NCI Thesaurus, managing towards an ontology

CENDI/NKOS Workshop
October 22, 2009
Gilberto Fragoso

Outline

- Background on EVS
- The NCI Thesaurus
- BiomedGT
- Editing Plug-in for Protege
- Semantic Media Wiki
  - supports collaborative development

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 Recommended Federal Standard Terminology for Anatomy
- 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

Issues and Challenges

- Content
  - many domains, users and uses
- Support
  - collecting input from end users and collaborators
- Tracking Provenance
- Use of Properties
  - standard terminology for properties (e.g. preferred_term vs term vs ?)
- QA and Editing Consistency

- New use cases
  - Query and reasoning against instance data on the Grid
  - Federation of ontologies and subontologies

NCI Thesaurus Evaluation 2006-2007

Goals

1. Review and report of OBO criteria and relevant ISO standards for semantic quality and federation of terminologies, semantic quality and consistency.
2. Review content and structure for compliance
3. Document examples of how compliance would be achieved

Among the Recommendations

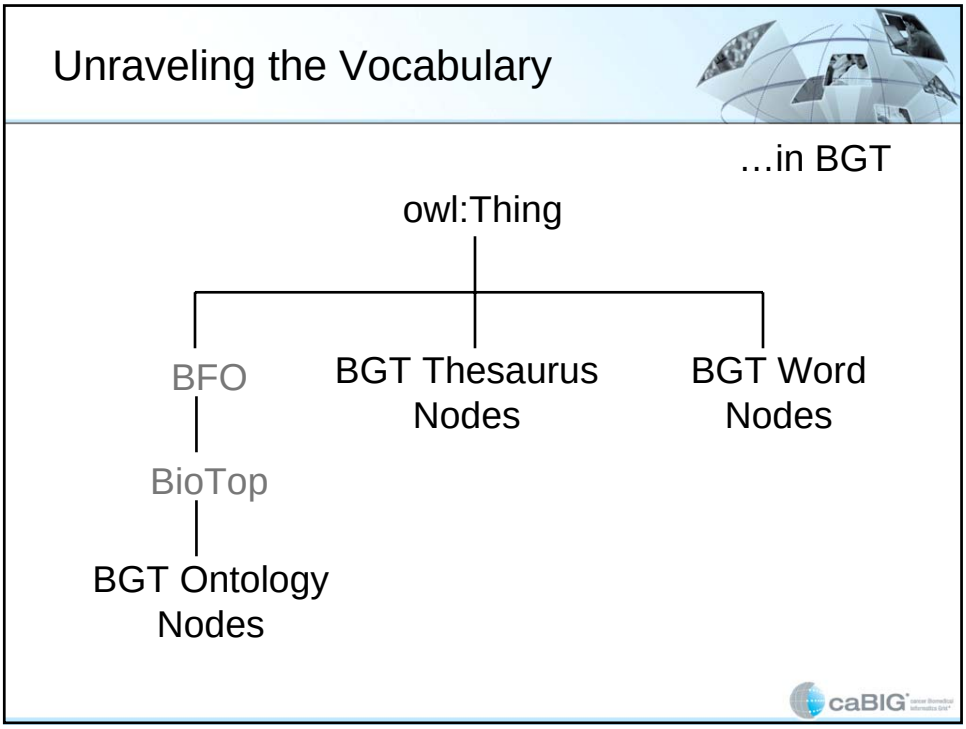
(Genesis of BiomedGT)

1) Unravel the vocabulary
   - Partition into:
     - Words and their definitions (Lexicon / Dictionary)
     - Categorization and navigational nodes (Thesauri)
     - Ontology

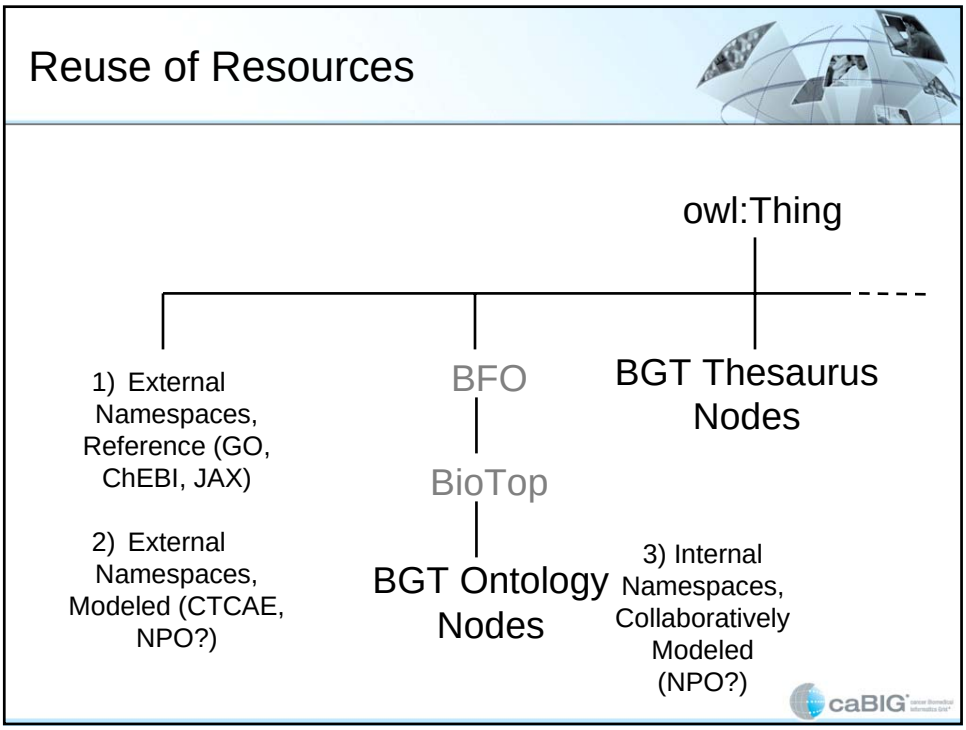
2) Identify external resources and:
   - Include ability to reference general upper level ontologies
   - Named relationships with other external resources

3) Enable collaboration: create environment where SME's can collaborate and discuss

Unraveling the Vocabulary



Reuse of Resources



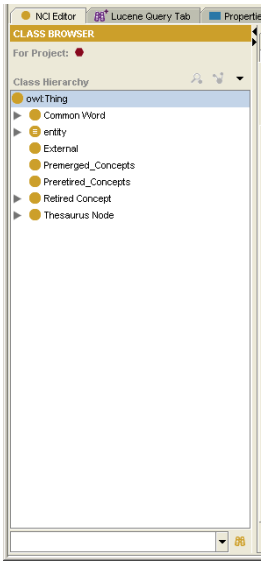
BGT Initial Development

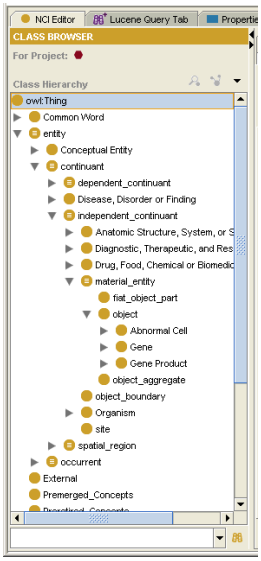
- NCIt content used to seed BGT
- Categorization of all BGT concepts as O, N, C
- Eliminate NCIt-specific branches and concepts
- Partition concepts – BFO, Thesaurus, and Words (ongoing)
- Import external vocabularies (ongoing)
- Refactor annotation properties (ongoing)
  - name and format
- Refactor object properties (standby)

Concept Removal and Federation

Branch Removed	Initial Federation	Removal after Refactoring
NCI Thesaurus Entity	GO	Biological Process
Chemotherapy Regimen or Agent Combination	chEBI	Biochemical Pathway
Experimental Organism Anatomical Concept	Sequence types and features ontology	Organism
Experimental Organism Disease	NCBI Taxon	Disease or Disorder
Medical Device Component or Accessory	Pathway Ontology	Molecular_Abnormality
Medical Product Usage and Evaluation	Protein Modification	Chemicals and Drugs
Reportable Event	PATO	
UML Entity	NCI Thesaurus Disease or Disorder Branch	
Business Rules	BFO	
Clinical or Research Facility		
Funding		
Information and Media		
Academic Degree		

Partitioning of BGT Concepts





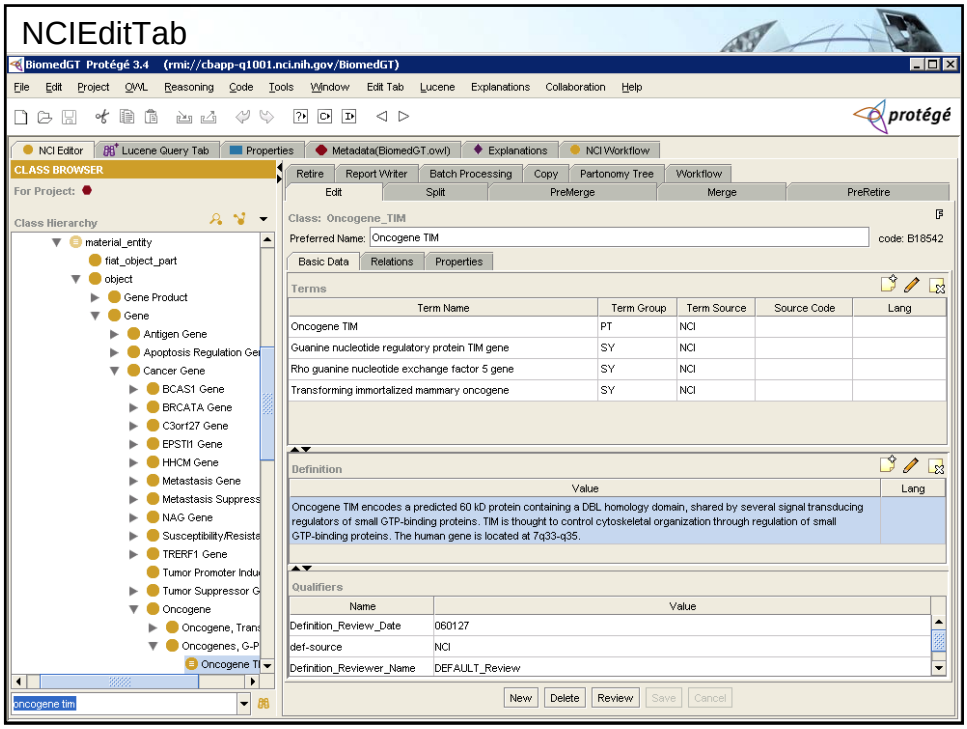
caBIG

Editing Tool Requirements

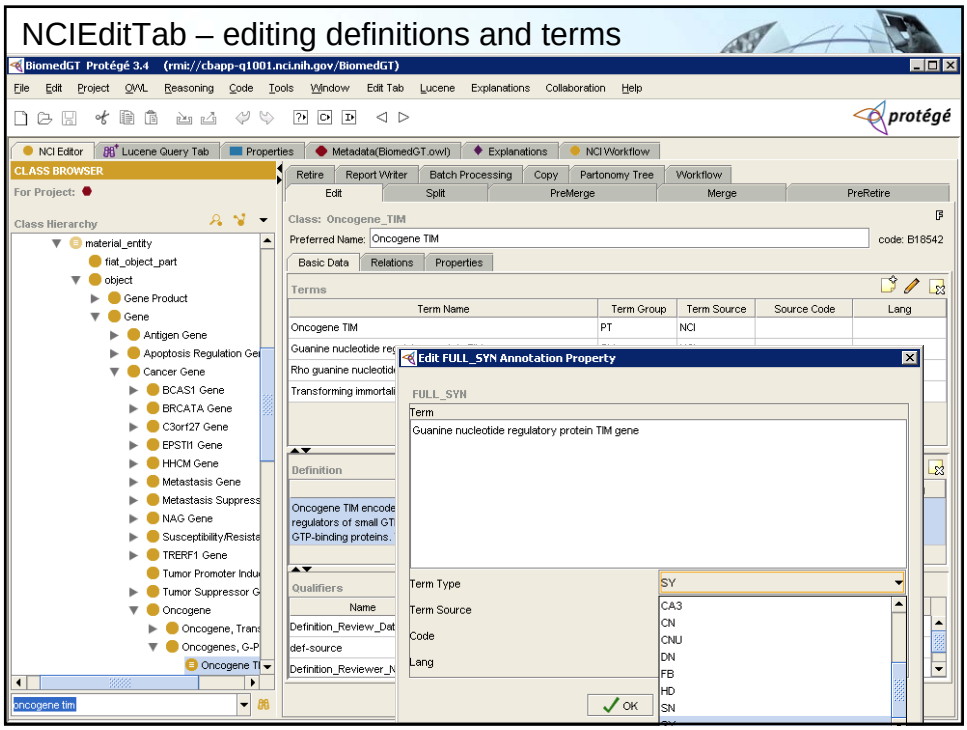
- Shared Data and Distributed Editing
- Reasoning
- GUI for Subject Matter Experts
  - search and reporting facilities
- Editing Consistency
  - basic content – preferred and alternative terms, definition
- Complex Operations
  - merge, split, retirement
  - tied to history tracking
- Rule Enforcement, Edit Checks
  - semantic type, no duplicate restrictions
- Support for Workflow, Editing Roles (manager, editor)

caBIG

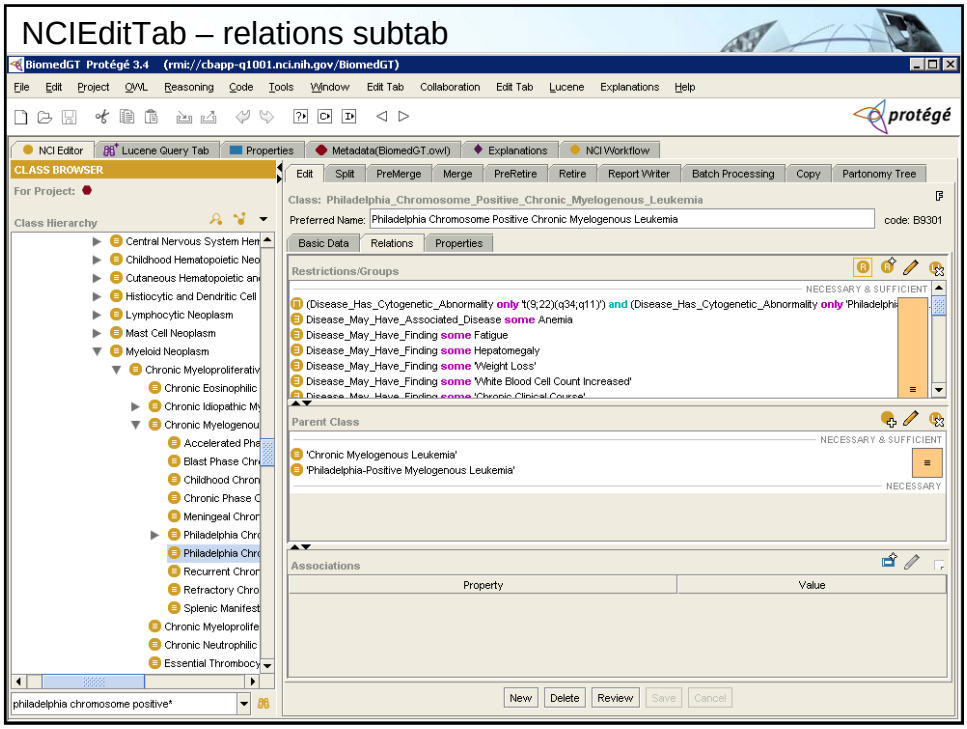
NCIEditTab



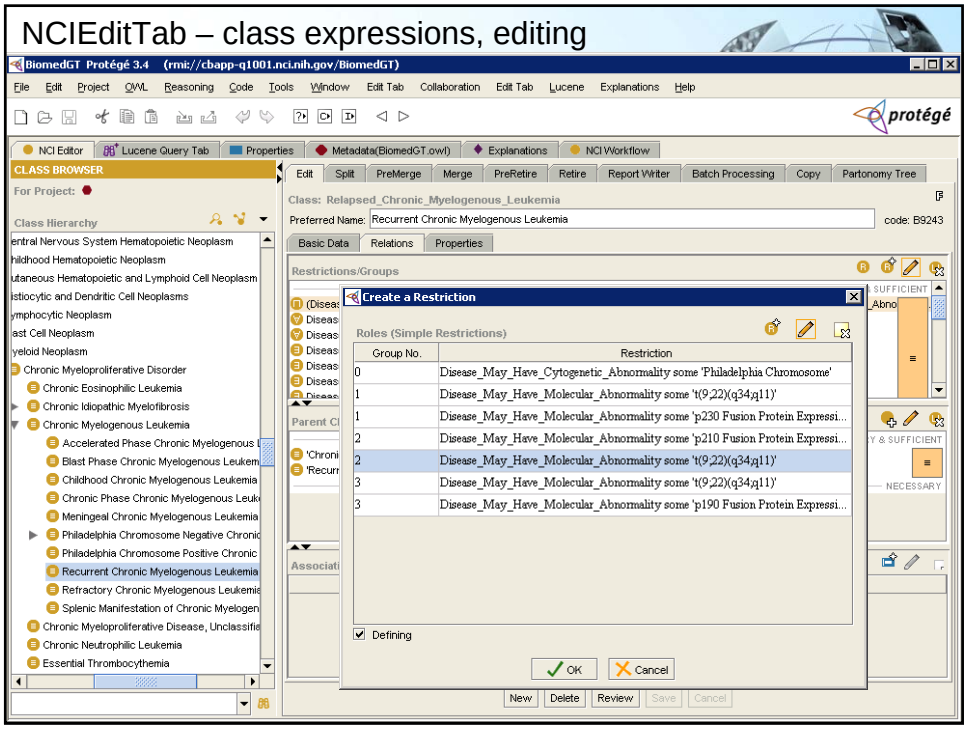
NCIEditTab – editing definitions and terms



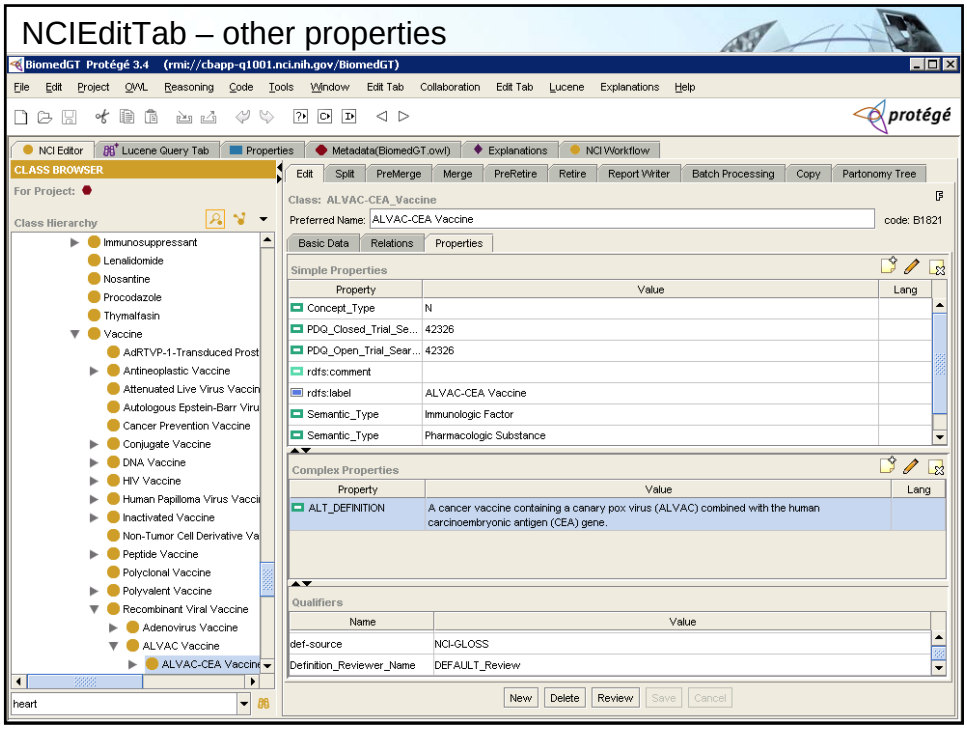
NCIEditTab – relations subtab



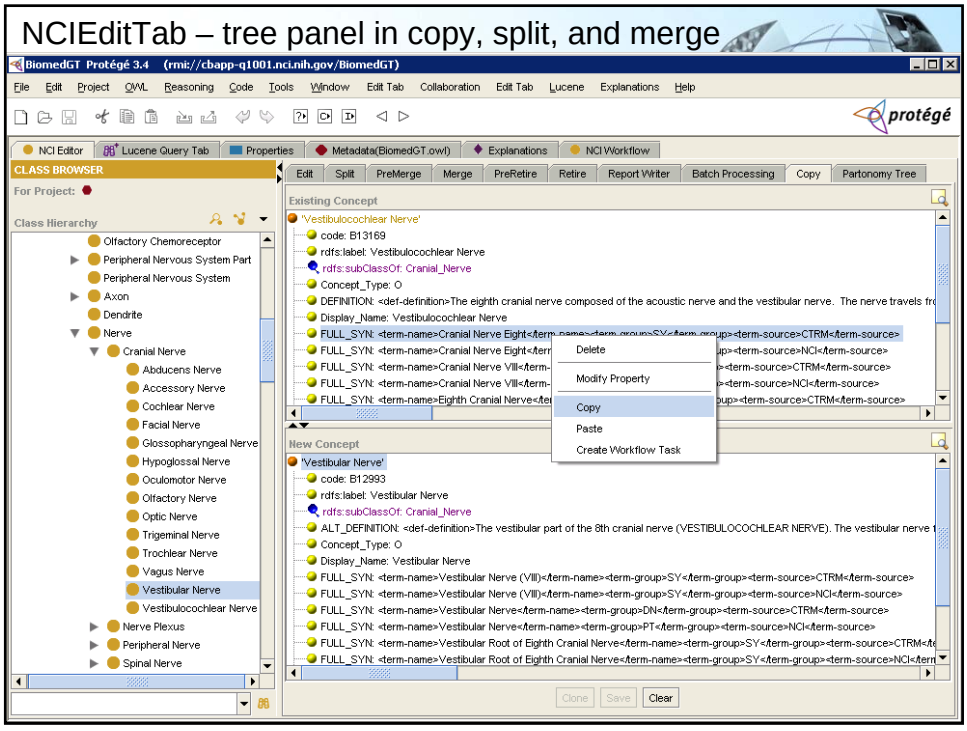
NCIEditTab – class expressions, editing



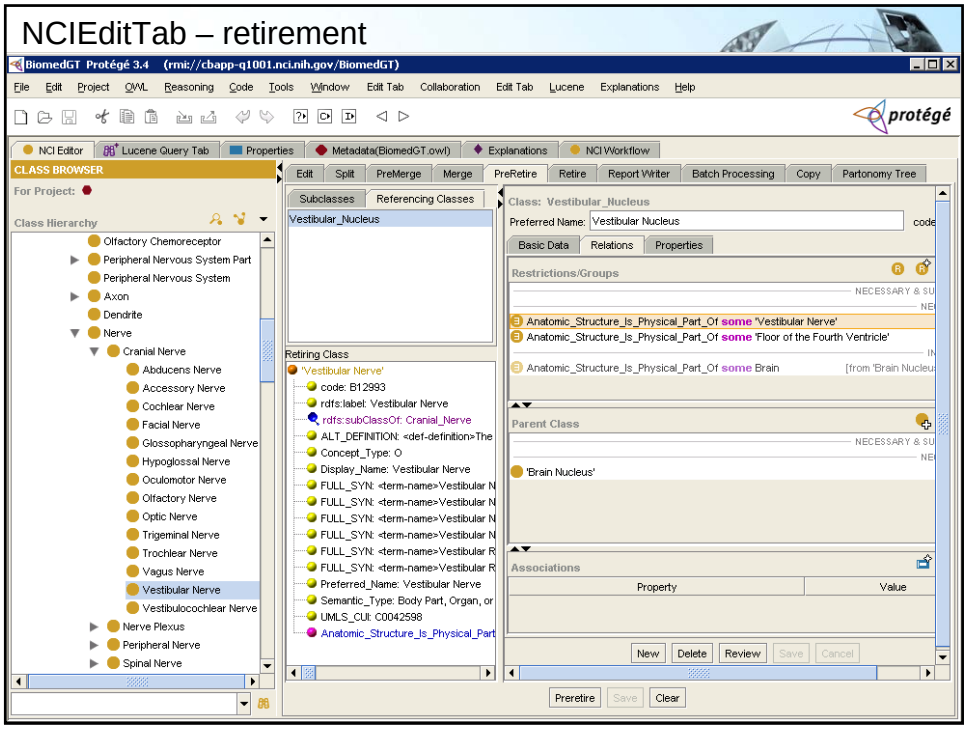
NCIEditTab – other properties



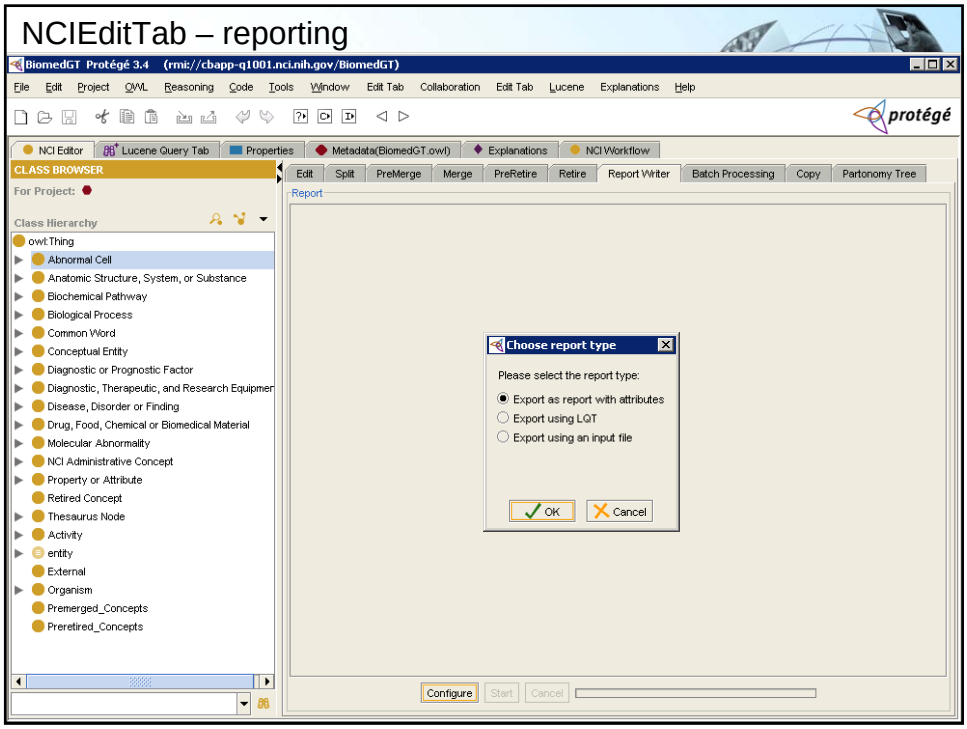
NCIEditTab – tree panel in copy, split, and merge



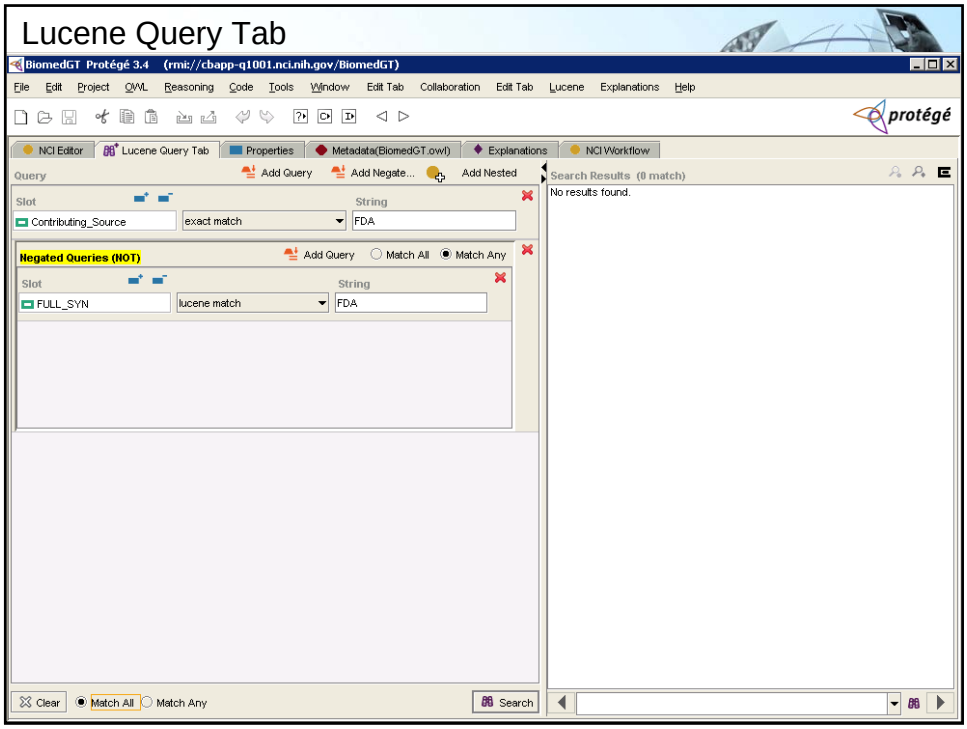
NCIEditTab – retirement



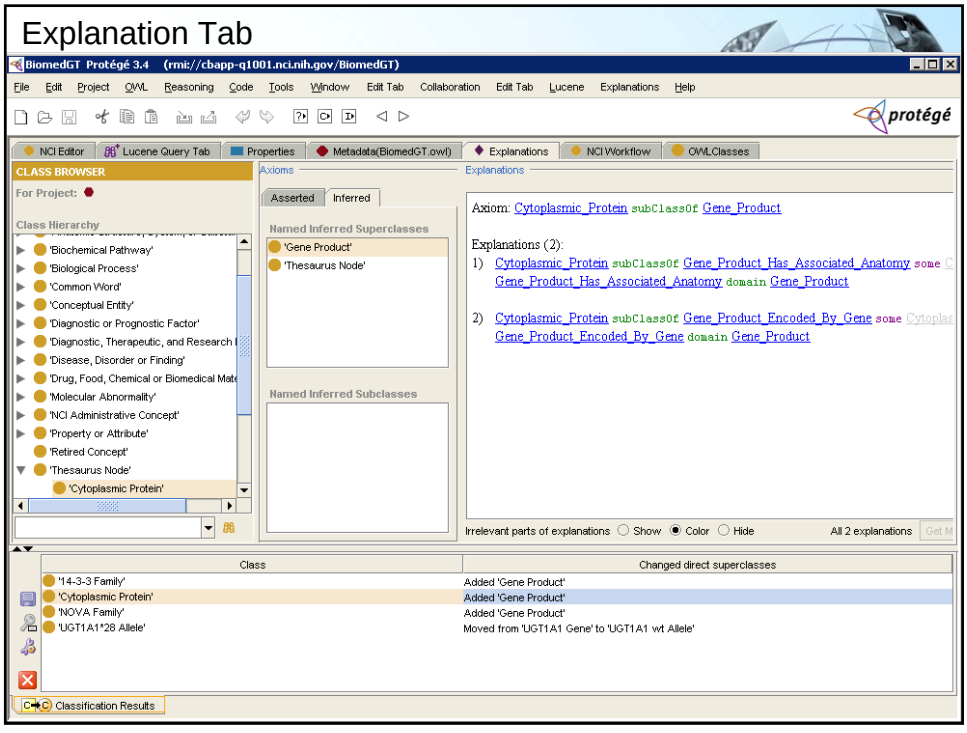
NCIEditTab – reporting



Lucene Query Tab

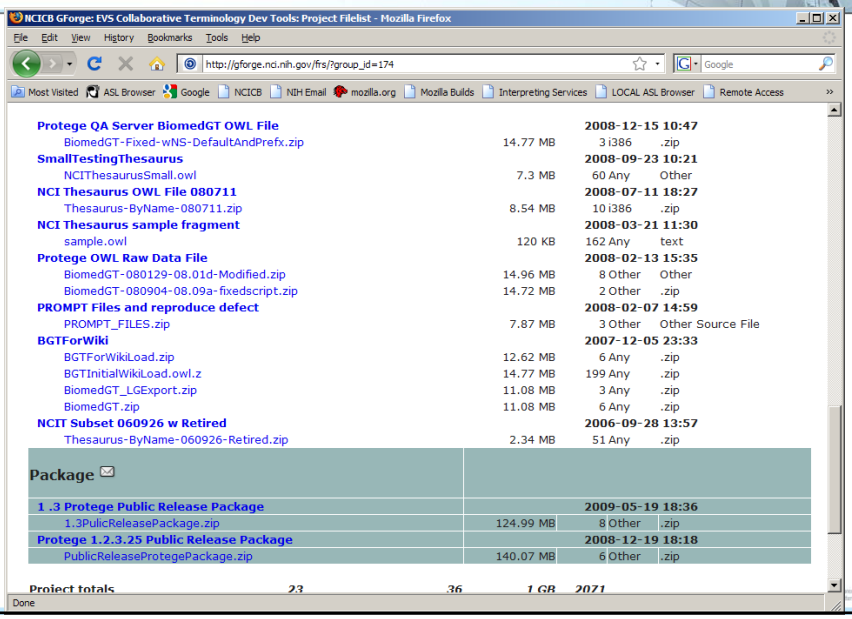


Explanation Tab

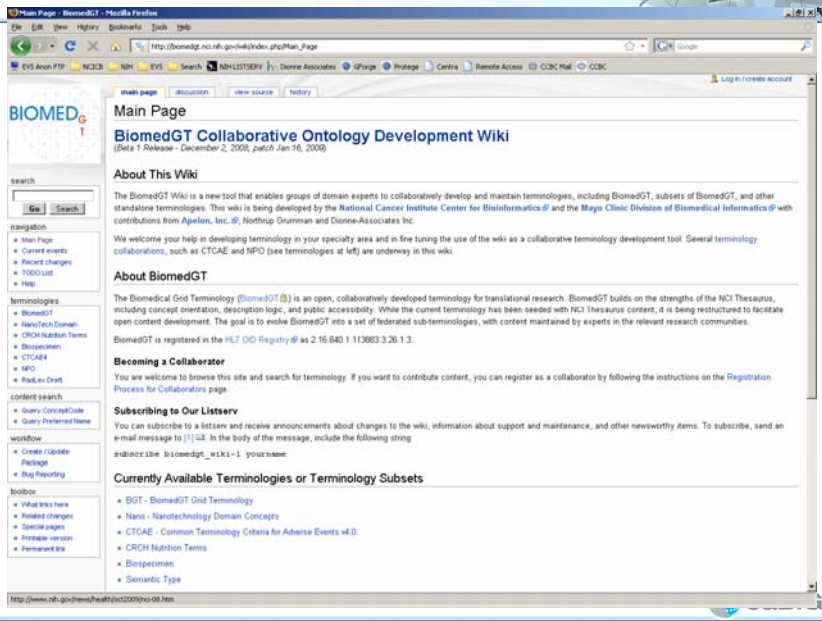


Download site

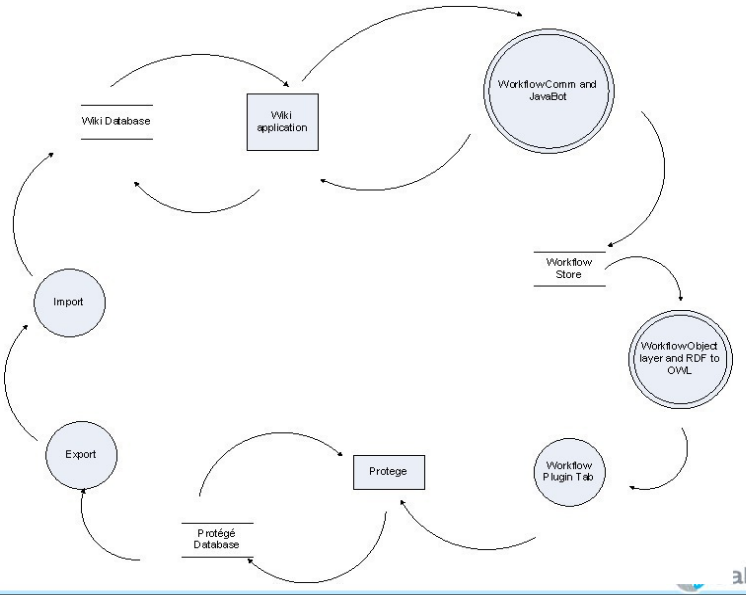
http://gforge.nci.nih.gov/frs/?group_id=174



SMW-based collaborative development



SMW- Protege Roundtrip



Acknowledgements

EVS Team

- **Protégé/ NCI Protégé/SMW programmers**
  - Stanford BMIR staff
  - Dionne Associates
  - Clark & Parsia
  - Mayo Clinic
- **Production and QA staff**
  - Steve Hunter (Ekagra)
  - M. Storey's group (UVic)
  - Tracy Safran, Rob Wynne, John Park
- **Editing**
  - Laura Roth
  - Lori Whiteman
  - Liz Hahn-Dantona and others (Lockheed Martin)
- **NCI staff**
  - Frank Hartel
  - Gilberto Fragoso
  - Sherri de Coronado
  - Margaret Haber
  - Larry Wright