Enterprise Vocabulary Development in Protege/OWL: Workflow and Concept History Requirements

Sherri de Coronado, Gilberto Fragoso NCI Center for Bioinformatics

The National Cancer Institute has developed the NCI Thesaurus, a biomedical vocabulary that provides consistent, unambiguous codes and definitions for concepts used in cancer research. It currently has about 34,000 concepts, in 20 hierarchies. A lexical component provides human usable definition and other information. Description logic is used to ensure complete, consistent and non-redundant definitions in the formal sense. NCI Thesaurus enables retrieval of information across a wide range of domains related to cancer research, with one goal being facilitating translational research. This vocabulary was originally built using Ontylog, a description logic developed explicitly for building large complex terminologies by Apelon, and implemented in the Apelon Terminology Development Environment (TDE). We have converted the Ontylog version of NCI Thesaurus to OWL, and are beginning to test the features of Protege OWL in a multi-user environment with a database backend to determine whether it could be used to support the operational needs of NCI Thesaurus development and maintenance.

NCI Thesaurus Conversion to OWL: To make NCI Thesaurus more widely available, we publish the NCI Thesaurus in OWL as one of our release formats. Our current conversion script generates the NCI Thesaurus in OWL Lite. Kinds and Concepts in Ontylog were modeled as classes in OWL, with a subclass relationships created between Kinds and the corresponding root Concepts in each Kind. Kinds are disjoint sets of concepts; however, we are not yet expressing this disjoint character in our OWL version. Roles in Ontylog are binary relations between concepts and are utilized by the classifier to determine class membership/subsumption between concepts in the taxonomies; these were best translated to OWL as restrictions on properties. Roles in Ontylog were therefore converted to ObjectProperty with domains and ranges; thus the need to create Kinds as classes in OWL during an automated conversion since in Ontylog domains/ranges are defined in terms of Kinds. The “some” role qualifier in Ontylog was expressed as someValuesFrom in OWL. Properties in Ontylog, which provide additional information about a concept that are not used in classification were converted to AnnotationProperties in OWL. There are no instances in Ontylog, since it was designed primarily for medical applications; in most medical applications the instances are data in medical record systems, or other repositories of patient data. This OWL conversion file is the basis for our testing of Protégé/OWL for possible use by NCI in operations, and for exploring the benefits of enriching the NCI Thesaurus with other DL constructs not available in Ontylog.

NCI Thesaurus Operational Requirements:

The NCI Thesaurus currently provides vocabulary support to the NCICB bioinformatics infrastructure that includes objects and data standards as well as vocabulary, and is used directly and indirectly by a number of applications. It operates on a monthly publication schedule: editing of distinct schemas by multiple domain experts is consolidated in one database, published to a test database and then promoted to production. The vocabulary is edited full time by a group of editors primarily at NCI, but also distributed around the country. Currently, each editor has a separate schema that resides on a central server. Edits are made to the individual copies. Each editor performs classification, and then exports a change set. The Work Flow Manager combines the changes from the separate schemas with the assistance of a work flow tool, deals with conflicts, and produces a new baseline. The editing must be halted from export of the change sets until the new baseline is released.

Protégé/OWL as the Editing Envronment for NCI Thesaurus

We are evaluating whether we could use Protégé OWL as an editing environment. The major facets of the evaluation include the following:

1. Test our ability to use Protégé OWL to model, including identifying whether it enables us to adopt a more expressive semantics without loss of maintainability;
2. Adapt or create a new workflow process that enables multi-user editing in a production environment, and then,
3. Either export back into Ontylog for publication in the Apelon Distributed Terminology Server (DTS), or provide for another terminology server to support programmatic access.

More specifically, the evaluation tasks include the following

1. Can we use Protégé OWL to effectively model NCI Thesaurus?
    a. Can we represent our current model adequately in Thesaurus OWL? Preliminary work indicates that we can but we need to avoid some DL constructions that will be present in upcoming releases of Ontylog.
    b. Will OWL facilitate expanding our current model in ways that will enhance the utility of the NCI Thesaurus to our end users?
2. Is Protégé OWL suitable for a production editing environment for a large vocabulary?
    a. Performance: Speed of concept retrieval from the backend database must be acceptable. NCI Thesaurus is currently about 34,000 concepts (classes) and undergoes heavy editing. Classification also has to be reasonably fast;
    b. Workflow:Will the multi-user backend for Protégé 2.0 provide an adequate (or better) alternative for the workflow process we currently use that includes individual schemas, changesets and conflict resolution? We must create a means for conflict resolution and/or review of editor's work performed in the context of a single editable database. We believe that Prompt and PromptViz may facilitate viewing changes between baselines, and review of weekly edits to the database.
    c. Editing Add-ons: In our current operational environment, we have extended the Apelon terminology development environment (TDE) to provide for important functions not currently available in the Protégé OWL tool. These include: editing restrictions such as allowing only a single editor to retire concepts, pre-retirement actions such as moving child concepts (classes) to other locations in the vocabulary structure; splits of concepts; and concept history, so that when a concept is retired, split or merged, end users will be aware, and be able to use pointers to new concepts as appropriate. Further, in the Protege environment we would need to be able to programmatically generate read-only codes (as AnnotationProperty) for newly-created classes that can be used subsequently by dependent applications to code artifacts in external repositories. These various functions will have to be replicated via a Protégé plugin that incorporates the Protégé OWL plugin, or by changes to the base Protégé code if necessary and appropriate, via collaboration with the Protégé developers.
    d. Visualization of semantic relationships : Our current editing environment focuses on visualizing is-a relationships. However, we have created a richer semantic network that can only be fully understood with better visualization tools. Again, we have extended our current tools to provide some assistance in visualizing partonomies, and other relationships, but the Protégé plugins being built may provide ready-made solutions. We will be evaluating these as well.
    e. Can we export back from Protégé to DTS for serving the vocabulary until there is a better public domain vocabulary server? As the vocabulary server (DTS) does not expose a number of DL constructions, it might be possible to export only the entities that users need in a terminology, while domain experts utilize all the capabilities of OWL to model the Thesaurus.

Progress to Date:

Modeling: We can and have implemented NCI Thesaurus in OWL. However, the current conversion utilizes only "someValuesFrom" and not "allValuesFrom". In some cases, we would

normally use “all” and will have to test this shortly. For example, Lung Carcinoma All Disease_Has_Primary_Anatomic_Site Lung would be a valid statement. Further, all of our concepts in the OWL Thesaurus are currently primitive, when in reality, some of them should be defined, i.e. we have modeled necessary and sufficient conditions to define the concepts in the areas of genes and gene products in Ontylog. Our evaluation will include making these changes in Protégé OWL, and determining whether the results are expected and useful. Second, we also believe we have identified necessary and sufficient conditions for modeling a subset of our disease terminology and will test the validity of that hypothesis. Remaining issues to consider include: 1) The update to Ontylog that enables Associations or non-defining roles, these are useful extensions and we would be interested in an OWL representation; 2) The update to Ontylog also enables use of the "poss", meaning "possibly" role modifier. OWL does not include this, we anticipate that in some cases these could be converted to "some" while in others the information might be represented in an AnnotationProperty.

Second, we have some use cases where negation, enumeration or cardinality would be useful but have not yet tried to implement this or just as importantly, determined what an end user would need to do differently to utilize the added semantics. If we test those features, we might have to rollback in order to export to Ontylog prior to serving the vocabulary.

Production Editing and Workflow: Our current limited tests indicate that the query time for a concept using the backend database is not fast enough, on the order of seconds. Classification also has to be reasonably fast – it currently takes about 10 minutes using Racer vs under one minute for the NCI Thesaurus with the Ontylog classifier.

We have begun working on editing add-ons via an NCIOWLClassTab that would enable us to adapt Protégé for workflow. For instance, we have started on an advanced search tool, and have identified an approach to create a Merge dialog box in Protégé/OWL for an editor to merge one OWL class to another. Further, we can hide the Delete button on the NCIOWLClass Tab programmatically based on user privilege. This meets one of our editing workflow requirements. We have also identified Protégé/OWL functionality to validate Frame names that we may be able to utilize for validating a class' Preferred Name according to our editing guidelines.

We have started testing the Wizard Plug in from Manchester, and think that this will be a very useful addition to workflow for batch loads and batch edits, which we do frequently with our extensions to TDE.

Visualization: We partially supported the development of Prompt Viz and its modification to work with Protégé/OWL in hopes that we can use it to help identify editing changes over time and from baseline to baseline. We are testing OWL Viz, Ontoviz and Jambalaya as well, with small test vocabularies, but have not begun to test with NCI Thesaurus. We hope that visualization of semantic relationships other than is-a will improve the usability of the NCI Thesaurus.

Summary:

NCI Thesaurus is a large complex vocabulary operating in a production environment. In order for us to contemplate switching from our current editing environment, we have to demonstrate that we can satisfy quite a large number of requirements, which we are just beginning to test. We have a long way to go before the evaluation is complete, but we haven't yet encountered insurmountable issues.