Protégé 3.4 Plug-in for Editing and Maintaining the NCI Thesaurus

Protégé Conference
June 23, 2009
Amsterdam
Sherri de Coronado, Gilberto Fragoso

caBIG cancer Biomedical Informatics Grid®

NCI Enterprise Vocabulary Services

Goal – Integration by Meaning

- EVS provides services and resources that assists to:
  - Integrate different conceptual frameworks for clinical, basic and translational research,
  - Create terminological and taxonomic conventions across systems

- Controlled Terminology Products
  - NCI Thesaurus – an ontology-like cancer-centric controlled terminology
  - NCI Metathesaurus – maps biomedical vocabularies
  - BiomedGT (Biomedical Grid Terminology - new)
  - External vocabularies maintained and served: MedDRA, HL7, NDF-RT, LOINC, GO, Zebrafish, RadLex, etc.

- Further info, see: https://wiki.nci.nih.gov/display/EVS/EVS+Wiki

caBIG

Products: NCI Thesaurus

- Reference Terminology for NCI, caBIG, Partners
  - Underpins caCORE, caGRID semantics

- A Federal Standard Terminology

- Public domain, open content license

- 80,000 "Concepts" hierarchically organized into domains

- Broad coverage of the cancer research and clinical domain including prevention and treatment trials
  - Neoplastic and other Diseases
  - Findings and Abnormalities
  - Anatomy, Tissues, Subcellular Structures
  - Agents, Drugs, Chemicals
  - Genes, Gene Products, Biological Processes
  - Animal Models – Mouse, other
  - Research techniques and management, apparatus, clinical and lab, radiology, imagery

caBIG

Products: NCI Thesaurus (2)

- Description-logic based

- Concept History

- Published Monthly

- Accessible via API, web browsers, downloadable files

- Transition to OWL begun in '03

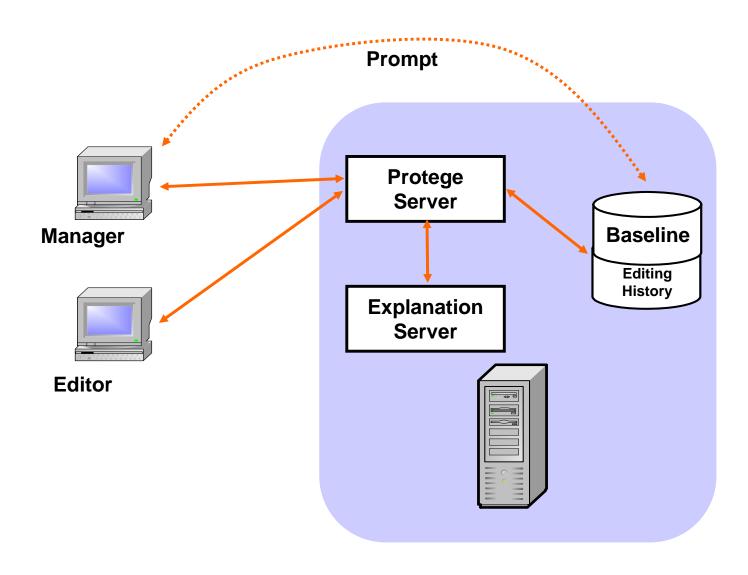# Editing Tool Requirements

- Shared Data and Distributed Editing

- Reasoning

- GUI for Subject Matter Experts
    - search and reporting facilities

- Editing Consistency
    - basic content – preferred and alternative terms, definition

- Complex Operations
    - merge, split, retirement
    - tied to history tracking

- Rule Enforcement, Edit Checks
    - no duplicate restrictions, semantic type

- Support for Workflow, Editing Roles (manager, editor)

Client – Server Environment

Prompt

Manager

Editor

Protege Server

Explanation Server

Baseline

Editing History

caBIG

Edit Checks Configured into SW

Table 1 – Selected edit checks built or configured into the software

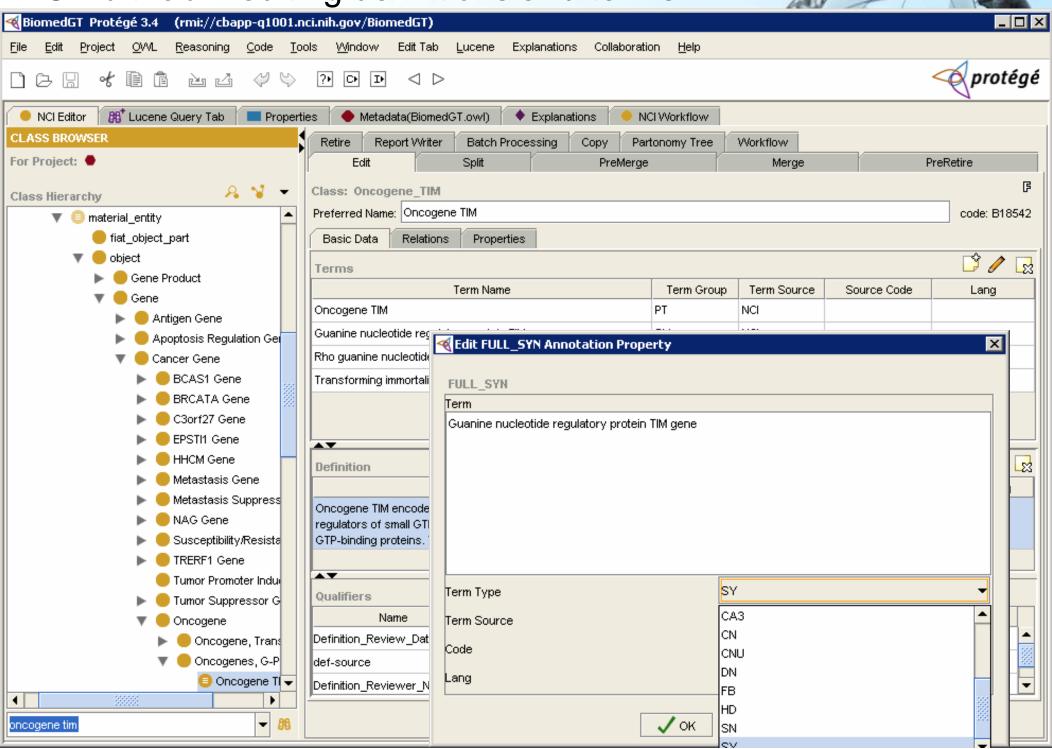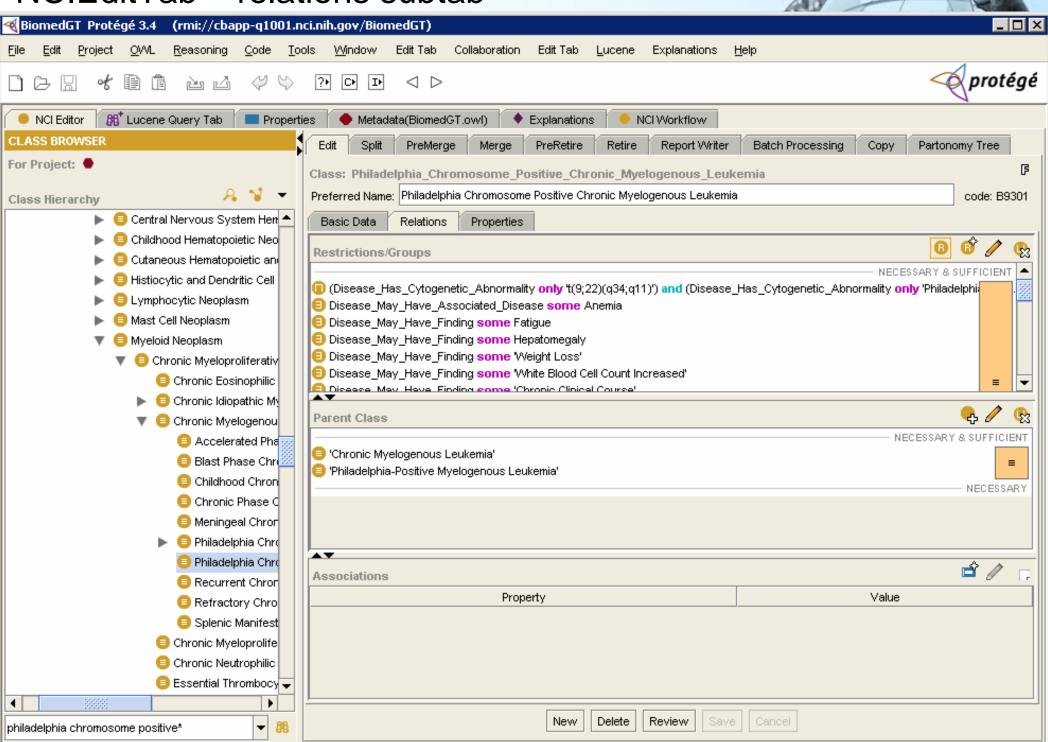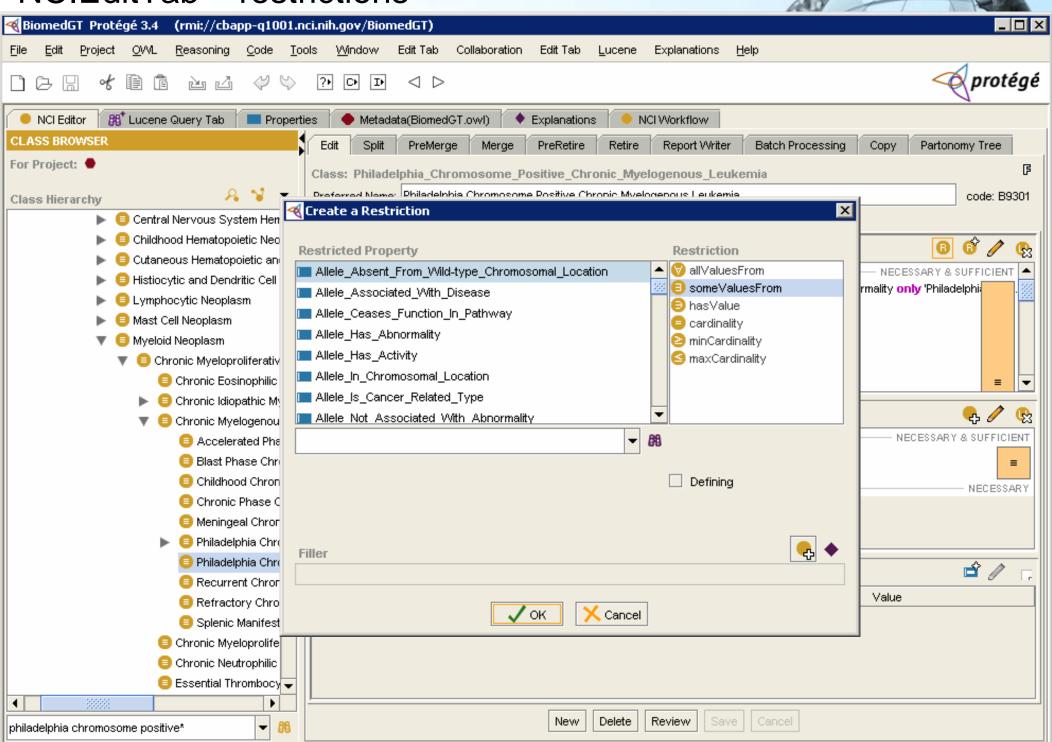| Edit Check Entity	Description
Concept Name	Cannot be changed (although preferred term can be); it must begin with letter or underscore.
Preferred Name	Concept must have one and only one Preferred Name; it must match the fully qualified synonym with group PT and source NCI.
Duplicates	Duplicate parents, roles and properties are not allowed.
Definition	Each must have 1 review date, 1 review name, 0 or 1 attributes; no characters less than utf-8 32 allowed; !,? or @ allowed, single spaces only in definitions unless preceded by these special characters.
Retired Concept	Only lead editor can retire concepts, although editors can pre-retire concepts. An editor's note should explain the retirement.
Merged Concept	Only lead editor can merge concepts, although editors can pre-merge concepts, and pre-merged concept must include an editor note with value of pre-merge annotation and an explanation if needed.
Split Concept	Check that newly created concept is a valid concept. All checks made during normal create are made during a split.
Other	Cannot create or maintain a restriction relationship that points at a retired, pre-retired or pre-merged class.

NCIEditTab Layout

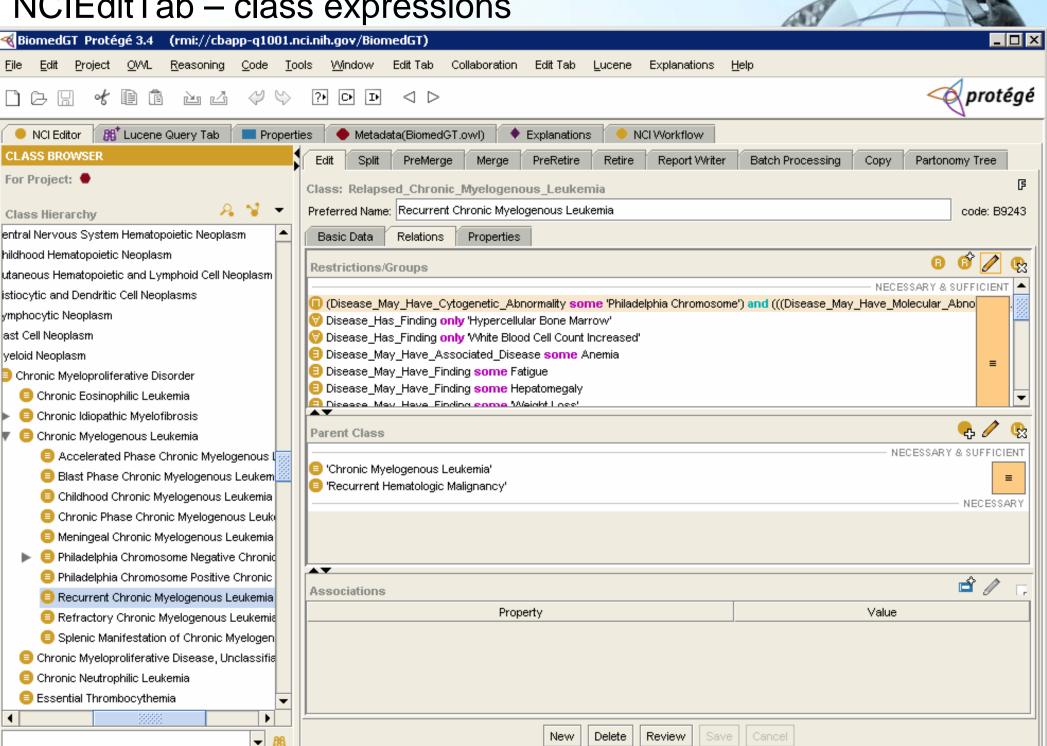# NCIEditTab – basic data

NCIEditTab – editing definitions and terms

NCIEditTab – relations subtab
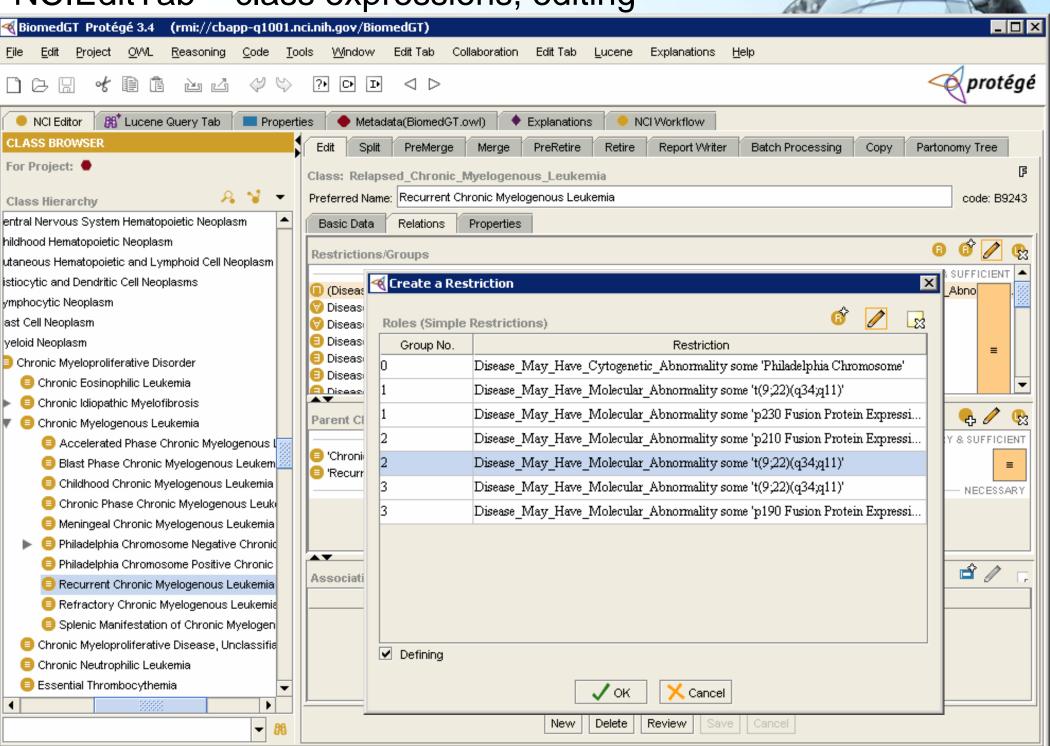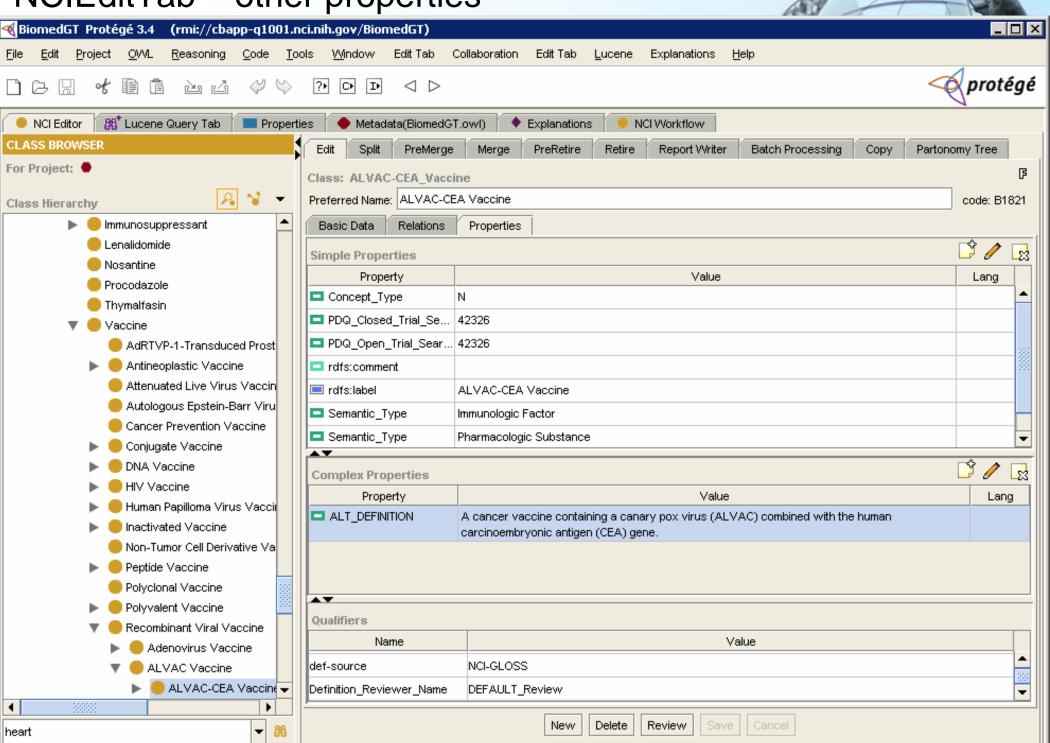
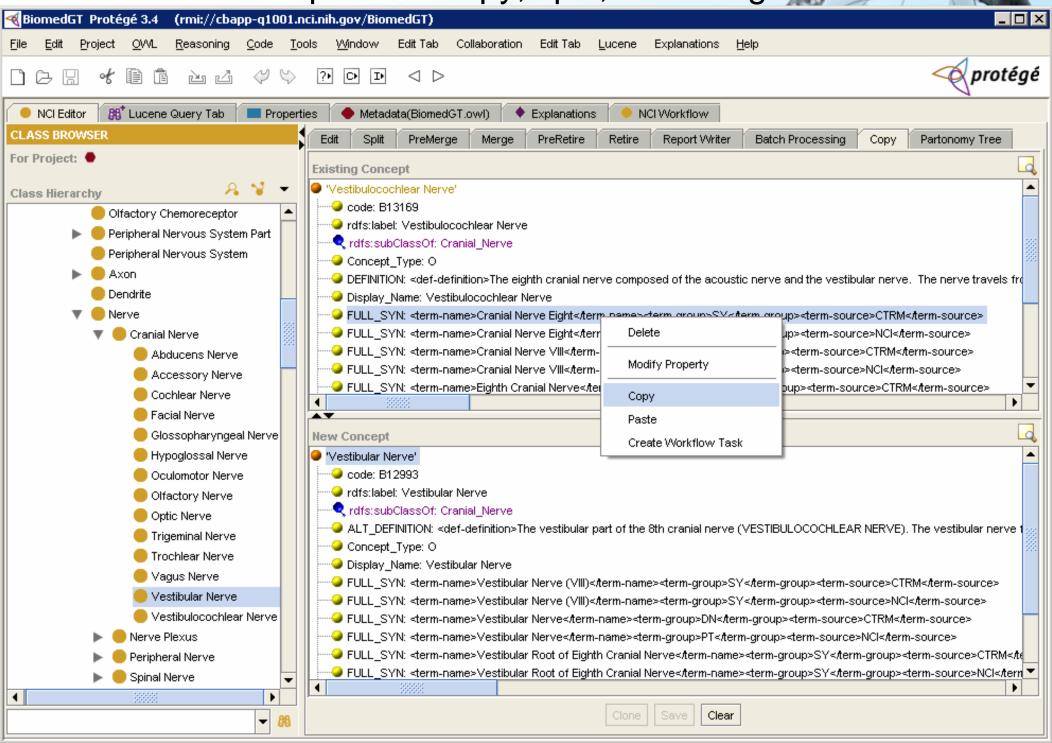# NCIEditTab – restrictions

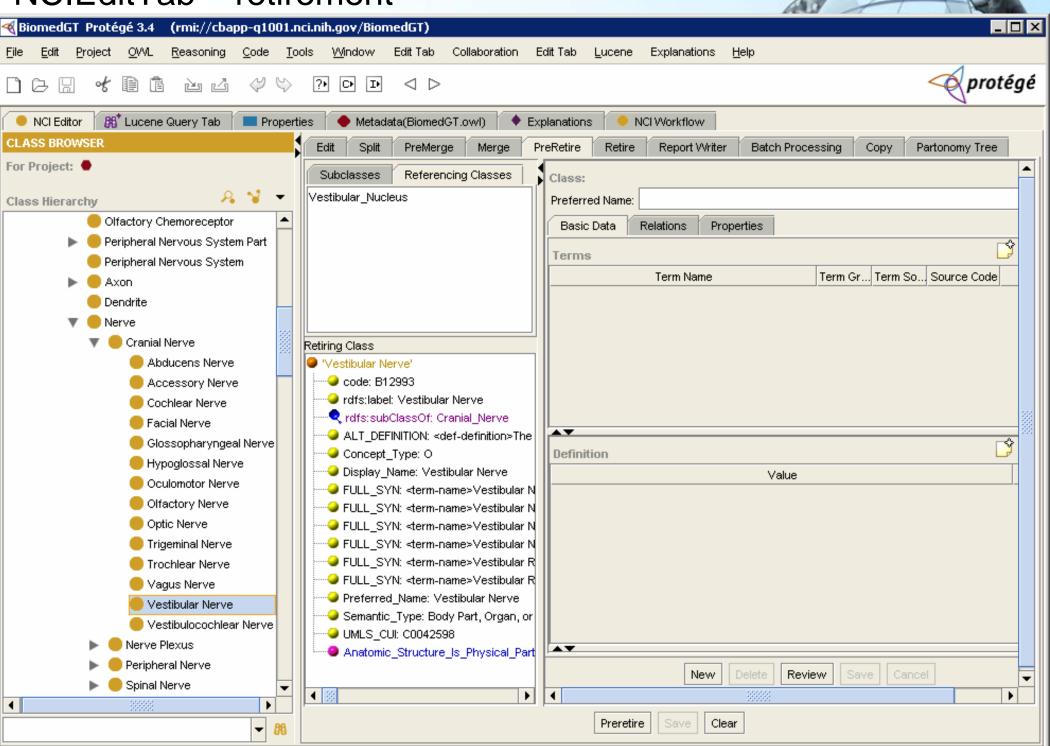# NCIEditTab – class expressions

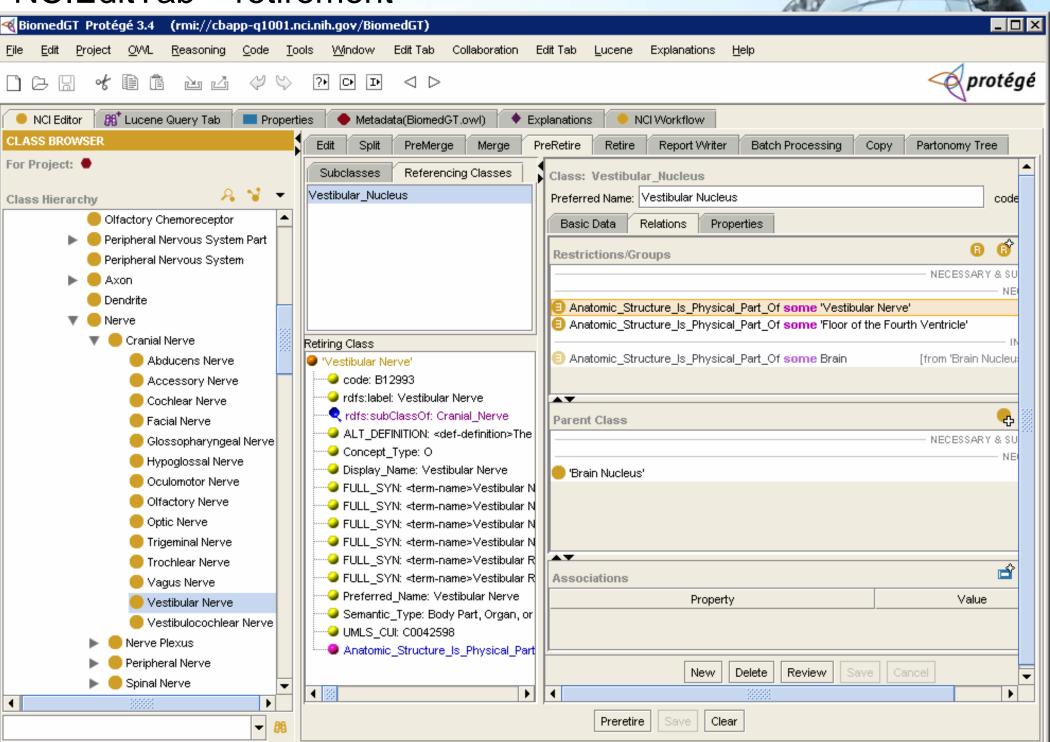# NCIEditTab – class expressions, editing

NCIEditTab – other properties
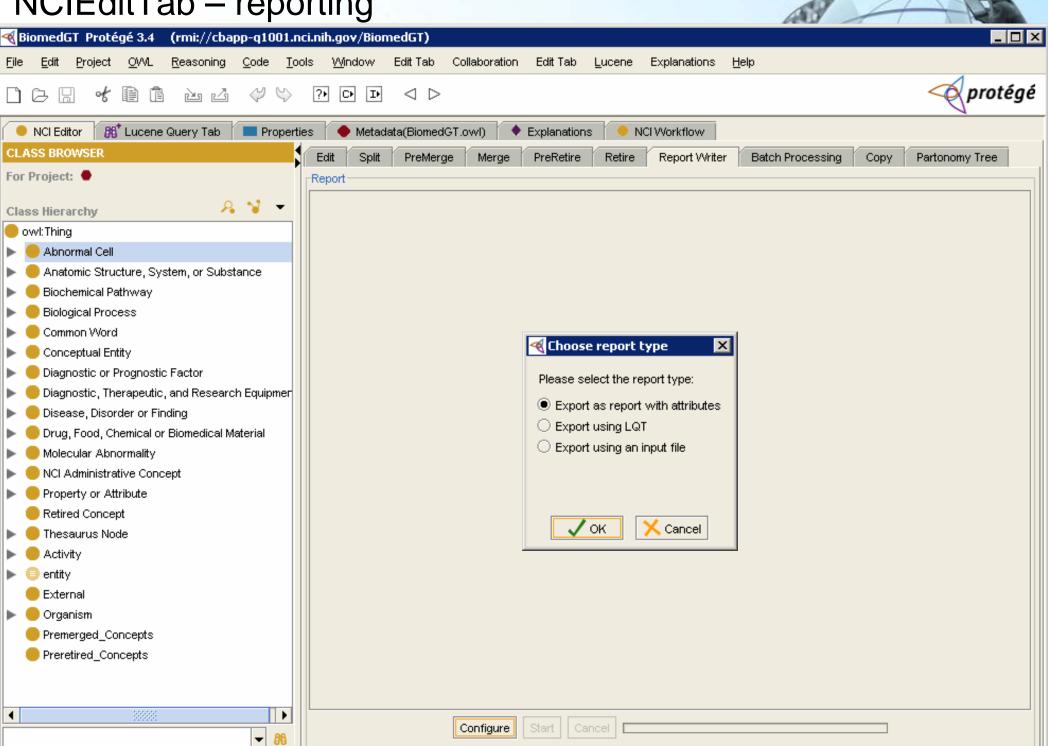
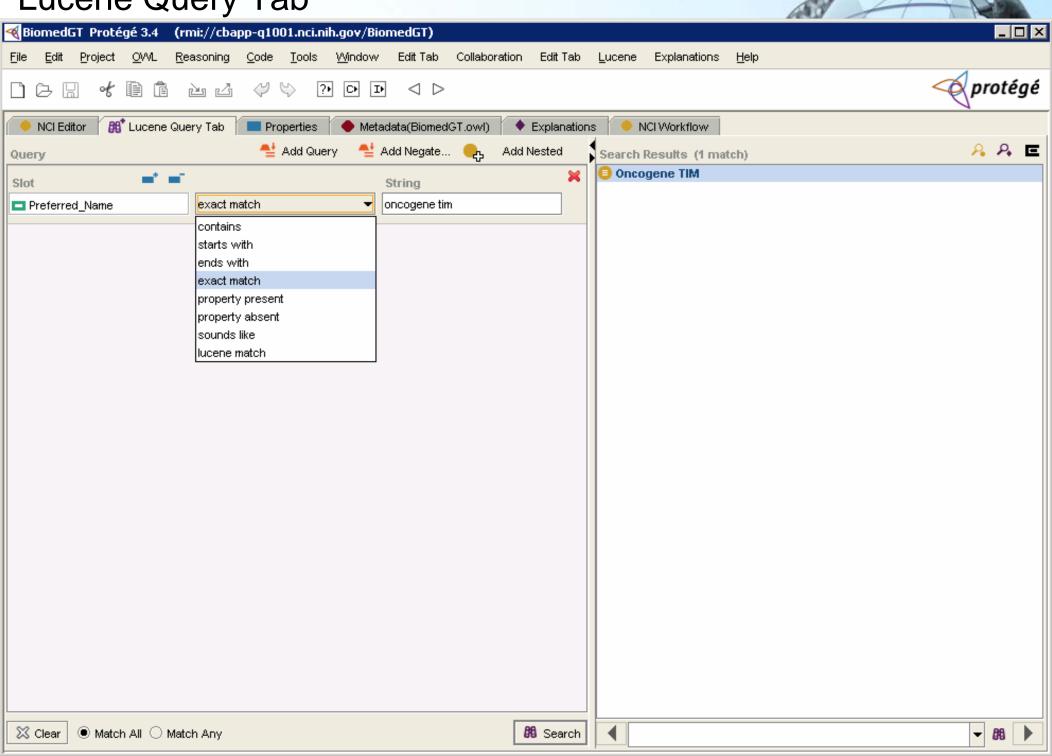# NCIEditTab – tree panel in copy, split, and merge

NCIEditTab – retirement

NCIEditTab – retirement

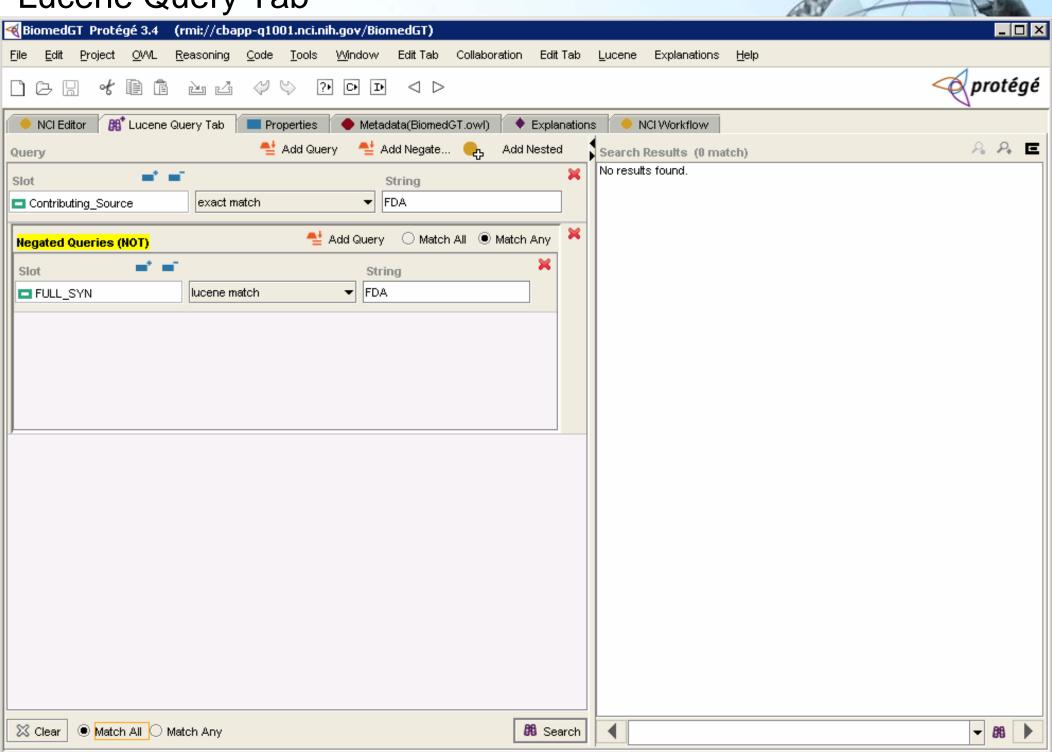# NCIEditTab – reporting

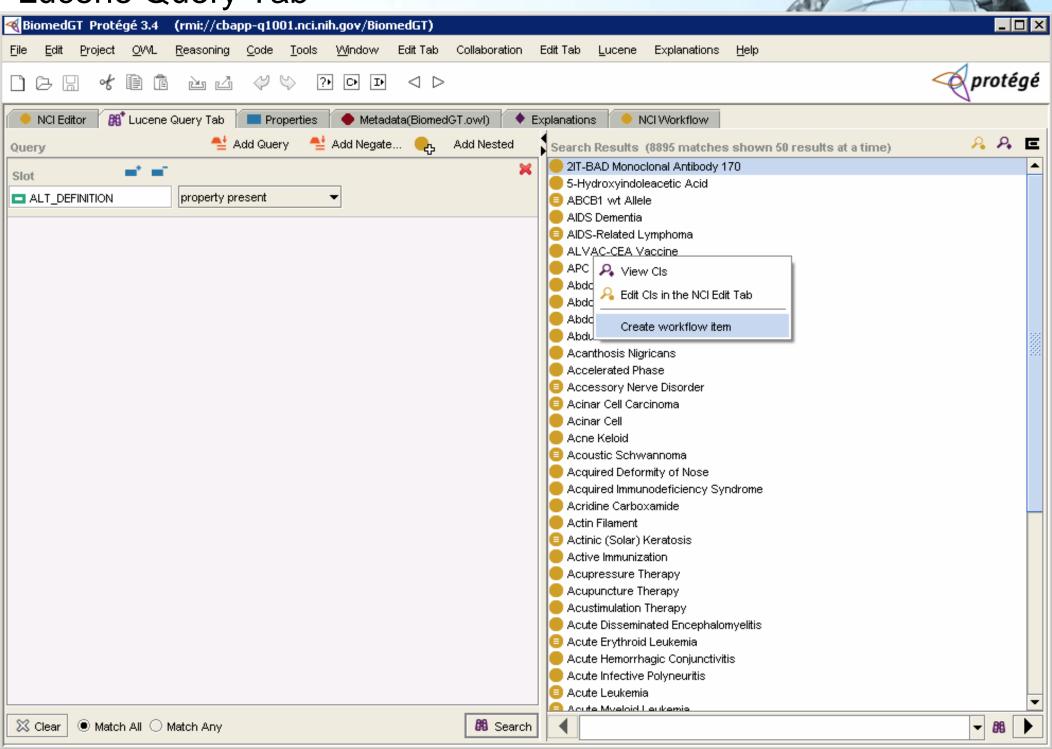# Lucene Query Tab

BiomedGT Protégé 3.4 (rmi://cbapp-q1001.nci.nih.gov/BiomedGT)

File   Edit   Project   OWL   Reasoning   Code   Tools   Window   Edit Tab   Collaboration   Edit Tab   Lucene   Explanations   Help

protégé

| NCI Editor | Lucene Query Tab | Properties | Metadata(BiomedGT.owl) | Explanations | NCI Workflow |

Query   Add Query   Add Negate...   Add Nested

Search Results (1 match)

Oncogene TIM

Slot

String

| Preferred_Name | exact match ▼ | oncogene tim |

contains
starts with
ends with
exact match
property present
property absent
sounds like
lucene match

Clear   ● Match All ○ Match Any   Search

Lucene Query Tab

Lucene Query Tab

NCI Workflow Tab

Classification and Explanation Service

- Maturing the Pellet reasoner - NCIt Classification
  - Prior to initial work: *non-terminating*
  - Improving resource efficiency: 9 hours
  - Algorithmic optimizations: 5 minutes
  - Incremental updates: *seconds*

- Explanation Service
  - Improve the efficiency of editors by identifying problems *and causes*

Modified from M. Smith, Clark&Parsia

OWL Classes Tab – classification views

Explanation Tab

Download site

Acknowledgements

EVS Team

- Protégé/ NCI Protégé programmers
  - Stanford BMIR staff
  - Dionne Associates
  - Clark & Parsia
- Production and QA staff
  - Steve Hunter (Ekagra)
  - M.A. Storey's group (UVic)
  - Tracy Safran, Rob Wynne, John Park

- NCI staff
  - Frank Hartel
  - Gilberto Fragoso
  - Sherri de Coronado
  - Margaret Haber
  - Larry Wright

caBIG® cancer Biomedical Informatics Grid®