



**caBIG**<sup>™</sup> cancer Biomedical  
Informatics Grid <sup>™</sup>

*An Initiative of the National Cancer Institute*

# Evaluation of Vocabulary Review Criteria

## ***Application to the NCI Thesaurus***

Olivier Bodenreider,  
National Library of Medicine

February 7, 2007

- **Evaluate vocabulary review criteria**
  - In the framework of caBIG
- **Application to one large-scale vocabulary**
  - NCI Thesaurus
- **Recommendations**
  - Applicability of the criteria
    - Operational definition
    - Scalability

# Proposed Criteria: High-level categories

1. URU (Understandability, Reproducibility, Usability)
2. Quality of Documentation
3. Maintenance and Extensions (Change Management)
4. Accessibility and Distribution
5. Intellectual Property Considerations
6. Considerations regarding Mapped Technologies
7. Quality Assurance and Quality Control
8. Concept Definitions
9. Community Acceptance
10. Reporting Requirements

- **Based on the review of the criteria**
  - Implemented by the Jackson Lab team
  - Applied to the Gene Ontology
- **Major difference**
  - Relies more on raw material
    - NCIT OWL file
  - Does not on rely on personal communication with the developers

- **Raw material: NCIT**
  - Tab-delimited ASCII file
  - Ontylog XML file
  - OWL file
- **Graphical interfaces**
  - EVS Terminology server
  - Protégé

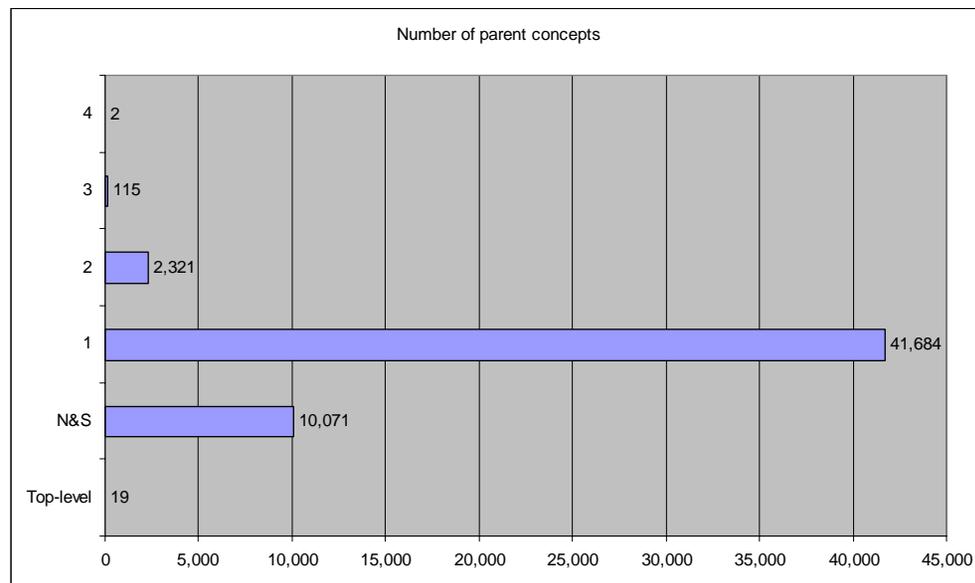
- **Version**
  - 06.09d
- **Size**
  - Concepts
    - Over 54,000 concepts
    - Over 150,000 concept names
    - 20 top-level categories
  - Relations
    - Isa
      - Over 56,000 relations
    - Associative
      - 87 relationship types
      - Over 56,000 relations

- **Documentation**
  - EVS chapter from the caCORE 3.1 documentation
- **Publications**
  - Several papers by the NCIT development team
    - In journals (JBI)
    - In conference proceedings (AMIA)
  - Two papers by outside research teams critiquing the NCIT
    - Ceusters et al., Methods 2005
    - Kumar et al., AIME 2005

- **Some programming involved**
  - Perl scripts
  - Parse the OWL file
  - Count the presence of properties
  - Extract values for some properties
  - Export to a matrix (concept x property) to enable specific checking
- **Statistical analysis of the concept x property matrix**

- **Evaluation of the review criteria**
  - Many criteria can be (easily) translated into an operational definition and checked
  - Some issues
    - Redundancy
    - Clarity
- **Compliance of the NCIT with the review criteria**
  - Mostly compliant
  - Difficult to evaluate thoroughly for some criteria because of its size

- **Fully compliant (or almost)**
  - Statement of purpose
    - Clearly stated in multiple publications
  - Concept orientation
    - Synonymy is explicitly represented
    - Each concept has one meaning and only one
    - Indirect evidence: 1:1 mapping to other concept-oriented terminologies
    - **Not all mappings to UMLS and NCI Meta are 1:1 (few exceptions)**
  - Concept permanence
    - History mechanism
  - Nonsemantic identifiers
  - Polyhierarchy
  - Accessibility and distribution
  - Intellectual property considerations
  - Community acceptance



- **Fully compliant (or almost) – continued**
  - Explicitness of relations
    - All relations have an explicit relationship
    - All relationships have a textual definition
  - Rejection of NEC
    - Not elsewhere classified (relative definition)
  - Multiple granularities
  - Consistent views
  - Graceful evolution
    - Well-documented history mechanism
    - Except for undocumented editorial policies
  - Maintenance and extension
    - Indirect evidence: growth in successive versions
    - Except for lack of documented editorial policy

- **Moderately compliant**
  - Formal definitions
    - Not always necessary and sufficient conditions
  - Composite concepts
    - No built-in mechanism for post-coordination
  - Documentation
    - No one place with the documentation
    - No real user manual  
(besides the EVS chapter from the caCORE 3.1 documentation)
  - Textual definitions
    - 62% of the concepts have a textual definition
    - Quality difficult to assess

- **Non-compliant**
  - Editorial policy
    - No documented editorial policy
    - Procedures for identifying and filling gaps
  - Context representation
    - No usage notes
    - But semantic categorization (semantic type)

- **Non evaluated**
  - Scope
  - Reporting requirements

- **An operational definition could be found for a large number of criteria**
  - E.g.,
    - Count the number of parents per concept
    - Check the presence of a value in the field *definition*
  - Scalability
  - Reproducibility
  - OWL formalism helps
    - Standard representation
    - Easily parseable
- **Some criteria are difficult to assess**
  - Internal consistency
  - Presence vs. quality of textual definitions
  - Indirect assessment of concept orientation (possible only in relation to other concept-oriented terminologies)

- **Redundancy issues**
  - Some criteria are redundant
    - E.g., criteria about revisions and extensions
- **Clarity issues**
  - Some criteria are not clearly defined
    - E.g., Atomic vs. composite concepts
- **Reinterpretation of some criteria**
  - E.g.,
    - Content coverage
    - Multiple views
  - Original criteria formulation vs. reformulation

- **Unresolved issues**
  - Demonstrating noncompliance vs. demonstrating compliance
  - Relative weight of the criteria
  - No surrogate for other kinds of evaluation
    - E.g., Ceusters et al.

- **Criteria**
  - Most of them are valid
  - Clarity issues with some
  - Redundancy in the list
  - Difficulty to operationalize some criteria
- **NCIT is compliant with most criteria**
  - Fully with most
  - Partly with some
  - Need to document the editorial policy



**caBIG**<sup>™</sup> cancer Biomedical  
Informatics Grid <sup>™</sup>

*An Initiative of the National Cancer Institute*

Contact: [olivier@nlm.nih.gov](mailto:olivier@nlm.nih.gov)

Web: [mor.nlm.nih.gov](http://mor.nlm.nih.gov)



*Olivier Bodenreider*

Lister Hill National Center  
for Biomedical Communications  
Bethesda, Maryland - USA