NCI Edit Tab: Protege 3.4 Plugin for Editing and Maintaining NCI Thesaurus

Sherri de Coronado, Gilberto Fragoso

NCI Center for Bioinformatics and Information Technology
National Institute of Health, Bethesda, USA

NCI Edit Tab is a package of plugins for Protégé 3.4 for supporting terminology editing and maintenance at the National Cancer Institute (NCI), publicly available for download from an SVN repository at the NCI's Gforge.[1] It provides for workflow, customized editing, unique code generation and history recording. There are also plugs for batch editing and loading as well as Lucene-based queries and reports. This suite of integrated terminology tools is being used for production development and maintenance of NCI Thesaurus and BiomedGT. The objective of this presentation is to introduce the features and some implementation details.

NCI Thesaurus (NCIt)[2] is a biomedical terminology currently containing about 80,000 concepts. It supplies a majority of the controlled terminology for the Cancer Bioinformatics Grid (caBIG)[3] as well as special terminology for FDA, CDISC and other organizations and other parts of the cancer research community. The terminology is published to the NCI implementation of the LexBIG terminology server, and accessible through web browsers, directly through the LexEVS[4] API or by download in several formats, including OWL DL.

NCIt is published monthly, and edited by a group of about 15 domain expert editors who need to work simultaneously. The NCI Edit Tab supports this workflow using Protégé in client server mode with a database backend for concurrent editing.

Curation*:* The NCI Edit Tab was designed to create a simpler user interface for domain expert editors than the native OWL interface in Protégé (Figure 1). It also supports the change of ontologies over time controlling those changes with notions of splitting and merging, retirement, change ontologies and history (Figure 2). As well, it is designed to provide a large number of edit checks and enforce a number of business rules in editing that improve the quality. A code generator plugin to automatically generate unique concept IDs has been included and meta level data such as who and when are captured for auditing purposes.

Workflow: Workflow managers are assigned specific privileges through the Metaproject, which enable them to create worklists through the workflow tab. The workflow manager is also responsible for doing a PROMPT cycle periodically, to review changes by other editors, and to produce a new baseline for further editing, or once a month, for publishing. The PROMPT (with a plugin that generates concept history) is used to review and accept or reject changes editors have made during a baseline. The concept history file is produced at the same time.

---

[1] https://gforge.nci.nih.gov/frs/?group_id=174

[2] See: http://bioportal.nci.nih.gov to browse NCI Thesaurus content or http://bioportal.biontology.org

[3] See: http://cabig.nci.nih.gov for further information about caBIG

[4] See: https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/LexBIG/EVS for more information on LexBIG and LexEVS.

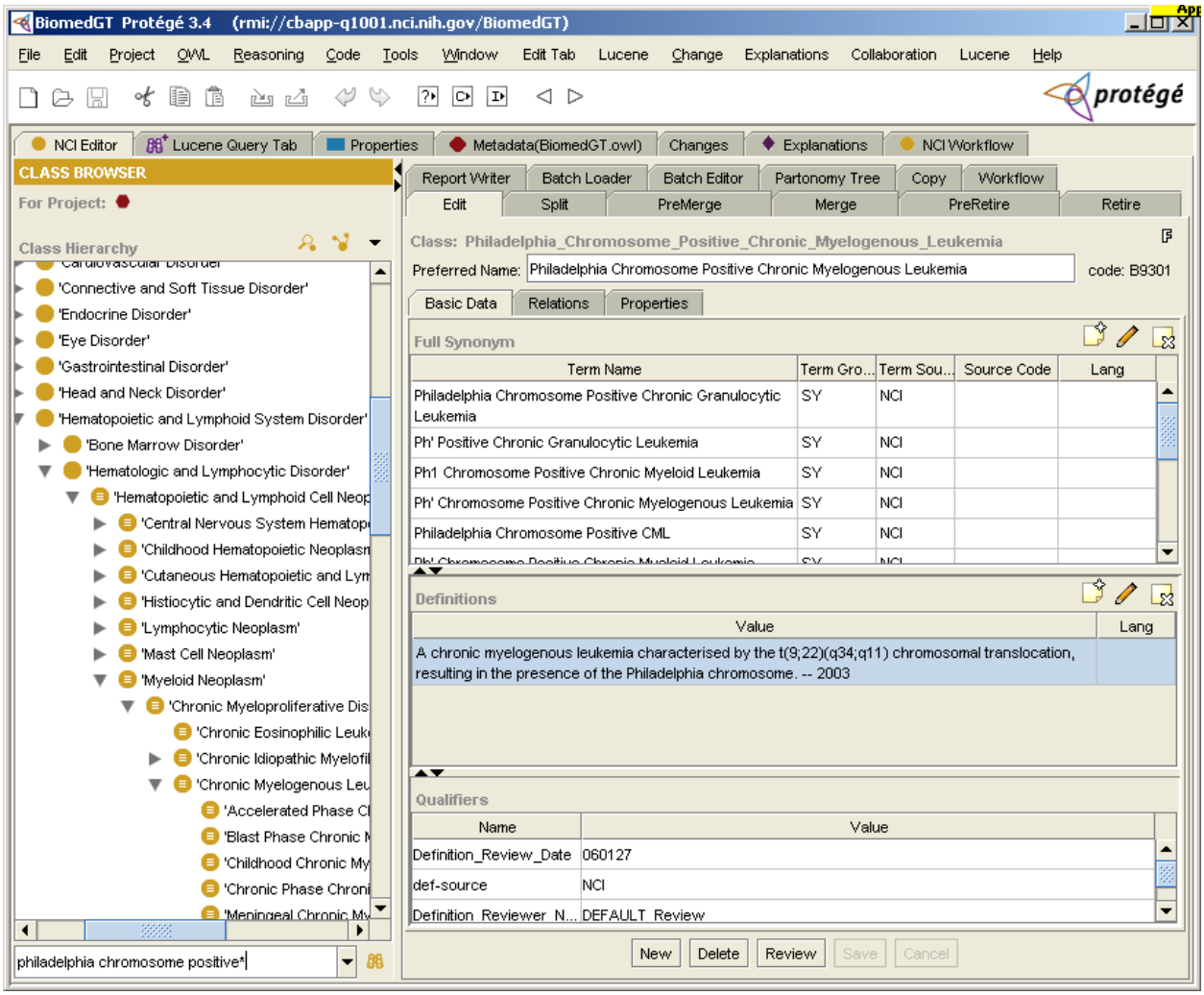

Figure 1 NCI Editor Plugin

Query and Reporting: The NCI Edit tab includes a ReportWriter subtab, which can generate concept reports of individual concepts or of the set of concepts found under a node in the class hierarchy. The reports present all the asserted annotations or restrictions on a class in a name-value pair format. In addition, the Lucene Query Tab has been enhanced with the reporting functionality found in the Query tab included in the base Protégé installation. These reports can include any user-selected property or restriction in a spreadsheet format, all annotations or restrictions found in a single row of the spreadsheet for any given class.

Batch Load and Edit: Batch loading and editing is extremely important for such a large ontology. Here, editors can work out the details of a large set of concepts ahead of time. Concepts (classes) can be created, with their parents, and codes assigned during the batch load process that uses a very simple input file format. The group of concepts can then be assigned terminological properties such as synonyms and definitions in a second for batch editing the already created concepts. Additional editing can then be done through the regular editing interface.

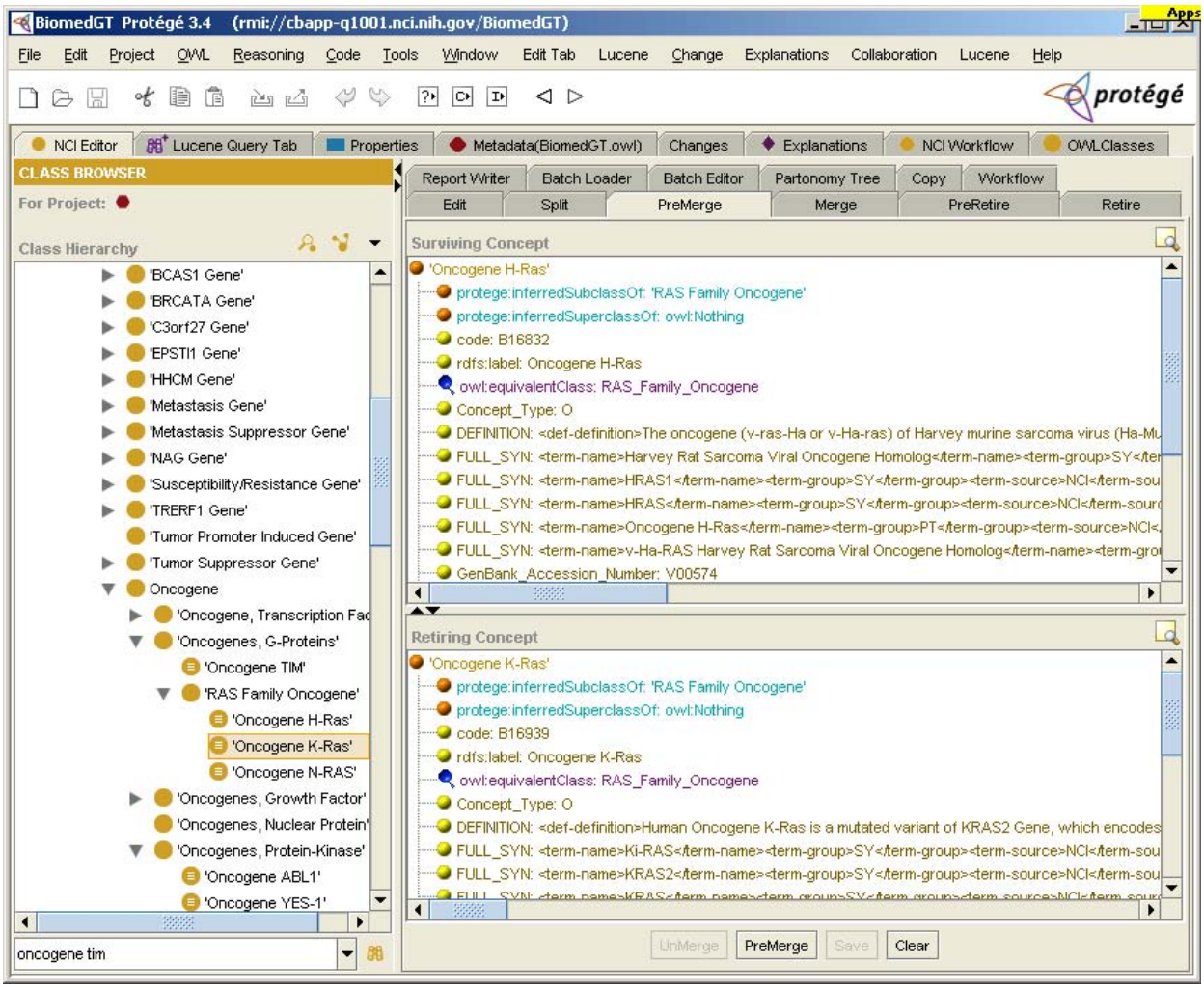

Figure 2 Support for merging concepts

Server Environment: The host server runs the Protégé server, the database engine (MySQL), and an explanation server that queries Pellet. The database is built with the Clark&Parsia schema, which ameliorates performance penalties on classification when the ontology utilizes database inclusion. Because of the size of the NCIt and the fact that it will continue to grow and that the NCI edits additional vocabularies (e.g. BiomedGT), hardware resources can become critical at various stages of the workflow cycle (e.g. PROMPT analysis, classification with Pellet). We are currently running the servers in a 64 bit Linux host with 8 CPUs and 32 GB of memory; however, even this amount of memory is tight on occasion, for instance when edits are performed that are later found to make the NCIt inconsistent and require that memory to the Pellet explanation server (Figure 3) be increased above the 8 GB allocated routinely.

The presenters will also demonstrate NCI Edit Tab during the demonstration session at the Protégé Conference.

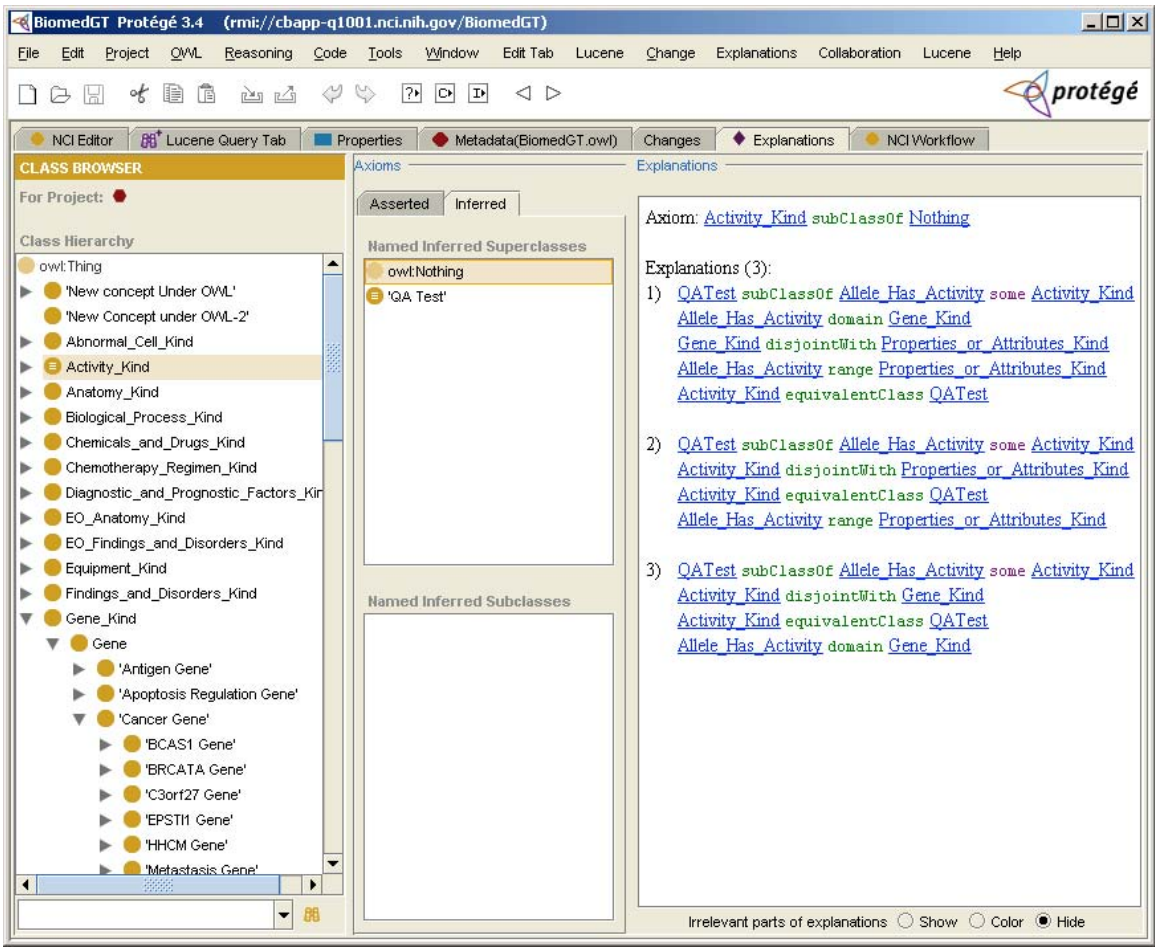

Figure 3 Support for explanation