NCI Thesaurus Semantics

The NCI Thesaurus is built using the Ontylog dialect of description logic (DL). The semantics of Ontylog (effective May 2004) are summarized in brief symbolic form at ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/OntylogSemantics.pdf. Ontylog is fairly widely used in biomedical terminology construction. Notably SNOMED/RT and SNOMED/CT are based on it, as is the US Veterans Administration National Drug File Reference Terminology (NDF-RT).

The NCI Thesaurus employs all the semantic constructions offered by Ontylog ***except*** modal restriction and right identity. The *concept* is the fundamental notion in Ontylog. In Ontylog concepts are abstract classes: there is no notion if an instance. Each concept denotes a semantic unit of meaning. Concepts are placed into is_a hierarchies, checked and potentially altered through a process of classification. Following the usual practice of description logics, classification is accomplished by subsumption testing across the acyclic graph structure in which each concept is a node. The graph's edges are semantic relationships among the concepts.

NCI employs the Apelon, Inc. Terminology Development Environment (TDE) to build the NCI Thesaurus, and the Apelon Distributed Terminology Server (DTS) to make the Thesaurus accessible to users via programming interface and Web browser. The TDE and DTS software products implement Ontylog DL in software. Because Ontylog is not widely supported by open free tools, we translate the Ontylog version of NCI Thesaurus into OWL, and then make either version available for download.

Each concept in NCI Thesaurus is either primitive (description limited to necessary conditions) or defined (description includes necessary and sufficient conditions). Subsumption of primitive concepts is established by the concepts' *defining super concepts*. Subsumption of defined concepts is established by its *direct super concepts*. Defining super concepts are manually determined, while direct super concepts are determined algorithmically during classification.

Originally all concepts in the NCI Thesaurus were primitive. As the terminology matures, the proportion of defined concepts has increased, but the bulk of the NCI Thesaurus remains composed of primitive concepts.. Top-level concepts will remain primitive as they express axiomatic knowledge used to infer the meaning of defined concepts, and assure that branching at the top of the hierarchy trees in the Thesaurus is well formed. Note that NCI Thesaurus supports concept *polyhierarchy*.

Many description logics make a distinction between generic concepts that describe sets and individual concepts 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. NCI Thesaurus does not support instances for two reasons. First, it is designed to be a large and complex vocabulary that will be employed by runtime systems supporting basic, translational and clinical research. In these systems instances are typically experimental data and research subject records, and are stored in databases where transaction semantics and other core system concerns are paramount. The kind of inferencing that is performed over the instances is readily performed over the generic concepts so there is no need for assertions about individuals in the Thesaurus. If Thesaurus supported exceptions, individuals might be required in the terminology. However, Ontylog has an enforced semantics, so there are no exceptions in the Thesaurus. Second, the distinction between generic and individual is considered a hard problem in mathematical foundations.

There are two types of semantic relationships among concepts in NCI Thesaurus:

1. *roles*: Roles are binary relationships between concept pairs that inherit.
2. *associations*: Associations are also binary relationships between concept pairs, but associations do not inherit.

The roles currently available for use in Thesaurus are listed in ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/Roles.pdf Note that not all roles have been instantiated in the Thesaurus. They are available, but some have not as yet have been used to restrict any concepts. Following DL convention, there is a domain and range value associated with each role. Ontylog uses the notion of a *kind* as values for domain and range. All concepts belong to one, and only one, kind. Kinds may be thought of as disjoint classes or as data types. A list of the kinds in the Thesaurus, and each kind's definition, is provided at ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/KindDefinitions.doc. A graphic is also available that represents the roles and their domains and ranges. We use a distinctly colored box for each kind. Arrows with role names represent roles[1]. Because many kinds and roles are required to satisfy the needs of cancer researchers, the graphic is large and visually complex. The current graphic is available in portable network graphic and Visio format at
ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/Tbox.vsd
We have adopted the practice of including the domain and range names in the name of each role (Disease_Has_Primary_Anatomic_Site has the domain Disease and the range Anatomy).

Associations do not have specific domain or range values. Associations no not define concepts. Semantically an association is simply a named relation between *specific* concepts. They are similar to Ontylog properties, which are discussed below. The distinction between Ontylog association and Ontylog property is that properties take strings as filler values, but concepts are the filler values of Associations. Associations are not inherited because the relationship might not hold for decedents. The associations currently available for use in Thesaurus are listed in ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/Associations.pdf

NCI Center for Bioinformatics relies on use cases to define the needs of our user communities for informatics resources. The NCI Thesaurus has adopted the discipline of listing the use case and individual need from within the use case to roles and kinds. The role and kind associations-to-use-cases enables us to track not only which community relies on each role and kind, but also the use that they make of it. The current list of use case to kind mappings ftp://ftp1.nci.nih.gov/pub/cacore/EVS/ThesaurusSemantics/RoleToUCmapping.pdf is provided as general information.

The NCI Thesaurus is designed, first and foremost, to be a *thesaurus* – "a controlled vocabulary arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators." NCI Thesaurus has some ontology-like features but NCI Thesaurus is not an ontology and is not designed or intended to one. Its primary role is that of a bridge for human to human communication across specialties and data resources. Much of this information is represented in what Ontylog calls *properties*. Properties are labeled values associated with a concept. The labeled values' filler values are strings and other literals. In most DLs, this sort of information would restrict individuals, not abstract classes as in Ontylog. Refer to the DTS User Guide for additional discussion.

[1] 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.