



The National Cancer Institute's Thésaurus and Ontology

Jennifer Golbeck^{a,*}, Gilberto Fragoso^b, Frank Hartel^b, Jim Hendler^a,
Jim Oberthaler^b, Bijan Parsia^a

^a Department of Computer Science, University of Maryland, A.V. Williams Building, College Park, MD 20742, USA

^b National Cancer Institute, Center for Bioinformatics, USA

Received 8 May 2003; received in revised form 23 July 2003; accepted 31 July 2003

1. Introduction

The NCI Thésaurus is a public domain description-logic-based terminology produced by the National Cancer Institute, and distributed as a component of the NCI Center for Bioinformatics caCORE distribution [1]. It is deep and complex compared to most broad clinical vocabularies, implementing rich semantic interrelationships between the nodes of its taxonomies. The semantic relationships in the Thésaurus are intended to facilitate translational research and to support the bioinformatics infrastructure of the Institute.

The NCI Thésaurus evolved from the NCI Metathésaurus. NCI Metathésaurus is based on the National Library of Medicine Unified Medical Language System (UMLS) Metathésaurus. The NCI Metathésaurus has been operational since 1999. A public version is available at <http://ncimeta.nci.nih.gov>.

The need for a comprehensive NCI-wide terminology arose because NCI staff requires access to

timely and accurate information about activities related to the scientific mission of the Institute. The collection, storage and retrieval of data related to NCI research programs is necessary to analyze, manage, and report about these activities. Though centralized coding of NCI-supported research-related activities met some of these needs, supplementary data coding had become common. This coding was assigned independently within various components of the Institute, and was frequently based on locally developed term lists or other informal vocabulary, making it difficult to find and combine information across programs.

The NCI source vocabulary within the NCI Metathésaurus encompasses the terminology used by the various offices and divisions within the Institute, with the goal of providing a common vocabulary to increase the interoperability of information systems. The NCI vocabulary provided not only an initial Institute-wide integrated vocabulary, but also rich mappings of NCI terminology to the numerous other biomedical vocabularies.

Useful as it is, however, the NCI Metathésaurus is not well suited to serve as a coding nomenclature.¹ The hierarchies of the NCI terminology in NCI

* Corresponding author. Tel.: +1-301-405-2662; fax: +1-301-405-6707.

E-mail addresses: golbeck@cs.umd.edu (J. Golbeck), fragosog@mail.nih.gov (G. Fragoso), hartel@mail.nih.gov (F. Hartel), hendler@cs.umd.edu (J. Hendler), oberthaj@mail.nih.gov (J. Oberthaler), bparsia@isr.umd.edu (B. Parsia).

¹ We define a *nomenclature* as addressing the symbol to object relation in the semiotic triangle.

Metathesaurus are not true IS_A hierarchies, for example. The NCI Thésaurus, which became operational in late 2001, is designed to address this need. It is intended for NCI offices and divisions to use the Thésaurus as a source of codes associated with terminology concepts to annotate data and other information artifacts and facilitate information retrieval. The NCI Thésaurus is not yet a true ontology,² as it contains numerous primitive concepts. It might be fairly described as a nomenclature with ontologic features. It is strongly ontology-like across several taxonomies, however, containing detailed semantic relationships among genes, diseases, drugs and chemicals, anatomy, organisms, and proteins. Among these categories, thousands of concepts are defined. Each concept's definition contains descriptive information, such as synonyms and English definitions, and explanations of how it relates to other concepts.

The Thésaurus has been expanded to support bioinformatics-related research efforts, and in this role, it is a source of information on asserted relationships between entities. With reference to a concept representing an animal model of a human disease, for example, the Thésaurus might describe relationships such as genetic abnormality is associated with the disease and the organisms in which it occurs. Although editors have defined a number of named ontologic relations to support the description–logic-based structure of the Thésaurus, additional relationships are considered for inclusion as required to support-dependent applications.

The detailed information provided about such a broad range of medical terms and topics is a rich and useful data source for the medical community. In an effort to make the knowledge in the Thésaurus more useful and accessible to the public, the National Cancer Institute and the University of Maryland, College Park have worked together to produce an OWL [2] ontology from the Thésaurus. In the following sections, this paper will describe the terminology development process at NCI, and the issues associated with converting a description–logic-based nomenclature to a semantically rich OWL ontology.

2. Ontology development

Given the rapid pace of research findings and clinical refinement related to cancer prevention and treatment, the content of the NCI Thésaurus evolves rapidly. To keep the information as up to date as possible, a new release is made every month. With nearly a dozen full time subject matter experts working on different parts of the terminology, a well-structured and complex process is used to ensure accuracy and coherence. NCI uses the Apelon Inc.'s Terminology Development Environment and Workflow Manager software tools [3] in the development process.

There are eight main steps in terminology development. The full process is diagrammed in Fig. 1, and the steps described below reference the numbered arcs in the figure. At the beginning of each cycle, the Lead Modeler, working from a local database with the current version of the Thésaurus creates a series of worklists which detail the concepts that need to be edited by each modeler (1). The worklists are exported through the Apelon Workflow Manager and assignments are made to each Modeler or Expert Modeler (2).

Each modeler has a local copy of the NCI Thésaurus. This prevents problems where one modeler may make a change that impacts changes made by another. Once each modeler makes changes to their local copies of the nomenclature (3), a Change Set that captures the editing performed in their local copy is sent back to the Workflow Manager (4). The changes are analyzed to identify any conflicts that may have arisen in the process. If any corrections are needed, new assignments are made for the expert modelers, and the cycle repeats until no conflicts are found (5). On a weekly basis, new baselines containing the consolidated changes are released to the modelers to update their local databases (6).

To make a full export of the new nomenclature available to the public on the web, several steps are still required. All of the changes are imported into a trial database and classified using description logic rules. If there are any unexpected results, problematic changes are either eliminated or corrected. Once classification is successful, comparisons and reviews are made between the new classified nomenclature and the previous version. In addition, the release candidate is also made available to developers of depen-

² We define an *ontology* as addressing the symbol to concept relation in the semiotic triangle.

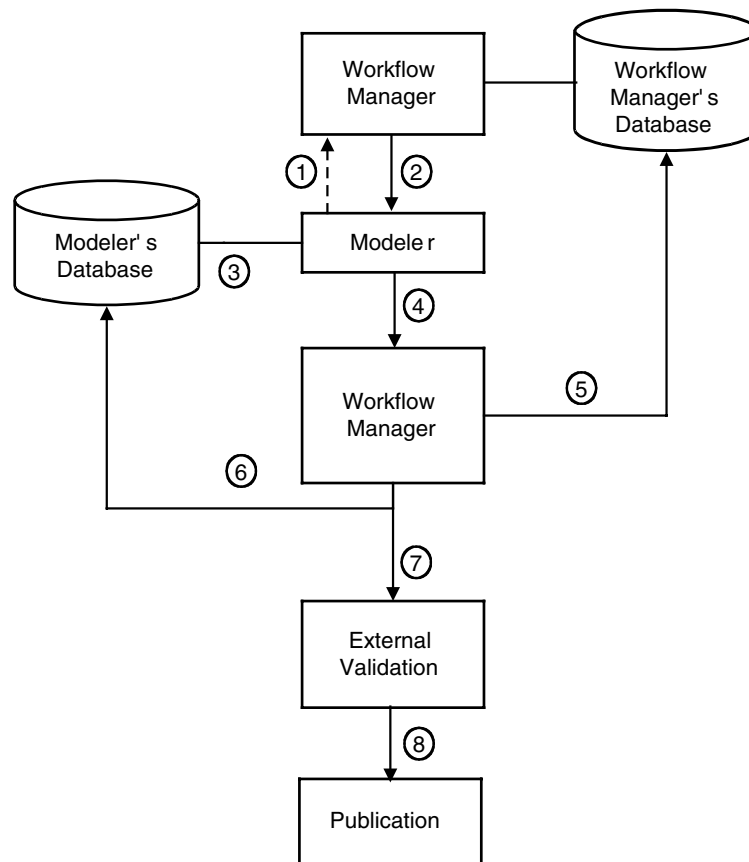


Fig. 1. Workflow diagram of the NCI Thesaurus editing and publication cycle. This is an abbreviated version of the editing workflow process. The numbers in the figure correspond to the numbers in the text above that describes the process.

dent projects for regression testing (7). When the final version is approved, it is published on the web (8) at <http://nciterns.nci.nih.gov>. As part of the caCORE distribution, it is periodically also made available in both a flat file and XML formats. The flat file, however, does not include all the elements of the nomenclature, such as role relations.³

The February 2003 release of NCI Thesaurus contained about 26,000 concepts and about 71,000 terms divided among 24 taxonomies. The taxonomies cover administrative, applied and basic science and clinical terminology.

3. Conversion to OWL

Converting the NCI Thesaurus to an ontology in OWL Lite is a multi-step process. At the end of the production cycle, the Thesaurus is exported in Apelon's Ontylog XML format. Though the semantics intended by the modelers were included in the XML, and recoverable using the Apelon server, the complex relationships are not apparent to any third-party software.

In creating an ontology that represented these semantics, the first choice was to decide among OWL's three sub-languages: OWL Lite, OWL DL, and OWL Full [2]. Though OWL Full is the most expressive of the three languages, the first two have several advantages. OWL DL and its formal subset, OWL Lite, are

³ These files can be downloaded from <ftp://ftp1.nci.nih.gov/pub/cacore/EVS/>.

guaranteed to be both computationally complete and decidable. The NCI Thésaurus is created using a DL tool, so the restrictions of these two variants of the language were a reasonable fit. Since the elements required to express the relationships in the Thésaurus fell within the restrictions of OWL Lite, we chose to encode in that sub-language.

One of the most important steps in transforming the Thésaurus into an ontology is to represent the concepts and their connections in a machine processable way. In our ontology, each concept is given a formal designator and the relationships between them are formalized in the base ontology language. This overcomes any ambiguity between natural language-based descriptive text, and formal concept names. In the NCI Thésaurus, names of entities are semantically rich. Some are simple one-word names, while others are long complex descriptions, e.g. “Anti-human chorionic gonadotropin vaccine/Gemcitabine” and “Transporter 1, ATP-Binding Cassette, Sub-Family B (MDR/TAP)”. To do this, the original concept names

were converted to proper RDF identifiers by removing any spaces and substituting illegal characters with underscores. Names that began with numbers were also prefixed with underscores to make them legal names. The names in their original forms are preserved as `rdfs:label`.

The four major types of definitions made in the Thésaurus are Kinds, Roles, Properties, and Concepts. Kinds are the top-level super classes for all of the concepts defined in the Thésaurus, and basically enumerate the different categories of concepts. Kinds include such things as Anatomy, Biological Processes, Chemicals and Drugs, and Diagnostic and Prognostic Factors. Each Kind is converted to an `owl:Class`.

A Concept also translates to an `owl:Class`, as it describes a specific concept under one of the Kind categories. The bulk of the Thésaurus comprises concept definitions, and this is also where the most complex semantics occur. As such, it was the most difficult part of the conversion. Consider the following abridged concept definition from the Thésaurus XML:

```
<conceptDef>
  <name>Chlorambucil</name>
  <code>C362</code>
  <id>362</id>
  <primitive/>
  <kind>Chemicals_and_Drugs_Kind</kind>

  <definingConcepts>
    <concept>
      Nitrogen Mustard Compound
    </concept>
  </definingConcepts>

  <definingRoles>
    <role>
      <some/>
      <name>
        Chemical_or_Drug_FDA_Approved_for_Disease
      </name>
      <value>
        Chronic Lymphocytic Leukemia
      </value>
    </role>
  </definingRoles>

  <properties>
    <property>
      <name>FULL_SYN</name>
      <value>
        <![CDATA [
          <term-name>
            phenylbutyric acid nitrogen
            mustard
          </term-name>
          <term-group>SY</term-group>
          <term-source>NCI</term-source>]]>
      </value>
    </property>

    <property>
      <name>Semantic_Type</name>
      <value>
        Pharmacologic Substance</value>
      </property>

    ...
  </properties>
</conceptDef>
```

As shown above, each Concept in the Thésaurus has three main types of associated data: defining concepts, defining roles, and properties. A “defining concept” is essentially a super class, so we used `rdfs:subClassOf` relationships here.

Roles translate to `rdf:Properties` in OWL. Roles describe how concepts relate to one. Generally, Roles have domains and ranges of Kinds, and the values are further limited to specific Concepts within concept definitions. The “Defining roles” within a concept definition provide these local restrictions on the ranges of roles. For example, the class “Chlorambucil”, a drug, has the defining role:

```
<Chemical_or_Drug_FDA_Approved_for_Disease>
    Chronic Lymphocytic Leukemia
</Chemical_or_Drug_FDA_Approved_for_Disease>
```

There is a class “Chronic.Lymphocytic.Leukemia” defined in the ontology, as well as the role property “Chemical_or_Drug_FDA_Approved_for_Disease.” In the OWL output, there is a local restriction on the range of that role to have `owl:someValuesFrom` the `Chronic.Lymphocytic.Leukemia` class.

```
<owl:Restriction>
    <owl:onProperty rdf:resource="#rChemical_or_Drug_FDA_Approved_for_Disease"/>
    <owl:someValuesFrom rdf:resource="#Chronic.Lymphocytic.Leukemia"/>
</owl:Restriction>
```

Properties in the Thésaurus, on the other hand, contain metadata that describes the class, but not its instantiations or subclasses. Properties are defined as `owl:AnnotationProperties` in the ontology. An `owl:AnnotationProperty` is a subclass of an `rdf:Property`, and, like `rdfs:comment` and `rdfs:label`,⁴ can be attached to any class, property or instance. This allows Properties from the Thésaurus to be associated directly to a Concept’s corresponding class, without violating the rules of OWL Lite.

In addition to explicitly named Properties, “code” and “id” attributes are special cases. Every entity

in the Thésaurus—Concepts, Kinds, Properties, and Roles—has an associated “code” and “id.” These are used as unique identifiers in the Apelon development software, and, as such, are not defined explicitly as Roles or Properties. This required a hard-coded definition of both entities. They were defined as `owl:AnnotationProperties`, just like other Properties in the Thésaurus, and this allowed them to be attached to all of the definitions in the ontology.

The example code above shows two Properties in addition to “code” and “id”. Each Property is defined to have a name and a value. The names of the properties correspond to actual property definitions included

elsewhere in the files. Values needed to be attached, via the Properties, to the new class. Since the Properties are `owl:AnnotationProperties`, this is done in the same way a property would be associated with an instance.

The final OWL ontology is made up of approximately 450,000 triples in a file that is over 33 MB.

A description and current OWL version is available for download at <http://www.mindswap.org/2003/CancerOntology>. Currently, we are exploring the inclusion of the OWL properties for controlling versions (including identification of deprecated classes and properties). When this is completed, publication of the Thésaurus to OWL will be incorporated into the NCI monthly production cycle.

4. Lessons

Countless organizations have their own intricate and well-developed thesauri that have been refined over many years. With consummate converters, these thesauri can become a strong and useful element on the semantic web. Like the Thésaurus from the National

⁴ `rdfs:comment` and `rdfs:label` are actually two pre-defined annotation properties in OWL. There are three others: `owl:versionInfo`, `rdfs:seeAlso`, and `rdfs:isDefinedBy`. More details are available in the OWL Reference at <http://www.w3.org/TR/owl-ref/>.

Cancer Institute, many are available in an unusual form of XML, or not in XML at all. As in this case, simple translations using standard XML parsers often will not be a reasonable solution for making the transition to RDF or OWL.

The most time consuming process in this conversion was a making careful analysis of the Thésaurus to understand the best way to translate it into OWL. In the terminology development process, Concepts are not strictly considered to be instances or classes. However, because of the hierarchical structure within the collection of Concepts, as well as the Roles that serve as local restrictions, they clearly translated best to classes in OWL.

Discovering the distinction between Roles (local restrictions on properties), and Properties (annotations to a specific class) was another important step. At this point in the development, we were forced to make a choice in which species of OWL to use. Developers attached properties that are not inherited to

concepts. If the translator directly attached Properties to classes, the ontology would be in OWL Full. For computational benefits, we decided to work in OWL Lite, and thus owl:AnnotationProperties were used.

For other conversions, these same types of distinctions and decisions must be made. The expressive power of a proprietary encoding can vary widely from that in OWL or RDF. Understanding the original semantics and engineering a solution that most closely duplicates it is critical for creating a useful and accurate ontology.

References

- [1] NCI Center for Bioinformatics caCORE : <http://ncicb.nci.nih.gov/NCICB/core>.
- [2] OWL Web Ontology Language: <http://www.w3.org/2001/sw/WebOnt>.
- [3] Apelon Inc.: <http://apelon.com>.