserted that a *Neoplastic B Immunoblast* is_a *Neoplastic Large B Cell* then a classifier would infer that *Diffuse Large B Cell Lymphoma* subsumes *Immunoblastic Lymphoma*, based on the comparison of the values for the role Disease_Has_Abnormal_Cell.

As relations between concepts, roles are what drive the building of an ontology in determining the structure of the graph. A concept is modeled in the semantics of Ontylog as a set of things. It is the roles as relations between these sets of things that determine how the subsumption graph is created. As one progresses down the graph through the is_a relation, each description adds more information as smaller and smaller sets of individuals are described. Having the roles determine the concept definitions and organizing the graph according to the subsumption relation fits well with how

[2] For example, Arthropods have segmented body, exoskeleton and jointed appendages. Arachnids have 4 pairs of jointed legs, 2 body segments and 1 pair of pedipalps. Lobsters, Beetles and Spiders are all arthropods, but only spiders are Arachnids.

Table 1
Ontylog language[a]

Constructor	Syntax	Semantics
Concept name	C	C^I (where $C^I \subseteq \Delta^I$)
Top	$\top$	Δ^I
Bottom	$\bot$	$\varnothing$
Conjunction	$C \sqcap D$	$C^I \cap D^I$
Disjunction	$C \sqcup D$	$C^I \cup D^I$
Universal restriction	$\forall R.C$	$\{x \mid \forall\, y\colon R^I(x,y) \rightarrow C^I(y)\}$
Existential restriction	$\exists R.C$	$\{x \mid \exists\, y\colon R^I(x,\ y) \wedge C^I(y)\}$
Modal restriction	$\Diamond R.C$	$\{x\ \Pr(\exists\, y\colon R^I(x,y) \wedge C^I(y)) > 0\}$
Role name	R	R^I (where $R^I \subseteq \Delta^I \times \Delta^I$)
Definitional or axiomatic constraint	**Syntax**	**Semantic constraint**
Concept definition	$C \doteq D$	$C^I \equiv D^I$
Concept subsumption axiom	$C \sqsubseteq D$	$C^I \subseteq D^I$
Role subsumption axiom	$R \sqsubseteq S$	$R^I \subseteq S^I$
Right identity axiom	$R \circ S \doteq R$	$(R \circ S)^I \equiv R^I$

See Appendix "Description Logic Terminology" from The Description Logic Handbook [18] for notation, syntax and semantics descriptions. Also see http://whatis.techtarget.com/definition/0,,sid9_gci803019,00.html for a table of mathematical symbols; http://www.unicode.org/charts/PDF/U27C0.pdf for a description of the modal operator symbols; and http://plato.stanford.edu/entries/logic-modal/ on the poss operator and modal logics.

[a] Note that disjunction, a feature currently under development can only be used in role values.

knowledge is organized when building ontologies based on classification graphs.

The phrase is_a refers to both the computed subsumption relation as well as the defining super concept relation that the classifier starts with when computing subsumption. We distinguish these by using the terms defining super concepts to refer to those is_a relations that are part of the concept's definition and direct super concepts to refer to those is_a relations that are discovered during classification. Using the example above, if *Diffuse Large B-Cell Lymphoma* subsumes *Immunoblastic Lymphoma* then this is an inferred relationship. Using an appropriate software tool,[3] such as the TDE, in the inferred view one would see *Immunoblastic Lymphoma* as a child of *Diffuse Large B-Cell Lymphoma* in a hierarchy viewer. If one switched to the defined view one would see it as a child of *Lymphoma*.

There are two pre-defined concepts, *Top* and *Bot*, which serve special roles in the graph. *Top* is the empty definition; it contains no information and so describes the entire universe, in other words *anything*. *Bot* is used when concepts are inconsistent, for example, *Vegetarian_Pizza_with_Meat*. *Bot* describes the empty set. These two concepts are used for technical purposes. *Top* allows the classifier a convenient place to start and *Bot* is used as a placeholder for a definition that is logically inconsistent. In Ontylog, these predefined concepts are not exposed to the user.

A reasonably precise definition of Ontylog in terms of logical syntax and the usual model theoretic semantics is provided in Table 1. Table 1 does not provide a complete definition of Ontylog. Ontylog has features outside most description logics, for example, kinds and role groups. We will discuss the basic operational semantics for these features in order to make clear how they are used in practice.

In the Ontylog implementation Apelon provides, the TDE product employs an intensional approach to classification that takes advantage of the compositional nature of the definitions. The intensional approach is not only the key to Ontylog's scalability but allows Ontylog to include features not found in many other DLs. In this approach concepts are compared pairwise as the classification process puts them in the graph. The nature of the walking algorithms is such that no inherited information is copied and only local information is used in comparing two concepts. These algorithms have been optimized over the years to be as fast as possible, using heuristics that have been developed from the observation of real world modelers. Ontylog also supports Kinds, and under certain conditions these allow the graph to be partitioned into smaller graphs for purposes of classification. Ontylog was designed to address certain real world practical problems in the construction and maintenance of large clinical medical terminologies. In description logics there is often a trade off between the expressivity of the logic and its performance and scalability. Tableaux methods enable more expressiveness, but do not scale as well with respect to time. Ontylog favors scalability in time, and the scalability is quite good.[4]

[3] As described below, NCI Thesaurus is available in OWL Lite format in which format it is viewable and editable in the publicly available Protégé with the OWL tab. Protégé and OWL are available at http://protege.stanford.edu/.

[4] Using a Dell Optiplex GX400 running Win2000, 1.7 GHz CPU, 256MB RAM, the TDE 3.0 Ontylog classifier classifies NCI Thesaurus in under a minute.

If the modelers create a healthy amount branching in the upper regions of the knowledge base, classification performance will be closer to $n \times \log(n)$ rather than the n^2 worst case. Depending on the complexity of the terms Ontylog can classify knowledge bases with up to 400K concepts in reasonable time (about 13 min).

Typically primitives are used at the top of a graph to help create a certain amount of branching. Top level branching assists human comprehension and may be vital to federation of ontologies [19]. As concepts can be defined in terms of other concepts, and those concepts can be defined in terms of others, one can imagine unwinding a concept all the way up to Top. If we do this for two given concepts we can see if they share a common set of primitives. The Ontylog classifier will never place a primitive above another concept. In order for one concept to subsume another they must share common primitives.

We will see an example in a later section of how kinds have been used to help organize NCI Thesaurus. We will also see that other features that were added to Ontylog to improve performance have turned out to have utility in helping modelers organize terminologies, and collaborate better with domain experts, specifically the use of right identities, and role groups, discussed below.

Returning to subsumption in Ontylog, we say that one concept A subsumes another B if A's primitives are a subset of B's primitives. For all role phrases in A, B must have that role phrase as well and the value restriction on the role belonging to A must subsume the value restriction of the role belonging to B. For example (∀ Disease_Has_Normal_Cell_Origin *Mature B-Cell*) is subsumed by (∀ Disease_Has_Normal_Cell_Origin *B-Cell*). When a new concept is added to the knowledge base the classifier walks the graph comparing it to other terms in order to find its immediate parents and immediate children. These are often referred to as "most specific subsumers" and "most general subsumees" [20].

As mentioned previously, NCI Thesaurus does not contain extensional knowledge. Many description logics distinguish between generic concepts that describe sets and individuals that describe actual instances or elements of the sets. They break the collection of terms up into a T-Box for the former and an A-Box for the latter. Ontylog does not support instances for two reasons. First, it was designed for building large and complex vocabularies and intended for embedding in runtime clinical systems. In these systems instances are typically constituents of individual patient records and are often stored in databases where transaction semantics and other core system concerns are paramount. The kind of inferencing that would be performed over the instances is readily performed over the generic concepts so there is no need for assertions about individuals. Secondly, Ontylog has an enforced semantics; there are no exceptions allowed, i.e., conditions applied to subconcepts that do not apply to parent concepts. Some other description logics allow exceptions in the instances. Exceptions are considered risky in medical knowledge [21]. Also, deciding where to draw the line between generics and individuals is extremely hard [22]. In fact, the distinction between generic and individual is considered a hard problem in mathematical foundations [23,24].

Kinds in Ontylog serve two purposes. First, they allow for the partitioning of concepts into disjoint collections. Second, kinds are used to define the domains and ranges of roles. It is easy to confuse Ontylog Kinds with the sets that provide semantics for concept definitions. Rather, think of each concept as having a unique kind associated with it. A concept can have only one kind associated with it. Kinds are introduced on the top-level primitive concepts and then propagated via inheritance. In the multiple inheritance case, checks occur that a concept does not inherit more than one kind. During this process of propagation, completion is computed—kinds are checked for consistency, roles that are multiply inherited are checked that their values are consistent, and the most restrictive value is collected. If there is not one then an anonymous conjunction is formed. Kinds are pair-wise disjoint. So, for example, one might have the kind disease and the kind drug. It's very clear that drugs and diseases are two entirely different things. One could readily argue that kinds are pretty much the same as disjoint primitives and from the viewpoint of the typical model theory semantics used in description logics this is an accurate characterization. However the semantic motivation for kinds is the view of kinds as a collection of primitive types with absolutely no assumptions about any model theory. It is more useful to view kinds as similar to typing in programming languages. The semantics of Ontylog in terms of type theory are not fully worked out and certainly outside the scope of this paper. However we mention it because although kinds were motivated by types in order to support right identities they turn out to have utility as a modeling and organizing device as described in the following section.

In Ontylog, roles are given a definition that states a domain and a range in terms of kinds. These domains and ranges restrict the concepts on which a role can be used. For example, one might define the relation *treats* that has the domain *drug* and the range *disease*. Role hierarchy is supported in Ontylog. Roles can have parents; a somewhat contrived example might be *cures* as a subrole of *treats*. From the fact that aspirin cures fever one could infer that aspirin treats fever. As a binary relation the ordering induced by giving a role a parent is exactly subset inclusion.

When a role is used as part of a concept definition it has a logical modifier and a value. The value is another concept that must belong to the range kind. In Ontylog

the logical modifier can be some, all, or poss (possibly). The most commonly used modifier is some. The modifier poss is a recent addition to handle the case of non-defining roles or what we call contingent roles. In Table 1 poss corresponds to the diamond modal operator. At times modelers would like to use roles as part of a definition and have them inherited but not want them to be defining, i.e., not have them affect the classification of the concept in question. In the newest version of Ontylog, the role modifier called poss (possibly) has been introduced expressly to address the need for this feature. When this modifier is used the role still is treated the same for purposes of inheritance, but the role does not participate in any subsumption tests. We have chosen the term poss because as the term implies this is indeed a modal operator. Modal logics typically involve the notions of "possibility" and "necessity". Like description logics they can be given classical set theoretic semantics but most often are given what are called Kripke or possible world semantics. The key idea is that possible worlds arise in an incremental fashion as knowledge grows. Things that may not be known (possible) in one world may be known in a later more complete world. Fuller discussion of these semantics is not appropriate for this paper; it suffices to appeal to the operational semantics that are readily accessible. Roles qualified with poss are inherited and completed for a concept but they do not participate in classification.

A typical partial set of definitions might look like the following:

Lymphoma $\subseteq \top$

Hodgkin's Lymphoma $\doteq$ *Lymphoma* $\sqcap$ (**some** Disease_Has_Normal_Cell_Origin (*B-Cell* $\sqcup$ *T-Cell* $\sqcup$ *NK-Cell*))

Nodular Lymphocyte Predominant Hodgkin's Lymphoma $\doteq$ *Hodgkin's Lymphoma* $\sqcap$ (**some** Disease_Has_Normal_Cell_Origin *Germinal Center B-Cell*)

Mature B-Cell Non Hodgkin's Lymphoma $\doteq$ *Lymphoma* $\sqcap$ (**some** Disease_Has_Normal_Cell_Origin *Mature B-Cell* $\sqcap$ **all** Disease_Has_Molecular_Abnormality *Clonal Immunoglobulin Kappa Light Chain Gene Rearrangement*)

Burkitt's Lymphoma $\doteq$ *Mature B Cell Non_Hodgkin's Lymphoma* $\sqcap$ (**poss** Disease_Has_Cytogenetic_Abnormality *t_8_14*)

Notice that the definition of *Hodgkin's Lymphoma* as "*Lymphoma* $\sqcap$ (**some** Disease_Has_Normal_Cell_Origin (*B-Cell* $\sqcup$ *T-Cell* $\sqcup$ *NK-Cell*))", illustrates that a concept definition is composed of a set of other named concepts, called defining superconcepts and a set of role phrases. Defining superconcepts themselves can be other definitions and so forth. As the classification process organizes them into a graph one can see that a concept might inherit many role constructs from its parents. As we move down the taxonomy the role values will become more specific in the subsumption order. Note however that in computing the is_a relation as basis for subsumption it is the is_a relation that is used on the role values. This will not pick up all subsumptions that one would want to make. In Ontylog, roles can also have what are called right-identities. Consider the concepts *Heart*, *Myocardium*, *Disease*, *Heart Disease*, *and Myocardium Atrophy*.

Assume the concepts have the following definitions:

(define-primitive-concept *Heart*)
(define-primitive-concept *Disease*)
(define-concept *Heart Disease* (is_a *Disease* and (**all** located_in *Heart*)))
(define-concept *Myocardium* (**all** part_of *Heart*))
(define-concept *Myocardium Atrophy* (is_a *Disease* and (**all** located_in *Myocardium*)))

Assume that *Disease*, *Heart Disease*, and *Myocardium Atrophy* are Disease_Kind and that *Heart* and *Myocardium* are Anatomy_Kind. Assume further that located_in is a role that maps Disease_Kind to Anatomy_Kind and part_of is a role that maps Anatomy_Kind to itself.

Normally, a DL classifier would look at *Myocardium Atrophy* and try to compare it with *Heart Disease* to infer that it is a subconcept. They both are defined in terms of the same primitive, *Disease*. However, when it compares (all located_in *Heart*) to (all located_in *Myocardium*) it sees that *Heart* does not subsume *Myocardium*. They are not in the is_a relation. Notice though that *Myocardium* is related to *Heart* through the part_of role. We would like to infer that since *Myocardium Atrophy* is located_in the *Myocardium* and *Myocardium* is part_of the *Heart* then *Myocardium Atrophy* is located_in the *Heart.* In logical terms, we would like to compose located_in with part_of and have part_of mean the same as located_ in. In a sense we would like to cancel part_of. Algebraically, a term that can be cancelled without changing the evaluation result is called a right identity (recall from arithmetic that 5 times 1 is equal to 5. In the ring of integers 1 acts as a multiplicative right identity).

The Ontylog classifier will make this inference provided the modeler states that part_of is a right identity for located_in. In this way, partonomies are introduced in a very elegant fashion. Partonomy is just another relation or role and it is linked to subsumption via this notion of the right identity. Right identities do add some computational costs to the classifier but in practice the additional cost is minimal and is greatly outweighed by the added functionality this provides to modelers.

Role groups are a feature in Ontylog that enable the modeler to group roles within a given concept. The idea seems very close to the use of nested definitions.[5] The treatment of role groups during classification is straightforward. Recall the rule for normal treatment of roles; to see if concept A subsumes concept B check that for every role of A, concept B also has that role, that their modifiers agree and that their values are in the subsumption order. For role groups, the same rule is used and checked before the normal role rule. For every role group on A, check that concept B has a role group that is the same and the groups subsume one another. To test if the groups subsume one another apply the normal rule for roles to the two groups.

Table 1 gives a predominantly first order set theoretic semantics for Ontylog and we are confident that a full and complete semantics can be offered. However, it is likely that some of Ontylog's features involve higher order logics. For example, the role groups might involve quantification over groups of roles. It is important to note that Apelon's TDE takes an intensional approach to implementation of the Ontylog classifier. Subsumption testing is a structural process that involves walking the graph of concepts already classified in an optimal way. We maintain that the intensional methods provide a more scalable approach to classification, and that any lack of completeness issues not handled are marginal cases that do not impact typical modeling usage [26].

3. T-Box as a framework for collaboration

The Ontylog DL entities, especially kinds and roles, have proven useful in bridging the structural worldview of the ontologist and the functional view of ontology users. We have capitalized on these entities in various artifacts that we use in collaboration with domain experts from biomedical fields. Since Ontylog DL does not support the T-Box/A-Box distinction, one might argue that there is no true T-Box in NCI Thesaurus. The kinds and roles together represent the Thesaurus' *ability* to represent intensional knowledge. The intensional knowledge itself comprises the instantiations of the role relationships among the concepts. In that sense, the whole Thesaurus is the T-Box. However, the Thesaurus as a whole is too hard to grasp: it would be unreasonable to ask domain experts to grasp the whole 30,000 odd concept space as a prerequisite for collaborating in developing and refining the Thesaurus. Needing a simplifying representation, we have found that the kinds and roles alone, without the instantiations, provide suitable simplification, which we will call a "pseudo T-Box". The pseudo T-Box is easy for domain experts in biomedicine to grasp. The pseudo T-Box is a useful framework for collaboration because it enables domain experts to determine if the NCI Thesaurus will be able to answer the questions they want it to answer.

Use Cases are a critical component of collaboration with NCI Thesaurus user communities. The needs or "problems" expressed in Use Cases are generally pretty clear expressions of the questions that users wish the ontology to be able to help them answer. By relating these Use Case problems to entities such as roles and kinds in the pseudo T-Box, an effective bridge can be constructed between the functional world of the users and the structural world of the ontologist.

Table 2 lists major tabular artifacts used at NCI to relate user needs or problem statements to pseudo T-Box model entities. These artifacts, together with a wireframe graphic (Fig. 1), are used throughout the NCI Thesaurus design and refinement process.

The following paragraphs describe the structure and use of each artifact.

3.1. Use Cases

We first ask domain experts to consider what they would do with the ontology, assuming it could provide the needed capabilities. In the context of their normal research, where would they want to rely on the ontology? What sort of information would they have in hand and what additional information would the ontology provide?

Subsequently, we examine the existing pseudo T-Box model to determine which of the Use Case problems it currently can address. Those needs that it cannot address are noted. We use a form similar to Table 3 to record these data. We add to the right columns notations of existing kinds and roles that the Use Case would require and notations of any new kinds or roles that might be needed to satisfy the Use Case, as well as issues needing further clarification and the actions by each party.

We then discuss that Use Case form with the domain expert to confirm that we understand the problems. If there appears to be common understanding we proceed, if not we revise the Use Case description until understanding is achieved.

We then map the Use Case problems to Roles and Kinds. Table 4 contains a fragment of these mappings from the March 2004 version of the NCI Thesaurus.[6] The complete current list of mappings may be found at ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics.

[5] The authors have been unable to establish any *formal* relationship between nested definitions and role groups. Role groups have been shown to affect the overall semantics of a given collection of concepts depending on the groupings. Spackman et al. [25].

[6] The rows of this table contain all Roles and Kinds in the pseudo T-Box model, and the numbers of all the Use Case Problems that depend on that Role or Kind. If no kind or role in the existing pseudoT-Box model appears to meet the need implied by a problem, a tentative role or kind is created in the table, and the number of the Use Case is recorded with the tentative entry.

Table 2
Tabular artifacts used in pseudo T-Box construction and refinement

Use Cases	Describe domain expert interaction with ontology to answer specific question or series of related questions of interest to expert
Table of Use Case to role and kind mappings	List of each individual need in each Use Case that implies the need for a role or kind
Table of roles with domain and range	List of each role name and its domain and range. Roles clustered according to role hierarchy
Concept hierarchies	Is_A hierarchy for concepts in a kind.
Kind definitions	English language description of the range of coverage of the kind: what is and is not included. Definitions must be unambiguous and clearly disjoint.

The five sorts of tabular artifacts produced during design of the NCI Thesaurus semantic model are enumerated in this table. The current set of artifacts are available for download from the NCI Center for Bioinformatics Web site. See: ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/.

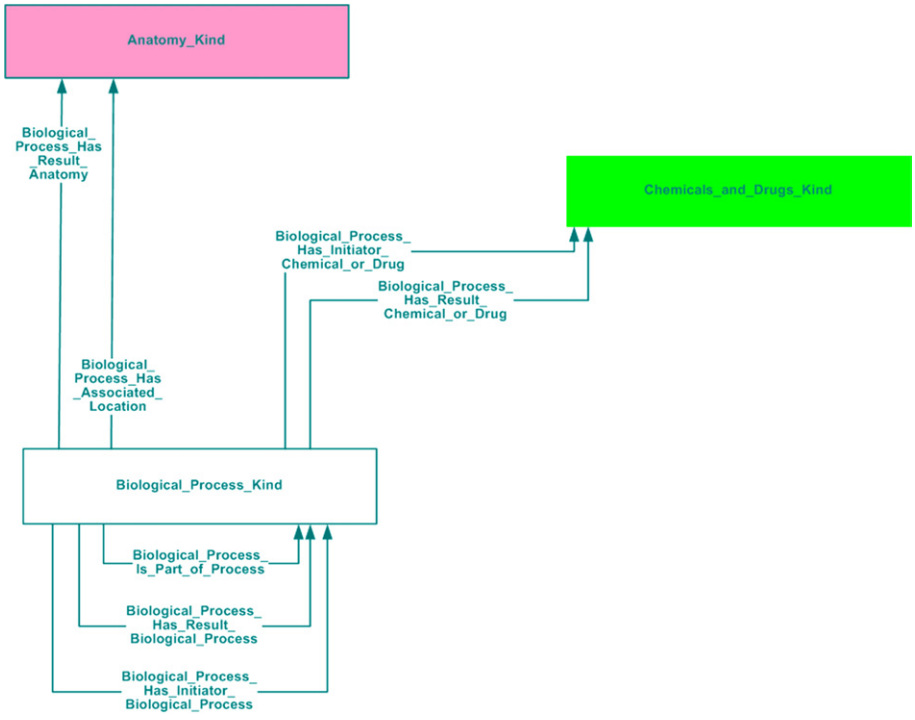

Fig. 1. The Biological Process layer of the NCI Thesaurus pseudo T-Box graphic.

Table 3
Use Case Format

Use Case number		Roles needed	Comments/steps
Use Case name	Use Case provider, organization and descriptive title for the Use Case		
Background	Précis of the scientific activity within which the Use Case occurs		
Problem statement(s)	Numbered list of the steps that the user needs to perform to complete the activity	Role(s) needed to satisfy need	Comments on role use
Solution	Summary of enhancements needed to enable the Use Case to be satisfied	Possible new roles needed to satisfy needs	Possible new kinds needed to satisfy needs
Follow-up actions	Numbered list of steps ontology team would need to perform to satisfy the Use Case and notations of requirements implied by Use Case that fall outside ontology domain	Specific issues requiring further discussion	List of steps to be taken by EVS or user community to address needs
Notes			

The fields of the NCI EVS Use Case form are described in this table. The purpose of each field is described briefly. See Table 8 for an example of a specific Use Case.

Table 4
Excerpt of spreadsheet mapping of roles to Use Cases

Role	Use Case: problem	Notes
Gene_Product_Malfunction_Associated_With_Disease	1.3 3.3 11.1	
Gene_Product_Plays_Role_in_Biological_Process	5.3	Appears to require 2nd role, Gene_Has_Function
Modality_Has_Associated_Attribute	8.5 10.5	etc
Modality_Has_Associated_Equipment	8.4 10.5	etc

In this example of mappings of Use Case to roles, the role Gene_Product_Malfunction_Associated_With_Disease is used to satisfy the third requirement of Use Case 1, the third requirement of Use Case 3, and the first requirement of Use Case 11. Similarly, the role Gene_Product_Plays_Role_in_Biological_Process is used to satisfy the third requirement of Use Case 5. However, it does not completely satisfy the requirement. An additional role appears to be needed to completely address the requirement. The current mappings of Use Cases to roles is available for download from the NCI Center for Bioinformatics Web site. See: ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/.

3.2. Concept hierarchies as a bridge into the T-box

We have found that it is better not to introduce the notions of roles and kinds into collaboration discussions with domain experts at this point. It works better to review with the domain expert those existing concept hierarchies that contain content relevant to his or her interests. The domain experts recognize the concepts in the hierarchies and very quickly can begin to assess the adequacy of the existing coverage. If coverage needs to be expanded, we make note of the needs, and draft expanded hierarchies will be reviewed with the domain expert in subsequent meetings. Discussion of the hierarchies frequently will elicit comments from the domain expert about the hierarchy structure. Not infrequently in biomedicine, there is no canonical determination of a concept's correct tree position. For example, meningococcal meningitis may be classified correctly as both a disease of the central nervous systems and a bacterial disease. So there are always things the experts will question. These discussions of why the hierarchies are structured as they are offer the opportunity to introduce the notions of roles, since the hierarchy position of defined concepts are the result of the concept's role restrictions.

Table 5 contains a sample of the Roles defined for the March 2004 build of NCI Thesaurus. (Go to ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics for the complete list).

If discussion of hierarchies leads the domain expert to express interest in roles, we discuss this table, or alternatively the wire frame graphic shown in Fig. 1. In either case, it is necessary to introduce the expert to kinds, as they are used to provide the domain and range of each role. Table 6 contains a few of the kind definitions for the March, 2004 build of NCI Thesaurus. (See the complete current list at ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics.) Kinds must be defined as unambiguous, non-overlapping sets of concepts. For example, the subdomain of microanatomy requires precise definition—What does it entail, and to which kind does it belong? Anatomy or Gene Product? Are structural proteins anatomical structures or gene products? In the NCI Thesaurus, the same concept cannot be both. Definitions of the kinds must make explicit how such distinctions will be made.

3.3. Graphical representations of T-Box

Domain experts sometimes do not comprehend readily the implications of the tables of role names and of kind definitions; therefore we also introduce the pseudo T-Box model in a graphical form. We use tables and graphics together because neither alone has proven entirely satisfactory. We have standardized on a box and arrow (wire frame) graphic to represent the T-Box model. We use a distinctly colored box for each kind and include the name of the kind. Arrows with role names

Table 5
Sample of the roles defined for the March 2004 release of NCI Thesaurus

Role name	Domain Kind	Range Kind
Anatomic_Structure_Has_Location	Anatomy_Kind	Anatomy_Kind
Anatomic_Structure_is_Physical_Part_of	Anatomy_Kind	Anatomy_Kind
Biological_Process_Has_Associated_Location	Biological_Process_Kind	Anatomy_Kind
Biological_Process_Has_Initiator_Chemical_or_Drug	Biological_Process_Kind	Chemicals_and_Drugs_Kind
Biological_Process_Has_Initiator_Process	Biological_Process_Kind	Biological_Process_Kind
Biological_Process_Has_Result_Anatomy	Biological_Process_Kind	Anatomy_Kind
Biological_Process_Has_Result_Biological_Process	Biological_Process_Kind	Biological_Process_Kind
Biological_Process_Has_Result_Chemical_or_Drug	Biological_Process_Kind	Chemicals_and_Drugs_Kind

Numerous roles included in the NCI Thesaurus Semantics model. Each role has a domain and range, both of which are constrained to be one and only one Kind. From time to time new roles are added. Not all roles are instantiated in the NCI Thesaurus; some are included in the model because they are expected to be needed in a future round of content creation. Occasionally roles are replaced or eliminated from the model. See: ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/ for a list of the roles in the current model.

Table 6
Some of the kinds defined for the March 2004 release of NCI Thesaurus

Kind	Description of kind's coverage
Gene_Products	Endogenous RNAs, proteins, protein complexes and riboprotein complexes. Excludes exogenous chemicals
Molecular_Abnormality	An enumeration of the molecular abnormalities that occur in human cells and tissues and non-human models of human cancer. Includes abnormalities such as translocation, polymorphism, underexpression, overexpression
Diagnostic_and_Prognostic_Factors	Characteristics of the organism or of a process that contributes to clinical diagnosis, treatment selection or prediction of clinical outcome.
Combination_Chemotherapy	Combinations of multiple drugs used in standard and clinical trial treatments. They do not currently specify order, dosage or dosing interval of the individual ingredients
Equipment	Supplies or apparatus used for cancer-related research, diagnosis or therapy
Organism	A living entity

It is vital to develop clear, obviously non-overlapping definitions in English language for each Kind. Neither editors nor users of the NCI Thesaurus could succeed if they were in doubt about the coverage of each Kind. Role semantics, hierarchy and concept meaning would all be undermined if Kinds were seen as overlapping. The current table of Kind definitions is available for download from the NCI Center for Bioinformatics Web site. See: ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/.

represent roles.[7] Because many kinds and roles are required to satisfy the needs of cancer researchers, the graphic is large and visually complex. It is helpful to use a graphics software application that supports layers. One layer of the wire frame graphic is provided in Fig. 1. (See the complete current graphic at ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics.)

We assign each kind to a separate layer. On the kind's layer we place the arrows for each of the roles that have the kind as their domain. We have adopted the practice of including the domain and range names in the name of each role, e.g., Disease_Has_Primary_Anatomic_Site has the domain *Disease_Kind* and the range *Anatomy_Kind*. Therefore, it is not necessary to display the range kinds on the layer, but some users are bothered by the range kinds not being visible, so we display them.

3.4. Other uses of collaboration artifacts

The materials described above are the foundation of the process of pseudo T-Box refinement and validation that has been developed at NCI. They are used not only for collaboration with domain experts, but also by ontology designers in their day to day work, for instance, in assessing whether the roles defined for a kind provide necessary and sufficient restrictions to define the concepts; useful subsumption relation within the kind; and assessing the vulnerability of the T-Box to assertions of circular logic with respect to concept restrictions.

3.5. Conversion of NCI thesaurus from ontylog to OWL

We publish the NCI Thesaurus in OWL among other formats in order to make it more widely available. The basic entities in Ontylog that we need to represent in OWL are concepts, roles, and properties. During the analysis phase we found that all the semantics and features of Ontylog that were in use by us could be represented in OWL without resorting to OWL Full, a major consideration because we wished the Thesaurus to be classifiable and OWL Full offers no computational guarantees. (See the OWL Language Guide, http://www.w3.org/TR/2004/REC-owl-guide-20040210/#OwlVarieties). In addition, it was desirable that the process be fully automated in order to minimize the impact to our production cycle. The current conversion results in an OWL Lite version of the Thesaurus starting from its Ontylog XML representation. The mapping from Ontylog to OWL XML elements is detailed in Table 7.

Properties in Ontylog are annotations assigned to a concept and are not inherited by subconcepts. The equivalent entity in OWL is the AnnotationProperty; this conversion is straightforward. AnnotationProperty is also utilized to store other Ontylog XML elements such as code and id which are specific to a concept (Table 7).

As detailed elsewhere, roles are binary relations between concepts and are utilized by the Ontylog classifier to establish class membership by subsumption. Roles are best represented in OWL as ObjectProperty, and role expressions are thus translated as restrictions on OWL properties. Because roles are defined in Ontylog with domain and range kinds, a fourth Ontylog entity, we include kinds in the OWL version as this facilitates the automated processing of the Thesaurus (in the conversion of roles to owl:ObjectProperty, see below). The some (existential) and all (universal) logical modifiers in role expressions are converted to the value constraints owl:someValuesFrom or owl:allValuesFrom (respectively) as the semantics are the same. The conversion of the Ontylog poss modifier has not yet been addressed.

Concepts can be represented as OWL classes or instances but the notion of an instance does not exist in

[7] In Ontylog, roles are unidirectional, so we use one-headed arrows. In working with the more expressive logics, such as SHIQ(D) logic of RACER, for example, we would use bidirectional arrows. Arrows take the color of their domain kind, and the arrowhead touches their range kind.

Table 7
Ontylog XML element conversions

Ontylog element	Owl element	Comment
kindDef	owl:Class	
roleDef	owl:ObjectProperty	
propertyDef	owl:AnnotationProperty	
conceptDef	owl:Class	
name	rdf:ID	Applies to the name subelement of *kindDef*, *roleDef*, *propertyDef*, and *conceptDef*. The names are modified to conform to xml ncname rules
name	rdfs:label	Because the *conceptDef name* contains some useful semantics, the original form is retained as an *rdfs:label*. No other name elements are retained in *rdfs:label*.
code	owl:AnnotationProperty	Defined as an *owl:AnnotationProperty* with *rdf:ID* = "code". Code values remain the same for each concept.
id	owl:AnnotationProperty	Defined as an *owl:AnnotationProperty* with *rdf:ID* = "ID". ID values remain the same for each concept.
definingConcepts	rdfs:subClassOf	The *concept* sub-element of *definingConcepts* is mapped to the *rdf:resource* attribute of the *rdfs:subClassOf* element.
domain	rdfs:domain	
range	rdfs:range	
definingRoles	owl:Restriction owl:onProperty owl:someValuesFrom OR owl:allValuesFrom	*definingRoles* are converted to owl restrictions on properties. The *name* child element of *definingRoles/role* is taken as the *rdf:resource* attribute of the *owl:onProperty* element. The *value* child element of *definingRoles/role* is taken as the *rdf:resource* attribute of the *owl:someValuesFrom* or *owl:allValuesFrom* element

Ontylog. As detailed elsewhere, we do not view Thesaurus concepts as instances, thus all concepts are represented as classes in OWL. Consequently, the concept hierarchy in the Thesaurus is retained in the OWL class hierarchy. There is one difference between the Thesaurus hierarchies in OWL and Ontylog; it has to do with kinds. Kinds can be thought of as types of things and can be represented in OWL as disjoint classes; every Ontylog concept has a kind. In OWL, we chose to represent kinds as root classes in the various domain taxonomies; the root concepts of the Thesaurus in its Ontylog format are subclassed from the appropriate kind class in OWL. The OWL hierarchy therefore has an additional level at the top, the "kind" root class. When modeling the Thesaurus in OWL, it is important to note that unique names do not imply unique concepts. As far as OWL is concerned a concept with a specified kind could also be a member of another kind. This is corrected by specifying all of the classes that represent kinds as disjoint, thus recovering the uniqueness of kinds. To date, however, Ontylog Kinds have not been declared as disjoint.

Including Ontylog Kinds as classes in OWL is not strictly necessary. However, roles are defined in Ontylog with domain and range kinds, so it is not straightforward to automatically convert role domains and ranges to ObjectProperty domains and ranges if no equivalent classes exist for kinds in the OWL Thesaurus. This is because not all the domains in the Thesaurus are represented by a single taxonomy of concepts. For instance, in the Ontylog Thesaurus the kind NCI_Kind is partitioned in two taxonomies with the root nodes *Conceptual_Entities* and *NCI_Administrative_Concepts*: in the absence of kind classes in the OWL Thesaurus, manual intervention would be required to assign an appropriate class for the domain or range of an ObjectProperty. If all the Ontylog Kinds contained only one taxonomy, kinds could be dispensed with in the OWL version.

As a final note, the notion of a concept being primitive versus defined is not being currently addressed, except that primitive Ontylog concepts contain an AnnotationProperty denoting this status.

4. Discussion

Collaboration with users of the NCI Thesaurus has suggested several points that probably are relevant to those working on ontologies across biomedicine.

First, good kind definitions are critical for domain experts to comprehend readily the ontologist's intentions regarding the structure of intensional knowledge and to see how the intensional knowledge can be used to answer questions so as to address Use Case problems.

Second, we have learned several things about defining kinds and believe they apply to disjoint classes in general:

- Definitions should take the form of simple declarative sentences. It is sometimes necessary to enumerate that certain subclasses are explicitly included or excluded. It is worth considerable effort to establish definitions that eliminate or reduce to a bare minimum the need for such enumerations.[8] The Thesaurus' kind definitions are available at ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/KindDefinitions.pdf.

[8] In Table 5, the explicit exclusion of exogneous chemicals in the definition of GeneProduct_Kind is an example of enumeration, as is the explicit inclusion of translocation, polymorphism, and under- and over-expression in the definition of the Molecular_Abnormality_Kind.

- Complicated kind definitions are likely to be misunderstood by modelers and users, resulting in disagreements about the kind to which a concept should belong.
- It is better initially to create more kinds than may ultimately be necessary, each covering a small semantic space. Kinds are easy to combine, but hard to split apart once they have been put into the T-Box and used to produce published versions of the ontology.
- Careful attention to the definition of the kinds has two immediate pay-offs for the ontology developer: in many cases good kind definitions recapitulate good hierarchy partitions; and definitions frequently suggest top-level hierarchy structure.

Third, what domain experts want to know about any biomedical ontology is the range of questions that they can rely on the ontology to answer. The number and diversity of the concepts in the ontology are relevant to the questions to which it can provide responses, but the number and form of the semantic relationships that the ontology asserts among concepts is the most important determinant. In biomedical research, interactions among entities have become the key data points that are used in hypothesis formation and results interpretation.

Fourth, ontology development is costly. The NCI Thesaurus is new, having been released initially for internal NCI use in 2001. Initially we included in the Thesaurus concepts that we simply believed to be important to NCI. However given the costs of ontology development one cannot go on doing that very long. Since no ontology is ever complete [27], a yardstick is needed to help decide where in the namespace to create fine granularity concept hierarchies and richly detailed inter-concept semantics on expressed user needs. We now require that all T-Box elements be required either explicitly or by implication by the needs of at least one user community.

Table 8 illustrates several of these points. The draft Use Case in Table 8 was developed after only two sessions between the EVS and Cancer Therapy Evaluation Program (CTEP) staff. Although CTEP is a part of NCI, the staff involved had no prior familiarity with EVS or Thesaurus. However they were able to easily grasp the meaning of the Diseases_and_Disorders, Gene_Products, and Molecular_Abnormality Kinds, the role relations among them, and the way in which these roles could be used to answer the questions that users of their software application would need to address. Good kind definitions and well organized hierarchies within the kinds, and roles that reflect cancer researchers' views of the important semantic relationships among biomedical entities made it possible for the CTEP and EVS staff to quickly convert the CTEP user's need for answers to questions into specifications for which roles to use for what, areas where additional content was needed within existing kinds, and plans for mutual support and collaboration. It is also worth noting that in this Use Case we decided to create a Pathway_Kind. Currently pathways are poorly defined conceptual entities and pathway naming is not standardized, and there isn't a good principle for organizing pathways into hierarchies. Nonetheless it was clear from the Use Case that molecular abnormalities and gene products will need to be related to pathways. Rather than wait until these issues were resolved, we felt that we could leverage KEGG http://www.genome.ad.jp/kegg/ and BioCarta http://www.biocarta.com/genes/index.asp to provide CTEP and other NCI Thesaurus users with some useful, albeit very tentative, representations of the known semantic relationships between pathways and other biomedical entities. Keeping in mind the we can easily merge kinds, we felt that the probability that in the long term our early work would need to be redone or reorganized, should not keep us from doing what we could today.

Fifth, ontologies or terminologies in science must be able to manage changing content. There are several facets to this issue. Changing concepts can be tracked with techniques like concept-level history [3]. The pseudo-T-Box artifacts we have described in this paper can also help to address another aspect of change management, by helping ontology users to understand what has changed and why. Further, the role and kind to Use Case:Problem Mapping table (Table 4) is useful for documenting not only which user community requested a T-Box feature, but also for documenting the point in time that the T-Box model exhibited the requested feature. Whenever changes to the underlying NCI Thesaurus model occur, a new set of pseudo T-Box artifacts are included on the FTP site in the documentation for the release, specifying all changes to roles and kinds.

Finally, developers of applications that utilize ontologies should be encouraged to develop suites of regression tests to validate specific data that their applications utilize. A change to the ontology by an editor might be considered relatively minor by the editor and be acceptable after peer review, yet it could significantly impact a dependent application over which the ontology builder has no control.

Looking to the future, there are several issues that we must address. First, we are aware that Ontylog does not support an important user requirement, role attribution. Workers in cancer research are demanding about the justification for assertions about causality, correlation or any relationship between data upon which they make decisions or base conclusions. For example, consider the assertion:

Bladder Cancer Specific Nuclear Matrix Protein BLCA-4
Gene_Product_is_Biomarker_Type Tumor Marker

Table 8
Example Use Case

DRAFT Use Case 11: disease concept relationship to molecular targets	Notes
Case	
CTEP/CIB Vocabulary Support	
Background	
When planning trials or designing/reviewing protocols, CTEP staff must decide if patient eligibility must depend on molecular targets. (Eligibility may be limited to those who over (under) express the target.)	
Problem	
To design a protocol CTEP CIB staff must learn which targets are relevant to a specific disease	
Solution	
EVS will include in NCI Thesaurus concepts to support this requirement, including:	
• a comprehensive taxonomy of oncologic diseases and related conditions,	11/1 Diseases_and_Disorders_Kind
• a comprehensive taxonomy of gene products, and	11/2 Gene_Products_Kind
• a comprehensive taxonomy of molecular abnormalities associated with oncologic diseases and disorders, and	11/3 Molecular_Abnormality_Kind
• semantic relationships relating oncologic diseases to the molecular abnormalities and gene products known to be relevant to the etiology, progression, treatment or other aspects of the disease	*Roles* Disease_Has_Cytogenetic_Abnormality Disease_Has_Molecular_Abnormality Disease_May_Have_Molecular_Abnormality Gene_Product_Is_Pathway_Element Gene_Product_Has_Malfunction_Type Gene_Product_Malfunction_Associated_With_Disease
Actions	
CTEP and other collaborators will provide the additional terminology needed in the Thesaurus and will consult on its inclusion, especially on semantic relationships among molecular disorders, gene products and diseases EVS editors will represent these relationships in the Thesaurus so that the application can retrieve molecular targets related to diseases EVS editors will continually review literature for mentions of molecular targets, especially when co-occurring with oncologic disease	

This draft Use Case is still under development. One of the NCI component organizations, the Cancer Therapy Evaluation Program's Clinical Investigations Branch is developing a software application that will require vocabulary support. The Use Case defines the needs of the application for terminology support, the kinds and roles that EVS will use to provide them, and the actions to be undertaken by EVS and CTEP.

We would like to be able to place a restriction on reified role assertions such as this to indicate a citation or other justification for the asserted relationship between the protein and its utility as a specific type of biomarker. For example:

Bladder Cancer Specific Nuclear Matrix Protein BLCA-4
Gene_Product_is_Biomarker_Type Tumor Marker
Has_Evidence J Urol 2000. 164:634-639

Currently Ontylog DL does not support such restrictions on reified roles. However, as reified roles are really complex concepts (see discussion of defining superconcepts, above), extension of the language definition to encompass restrictions on reified roles is possible.

Other formalisms that our Use Cases suggest, but that Ontylog cannot currently represent, include cardinality and enumeration. In cancer biology enumeration in the sense of a one-of operator is especially important. Consider the example, of Leukemia, a disease characterized by the presence of primitive or atypical myeloid or lymphoid cells in the bone marrow and the blood. Ability to place a one-of restriction on such a concept would be useful.

Leukemia $\doteq$ Disease_Has_Abnormal_Cell
One_Of (Myloid Cell, Lymphoid_Cell)

While these needs could be met by using a DL such as the one implemented in RACER, that is not a completely satisfactory solution. Ontylog can express useful constructions that other DLs cannot, such as the poss role qualifier. (We will experiment with representing this Kripke-like qualifier in OWL.) In the rapidly advancing sciences covered by NCI Thesaurus, the need to express possibility, in the sense of a future world, is a recurring motif.

Right Identities also are unique to Ontylog. NCI has not used the Right Identity construction, but the College of American Pathologists has used it in SNOMED/RT [Spackman, personal communication]. Features such as Right Identity contribute to Ontylog's speed of classification. In our hands, classification of the NCI Thesaurus

using Ontylog takes under a minute, while it takes about 12 min using RACER. Rapid classification enables ontology modelers to reclassify as they edit. Given that no classifier offers very powerful debugging to help ontologists to locate the source of errors in the ontology, frequent reclassification has something to recommend it as a way to help find loops and other problems with an ontology before they cause serious problems.

5. Conclusion

In conclusion, the NCI Thesaurus is a large, deployed and evolving biomedical vocabulary that intersects with multiple research domains. Such vocabularies require collaborating with subject matter experts to make them useful to various groups of end users. Ontylog DL is well suited for constructing big biomedical vocabularies and the Ontylog constructs kind and role can be and have been used to facilitate this collaboration. The collaboration process described in this paper is being used successfully, as validated by collaboration with and use by communities of scientists such as those associated with the Cancer Bioinformatics Grid and by adoption of NCI Thesaurus by groups such as one of the US government's Consolidated Health Initiative's government-wide standard terminologies. The artifacts and processes developed by NCI for collaboration may be useful in other domains.

Acknowledgments

The authors thank other members of the EVS team, notably, Margaret Haber, Larry Wright and Nick Sioutos, as well as Peter Covitz, Jim Oberthaler, Mark Tuttle, Tony Weida, John Carter, and Mark Musen for offering valuable comments, Jason Weis who implemented role groups, Eric Mays for many years of collaboration on the Ontylog Classifier and Jim Hendler and Bijan Parsia for their help in developing the OWL converter.

References

[1] Nardi D, Brachman J. An introduction to description logics. In: Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schnider P, editors. The description logic handbook. Cambridge: Cambridge Press; 2003. p. 1–40.

[2] de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW: NCI Thesaurus: using science based terminology to integrate cancer research results. Proceeding of the 11th World Congress on Medical Informatics (Medinfo 2004) 2004, P. 33–7.

[3] Hartel FW, Fragoso G, Ong K, Dionne R. Enhancing quality of retrieval through concept edit history. In: Musen M, editor. Biomedical and health informatics: from foundations to applications. Proceedings/AMIA; 2003.

[4] Covitz P, Hartel FW, Schaefer C, de Coronado S, Fragoso G, Sahni H, et al. caCORE: A common infrastructure for cancer informatics. Bioinformatics 2003;19:2402–12.

[5] Golbeck J, Fragoso G, Hartel FW, Hendler J, Oberthaler J, Parsia B. The National Cancer Institute's Thesaurus and Ontology. J Web Semantics 2003;1:75–80.

[6] Nelson SJ, Tuttle MS, Cole WG, Sherertz D, Sperzel WD, Erlbaum MS, et al. From meaning to term: semantic locality in the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care 1991:209–13.

[7] Baader F, Nutt W. Basic description logics. In: Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schnider P, editors. The description logic handbook. Cambridge: Cambridge University Press; 2003. p. 43–95.

[8] Hartel FW, de Coronado S. Information Standards within NCI. In: Silva JS, Ball MJ, Chute CG, Douglas JV, Langlotz C, Niland J, Scherlis W, editors. Cancer informatics: essential technologies for clinical trials. New York: Springer; 2002. p. 135–56.

[9] Minsky M. A framework for representing knowledge, MIT-AI Laboratory Memo 306, June, 1974. Reprinted from P. Winston (ed.), The psychology of computer vision. NY, McGraw-Hill, 1975.

[10] Bobrow DG, Winograd T. An overview of KRL, a knowledge representation language. Cogn Sci 1977;1(1):3–46.

[11] Baader F, Hollunder H. KRIS: knowledge representation and inference system. SIGART Bull 1991;3(2):8–14.

[12] Brachman RJ, McGuinness DL, Patel-Schneider P, Alperin Resnick L, Borgida A. Living with CLASSIC: when and how to use a KL-ONE-like language. In: Sowa, editor. Principles of semantic networks. Los Altos, CA: Morgan Kaufmann; 1991.

[13] MacGregor R. Inside the LOOM description classifier. SIGART Bull 1991;3(2):88–92.

[14] Mays E, Weida R, Dionne R, Laker M, White B, Liang C, et al. Scalable and expressive medical terminologies. Proceedings/AMIA Annu Fall Symp 1996:259–63.

[15] McGuinness DL. Ontologies come of age. In: Fensel D, Hendler J, Lieberman H, Wahlster W, editors. The semantic web: why, what, and how. Cambridge: MIT Press; 2001. Available from: http://www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-abstract.html.

[16] Rector A, Rogers J, Roberts A, Wroe C. Scale and context issues in ontologies to link health and bio-informatics, Proceedings/AMIA Fall Symp, 2002; p. 624. Available from: http://www.cs.man.ac.uk/mig/rector@cs.man.ac.uk.

[17] Mays E, Dionne R, Weida R. K-Rep system overview. SIGART Bull 1991;3(2):93–7.

[18] Baader F. Appendix: description logic terminology. In: Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schnider P, editors. The description logic handbook. Cambridge: Cambridge Press; 2003. p. 485–95.

[19] Rector A. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. K-CAP '03, October 23–25, 2003 Sanibel Island, FL, p. 121–8.

[20] Baader F, Küsters R. Computing the least common subsumer and the most specific concept in the presence of cyclic ALN-concept descriptions. LTCS-Report 98-06, LuFg Theoretical Computer Science, RWTH Aachen, Germany; 1998.

[21] Personal communications with Drs. Fedack, Rose, Campbell and Spackman at SNOMED modelers meetings.

[22] McGuiness D, van Harmelen F, editors. Available from: http://www.w3.org/TR/owl-features/.

[23] Goldblatt R. Set concepts and validity in Topoi: the categorial analysis of logic. Amsterdam: Elsevier; 1984. p. 212.

[24] Scott D. Advice on modal logic. In: Lambert K, editor. Philosophical problems in logic. Dordrecht: Reidel; 1970. p. 143–73.

[25] Spackman K, Dionne R, Mays E, Weis J. Role Grouping as an extension to the description logic of ontylog, mobivated by concept modeling in SNOMED, Proceedings/AMIA Fall Symp; 2002.

[26] Sanner S. Towards practical taxonomic classification for description logics on the semantic web. Technical Report, Stanford University Knowledge Systems Lab KSL-03-06; 2003.

[27] Baader F, Nutt W. Basic description logics. In: Baader F, Calvanese D, McGuinness D, Nardi D, Patel P-Schnider, editors. The description logic handbook. Cambridge: Cambridge University Press; 2003. p. 43–95.