

The background image shows an outdoor setting, possibly a park or waterfront area. There are several trees, some with bare branches and some with green leaves. In the distance, there is a building with a traditional Chinese-style roof. The sky is overcast. The text is overlaid on this image.

Federating terminological  
resources from the BiomedGT  
perspective.

(or the redemption of “Ontology”)

Harold Solbrig  
Technical Specialist  
Mayo Clinic

September 4, 2008

# Outline

- Purpose of this Presentation
- Brief introduction to the NCI Thesaurus
- Ceusters' critique of the Thesaurus
- Evaluation, Recommendations and New Approach
- Lessons Learned

# Outline

- **Purpose of this Presentation**
- Brief introduction to the NCI Thesaurus
- Ceusters' critique of the Thesaurus
- Evaluation, Recommendations and New Approach
- Lessons Learned

# Berners Lee

The Semantic Web gives you an especially powerful form of data integration. It does this by using URIs, and by *connecting your raw data (in databases, XML documents, etc) **to a model of the real things*** (like customers, products, etc.) which your business uses. Any system which does one without the other won't get the effect of *allowing data from one application to be used in unexpected new ways by other applications*. And any system which does the same thing but doesn't use the common standards isn't going to be compatible, and so isn't going to be part of it.

# Berners Lee

The Semantic Web is not about the meaning of English documents. It's not about marking up existing HTML documents to let a computer understand what they say. It's not about the artificial intelligence areas of machine learning or natural language understanding -- they use the word semantics with a different meaning. It is about the data which currently is in relational databases, XML documents, spreadsheets, and proprietary format data files, and all of which would be useful to have access to as one huge database.

# Purpose

“How do you establish semantic equivalence? A standard criteria set, but it could be different for different domains. In aircraft, tolerances are strict, on ground, different rule sets apply.”

“Three domains but the same object – to establish mappings, business rules are different, aimings are different”

# Purpose

Thesis: Interoperable ontology models, interoperable ontology is only part of the task

Quality in ontology content is equally important – this can't just left as an "exercise for the users". Ontology construction process, purpose, etc. must be a controlled and regulation discipline in its own right.

# Purpose

This talk focuses on Level 3 – the Ontology instance level.

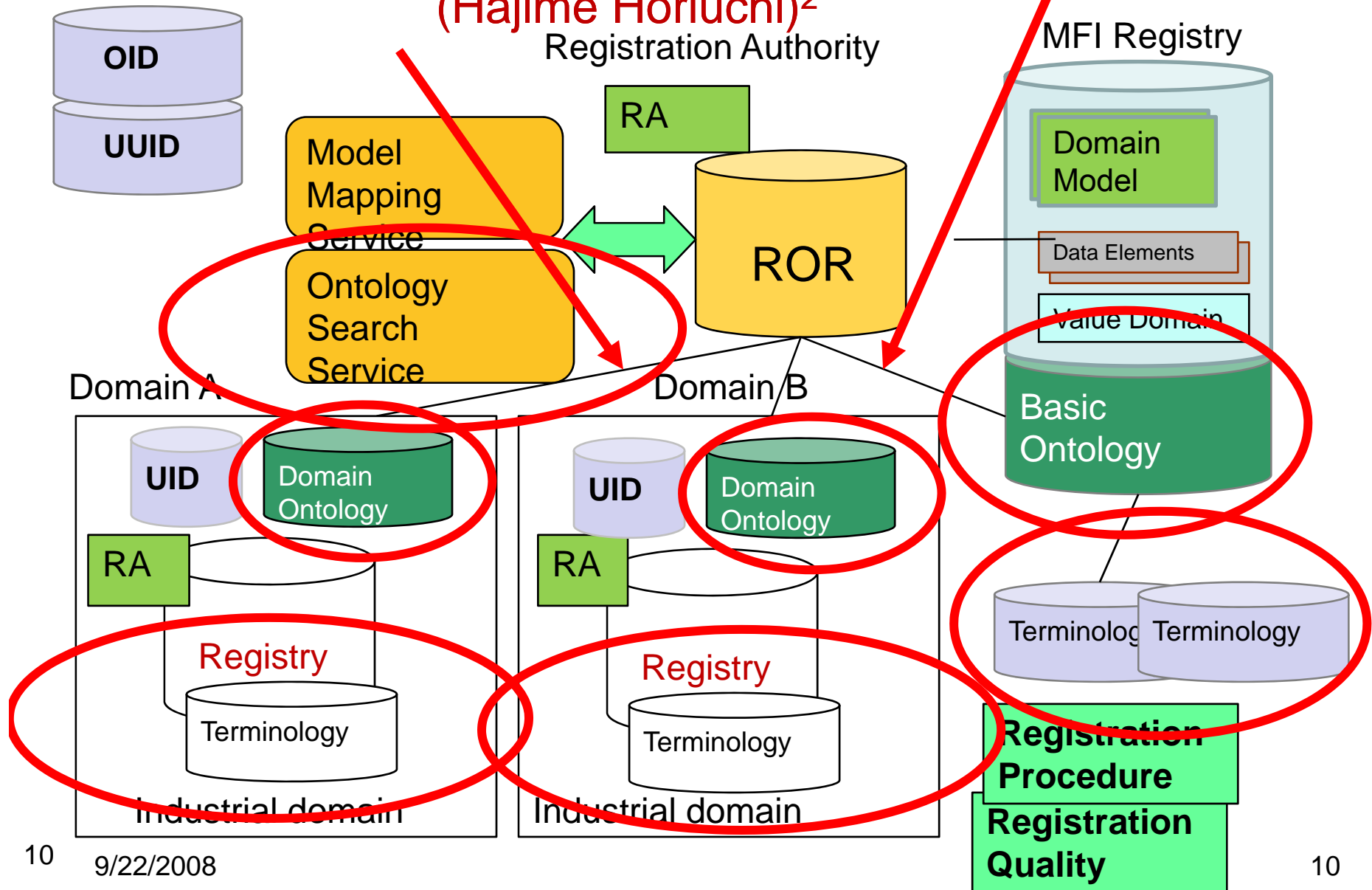
# Purpose

**Ontology:** “A specification of the kinds of entities that exist or may exist in some domain or subject area.”<sup>1</sup>

- Information models describe data that identify, document and characterize entities in some domain of subject area.
- Ontology, if done right, provides the foundation for information (model) interoperability.
- This talk is about some steps that the National Cancer Institute is taking to try to do ontology right (or at least better).

# Ontology and Terminology

(Hajime Horiuchi)<sup>2</sup>



# Outline

- Purpose of this Presentation
- **Brief introduction to the NCI Thesaurus**
- Ceusters' critique of the Thesaurus
- Evaluation, Recommendations and New Approach
- Lessons Learned

# The NCI Thesaurus

- Started in 1999
- A major effort to integrate molecular and clinical cancer-related information within a unified biomedical informatics framework, with controlled terminology as its foundational layer.<sup>3</sup>
- “Designed to meet the growing need for accurate, comprehensive, and shared terminology, covering topics including: cancers, findings, drugs, therapies, anatomy, genes, pathways, cellular and subcellular processes, proteins, and experimental organisms.”<sup>3</sup>

# The NCI Thesaurus









































- “... a partial model of how these things relate to each other, responding to actual user needs and implemented in a deductive logic framework ...”<sup>3</sup>

# The NCI Thesaurus

- As of 2008-08:
  - ~ 64,000 entries
    - ~58, 000 “primitive” (necessary, but not sufficient)
  - ~ 190,000 terms
  - ~ 46,000 textual definitions
  - ~ 48,000 asserted relations
- Created using Apelon TDE editor
  - OWL DL Version available as well
- Part of the caCORE infrastructure stack  
<http://ncicb.nci.nih.gov/NCICB/infrastructure>
- Open caBIO API and browsed via the web  
<http://nciterms.nci.nih.gov>

# NCI Thesaurus Top Node

## NCI\_Thesaurus Taxonomy

-   Abnormal Cell
-   Activity
-   Anatomic Structure, System, or Substance
-   Biochemical Pathway
-   Biological Process
-   Chemotherapy Regimen or Agent Combination
-   Conceptual Entity
-   Diagnostic, Therapeutic, and Research Equipment
-   Diagnostic or Prognostic Factor
-   Disease, Disorder or Finding
-   [Drug, Food, Chemical or Biomedical Material](#)
-   Experimental Organism Anatomical Concept
-   Experimental Organism Diagnosis
-   Gene
-   Gene Product
-   Molecular Abnormality
-   NCI Administrative Concept
-   Organism
-   Property or Attribute
-   Retired Concept

# NCI Thesaurus Sample Page

The screenshot shows the NCI Terminology Browser interface. The browser title is "NCI Terminology Browser - Mozilla Firefox". The address bar shows the URL: [http://nciterns.nci.nih.gov/NCIBrowser/GetRoleAndProperty.do?bookmarktag=1&conceptname=Prostate\\_Angiosarcoma](http://nciterns.nci.nih.gov/NCIBrowser/GetRoleAndProperty.do?bookmarktag=1&conceptname=Prostate_Angiosarcoma). The page header includes the National Cancer Institute logo and the text "National Cancer Institute" and "U.S. National Institutes of Health". The main navigation bar contains links: [HELP](#), [RESULTS](#), [CUSTOMIZE](#), [ABOUT](#), [BROWSE HIERARCHY](#), and [LOGOUT](#). The left sidebar has a search section with "Quick Search" and "Advanced Search" tabs. The "Quick Search" tab is active, showing "Max Results: 25" and a search input field containing "prostate". Below the search field, it says "Concepts visited (during this session): Prostate Angiosarcoma". Underneath is a "QUICK LINKS" section with links for "EVS HOME", "NCICB HOME", "NCI HOME", and "KNOWN ISSUES". The main content area is titled "Concept Details" and "Bookmark this page". It features a red star icon and the text "Prostate Angiosarcoma". To the right of this text are links for "Printable Page", "History", and "Graph". Below this is an "Identifiers:" section with a table:

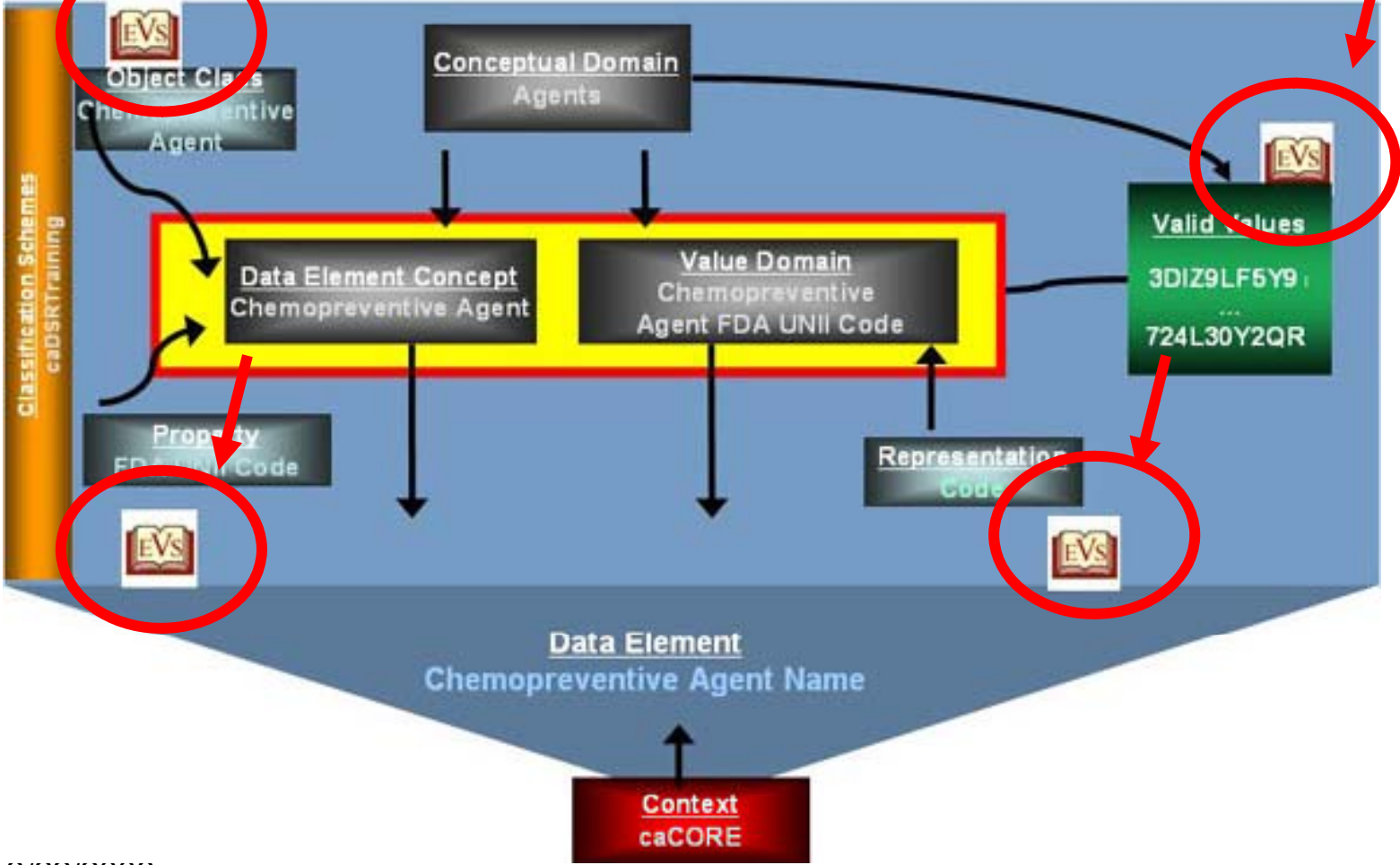
name	Prostate_Angiosarcoma
code	C5528

Below the identifiers is a "Relationships to other concepts:" section. It lists various relationships, each with a red star icon and a red document icon:

- Disease\_Excludes\_Finding: Indolent Clinical Course
- Disease\_Has\_Associated\_Anatomic\_Site: Cardiovascular System
- Disease\_Has\_Primary\_Anatomic\_Site: Blood Vessel
- Disease\_Has\_Associated\_Anatomic\_Site: Connective and Soft Tissue
- Disease\_Has\_Finding: Multinodular Mass
- Disease\_Excludes\_Normal\_Tissue\_Origin: Epithelial Tissue
- Disease\_Excludes\_Primary\_Anatomic\_Site: Bone
- Disease\_Has\_Primary\_Anatomic\_Site: Vascular System
- Disease\_Has\_Finding: Vascular Channel Formation
- Disease\_Has\_Associated\_Anatomic\_Site: Vascular System
- Disease\_Excludes\_Normal\_Cell\_Origin: Neuron, Neuroepithelial Cell, and Supporting Cell of the Nervous System

# NCI Thesaurus

caDSR Implementation – ISO 11179



# Outline

- Purpose of this Presentation
- Brief introduction to the NCI Thesaurus
- **Ceusters' critique of the Thesaurus**
- Evaluation, Recommendations and New Approach
- Lessons Learned

# Ceuster's Critique

## A Terminological and Ontological Analysis of the NCI Thesaurus

- Ceusters / Smith / Goldberg, *Methods of Information in Medicine*, 44 (2005), 498-507

# Ceusters' Critique

Version 04.08b of the NCI Thesaurus suffers from the same broad range of problems that have been observed in other biomedical terminologies. For its further development, we recommend the use of a more principled approach that allows the Thesaurus to be tested not just for internal consistency but also for its degree of correspondence to that part of reality which it is designed to represent.

# Ceuster's Critique (Excerpts)

We have measured the NCIT's qualities along three lines:

- 1) Conformity with relevant terminological standards put forward by ISO
- 2) Ontological principles
- 3) Appropriateness of OWL as a knowledge exchange format.

# Ceusters' Critique ISO Conformance

- ISO 704:2000 Terminology work – Principles and methods
- ISO 860:1996 Terminology work – Harmonization of concepts and terms
- ISO 1087-1:2000 Terminology work – Vocabulary – Part 1: Theory and application
- ISO 15188:2001 Project management guidelines for terminology standardization
- ISO 1087-2:2000 Terminology work – Vocabulary – Part 2: Computer applications
- ISO 12620:1999 Computer applications in terminology – Data Categories
- ISO 16642:2003 Computer applications in terminology – Terminological markup framework
- ISO 2788:1986 Documentation – Guidelines for the establishment and development of monolingual thesauri

# Ceusters' Critique ISO Conformance Findings

...very many NCI concepts would benefit from a clear definition, since it is often hard to grasp what they stand for in reality and browsing the hierarchy often gives no further clues.

▪

# Ceusters' Critique ISO Conformance Findings

**Ontology:** *“The word ontology has a long history in philosophy, in which it refers to the study of being as such. In information science, an ontology is an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships among them.”*

# Ceusters' Critique ISO Conformance Findings

## **Antitubercular Agent:**

*Drugs used in the treatment of tuberculosis. They are divided into two main classes: 'first-line' agents, those with the greatest efficacy and acceptable degrees of toxicity used successfully in the great majority of cases; and 'second-line' drugs used in drug-resistant cases or those in which some other patient-related condition has compromised the effectiveness of primary therapy.*

# Ceusters' Critique ISO Conformance Findings

***Disease Progression*** enjoys three definitions:

- (1) *Cancer that continues to grow or spread.*
- (2) *Increase in the size of a tumor or spread of cancer in the body.*
- (3) *The worsening of a disease over time.*

# Cuesters' Critique Ontological Principles

## The BFO Ontology

- Universals / Particulars
- Continuants / Occurrents
  - Parthood can't cross boundary
  - Is “incision” the act or the result of the act?
- Independent / Dependent continuant

# Cuesters' Critique Ontological Principles

## Top level category “Conceptual Entities”

...they are mostly not abstract at all, but rather highly concrete, including: *action, change, color, death, event, fluid, injection, temperature* (and many others in similar vein).

Moreover the definition itself contravenes the principle mentioned already above to the effect that definitions should define *concepts* and not *words*. (We hasten to point out that the NCIT is not by any means alone in having troubles with the weasel phrase “conceptual entity”.)

# Cuesters' Critique Ontological Principles

- **Biological Function / Biological Process:**
- **Anatomic Structure, System, or Substance / Anatomic Structures and Systems:**
- **Organism / Organisms:**

# Cuesters' Critique Ontological Principles

The most fundamental problem for the NCIT, however, is the unprincipled way in which its class hierarchy is built up. For this means that it ignores the basic ontological distinctions between continuants and occurrents on the one hand, and dependent and independent entities on the other.

# Ceusters Critique OWL Usage

Some vs. All

- *Human-Organ HasLocation allValuesFrom Human-Body-Region*
- *Primitive vs. Defined (most primitive)*
- *Formal vs. “Verbal” definition*

# Outline

- Purpose of this Presentation
- Brief introduction to the NCI Thesaurus
- Ceusters' critique of the Thesaurus
- **Evaluation, Recommendations and New Approach**
- Lessons Learned

# 2007 Evaluation

Contract Issued in 2006 – primary goal was to start with Ceusters' Critique and create *and demonstrate* recommendations on how to correct the various issues

Primary work done by Apelon (Harold Solbrig, Frank Din) 2006-2007

# 2007 Evaluation Deliverables (partial)

- Summarize NCBO as relates to ontology content and federation
- Review and summarize current literature and standards relevant to the quality and federation aspects of ontologies
- Use the results of this review to create an Ontological Criteria Report that lists and describes the goals that we believe that the NCI Thesaurus must meet,
  - will include requirements for semantic, ontological and terminological quality and consistency
  - prioritized into high, medium and low categories to reflect the order that we believe that the issues need to be addressed.

# Evaluation

## Elements considered vital

Increase the ability to utilize the resources (SME's and curators) available to NCI *today* in a way that

- Focuses on their primary area of expertise
- Is as lossless as possible

# Elements considered vital Focus on area of Expertise

## Capturing *subject matter* knowledge

- Focus on capturing *information*
  - Not on the exercise of technology skills
- Capture the complete picture
  - Record even if the tooling doesn't currently support it
  - Don't waste time trying to make something fit that the current tools can't deal with
  - Minimize underlying assumptions – SME's, by nature, have a tacit contextual base. This needs to be made as explicit as possible by asking questions, requiring complete documentation.

# Elements considered vital

## Minimize Loss

- Proceed one step at a time, but try to take as few steps backwards as practical
  - Record the information while you've got the SME at hand
  - Get sufficient detail that *non-SME*'s should be able to (re-)encode the information into different computational paradigms without major re-recording expenses.

# NCIt (2006)

- 82% of the entries are marked as *primitive*
  - they define some necessary conditions for membership, but not the sufficient conditions
    - This is not a *bad* thing – some would argue that the number is suspiciously low.
    - It does mean, however, that the entries are formally *described*, not *fully defined*
      - Which means that we are at least partially dependent on the textual terms, definitions, synonyms to understand the author's intent

# NCIt (2006)

- 39% of the primitive entries currently lack definitions
  - This is sometimes inappropriate, as there are some things that shouldn't be defined (e.g. chromosome locations, drug packaging, etc)
  - There are still a lot, however, that *should* have been defined, and after-the-fact analysis doesn't always work
  - Formalization requires understanding of intent
    - not just names.

# NCIt (2006)

- Where formal definitions do exist, there isn't a clear partition between *defining* and *assertional* knowledge.

– ABI2\_wt\_Allele isA ABI2\_Gene that

- Is found in organism "Human"
- Is locate at \_2q33
- Has location \_2\_20401886668-204117837
- Is (can be?) an element in the Actin\_Branching\_Pathway

Defining

Assertional

# NCIt (2006)

- Attempts have been made to formalize semantics that are unavailable to the current tools:
  - Allele\_Ceases\_Function\_In\_Pathway
  - Disease\_Excludes\_Abnormal\_Cell
  - Disease\_May\_Have\_Abnormal\_Cell
- This information is useful, but the encoding masks the original intent.

# Evaluation Recommendations

1. Traceability and reproducibility
2. Rules and techniques for consistent and precise definitions, designations and annotations
3. Semi-automated methodologies to validate and align definitions with corresponding formal relationship structures
4. Adopt a meta-model for the ontological resources

# General Goals (continued)

5. Differentiate thesaurus and ontology
6. Subdivide the thesaurus by subject field / discipline
7. Organize the ontology horizontally by subject field and vertically by ontological “meta-type”
8. Adopt faceted / dimensional classification schemes

# Goal 1: Traceability and reproducibility

Preserve links to the source material, process, context, etc.

- Reference documents
- Context of request
- Who did the entry, with what tools

# Goal 1

## Motivations

**Cost** - A significant portion of the cost of assembling a terminological resource is in gathering the information to be organized.

**Verifiability** - Preserving the input information allows validation, verification and subsequent processing (ala. Columbia Semantic Net studies)

# Goal 1

## Proposed Solution(s)

- Semantic MediaWiki – record dialog, history, audit trail
- “Semantic SVN”
- Tools for consistent and accurate references
  - URI’s for documents and other resources
  - Online dereferencing where possible

Main Page - LexWiki - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://biomedgt.nci.nih.gov/index.php/Main\_Page

main page discussion view source history

# Main Page

## BiomedGT Collaborative Ontology Development Wiki

(Alpha 2 Release Jul 15 2008)

### About BiomedGT

The Biomedical Grid Terminology ([BiomedGT](#)) is an open, collaboratively developed terminology for translational research. BiomedGT bu orientation, description logic, and public accessibility. While the current terminology has been seeded with NCI Thesaurus content, it is bei is to evolve BiomedGT into a set of federated sub-terminologies, with content maintained by experts in the relevant research communities.

**NOTE:** The original biomedgt.org site will be revamped to carry a current image of the NCI Thesaurus. It can still be found at <http://biomedg>

### About This Wiki

The BiomedGT Wiki is a new tool that enables groups of domain experts to collaboratively develop and maintain terminologies, including Bi terminologies. This wiki is being developed by the [National Cancer Institute Center for Bioinformatics](#), [Apelon, Inc.](#), and the [May](#) your help in developing terminology in your specialty area and in fine tuning the use of the wiki as a collaborative terminology development 1

The terminology content developed in this wiki will be transferred to and integrated into the BiomedGT terminology and/or converted into Le: Cancer Bioinformatics Grid (caBIG) community and other interested users. After integration, the content in the wiki will be refreshed for coll

### Becoming a Collaborator

You are welcome to browse this site and search for terminology. If you want to contribute content, you can register as a collaborator by foll [Collaborators](#) page.

### Subscribing to Our Listserv

You can subscribe to a listserv and receive announcements about changes to the wiki, information about support and maintenance, and otl [listserv@list.nih.gov](mailto:listserv@list.nih.gov). In the body of the message, include the following string:

search

Go Search

navigation

- Main Page
- Current events
- Recent changes
- TODO List
- Help

terminologies

- BiomedGT
- CTCAE Subject Terms
- NanoTech Domain
- CRCH Nutrition Terms
- Biospecimen
- Semantic Type
- CTCAE\_Category

content search

- Query ConceptCode
- Query Preferred Name

workflow


[view source](#)

## Category: BGT Oncogene K Ras(B16939)

**BGT\_Oncogene\_K\_Ras(B16939)**

### Lexical

**Concept Code:** B16939 (O)

**Preferred Name:** Oncogene K-Ras

**Coding Scheme:** [BGT \(01.01\)](#)

**Synonym:** **Oncogene K-Ras** (PT) (Source: NCI) / **Oncogene, K-Ras-2** (SY) (Source: NCI) / **Ki-RAS** (SY) (Source: NCI) / **KRAS2** (SY) (Source: NCI) / **Rat Sarcoma 2 Viral Oncogene Homolog** (SY) (Source: NCI) / **Oncogene, K-Ras** (SY) (Source: NCI) / **KRAS** (SY) (Source: NCI)

**Definition:** Human Oncogene K-Ras is a mutated variant of KRAS2 Gene, which encodes two alternative isoforms of monomeric p21 K-R signal transduction that alternates between inactive GDP-bound and active GTP-bound forms. K-Ras is activated by a guanine nucleotide-Mitogen-stimulated RAS stabilizes MYC protein and enhances MYC accumulation by the RAS/RAF/MAPK pathway, which appears to in a variety of human tumors, mutations of specific amino acids activate RAS to transform cells. KRAS is involved in malignancy much more function. (Source: NCI)

**URI:** urn:oid:2.16.840.1.113883.3.26.1.2:B16939

### Properties

**GenBank\_Accession\_Number:** M54968

**Locus\_ID:** 3845

**OMIM\_Number:** 190070

**Semantic\_Type:** Gene or Genome

**UMLS\_CUI:** C0022457

### Associations

**Parent:** [RAS\\_Family\\_Oncogene](#)

Every instance of [Oncogene\\_K\\_Ras Gene\\_Associated\\_With\\_Disease](#) at least one instance of [Malignant\\_Colorectal\\_Neoplasm](#).

Every instance of [Oncogene\\_K\\_Ras Gene\\_Associated\\_With\\_Disease](#) at least one instance of [Colorectal\\_Adenoma](#).

search




navigation

- [Main Page](#)
- [Current events](#)
- [Recent changes](#)
- [TODO List](#)
- [Help](#)

terminologies

- [BiomedGT](#)
- [CTCAE Subject Terms](#)
- [NanoTech Domain](#)
- [CRCH Nutrition Terms](#)
- [Biospecimen](#)
- [Semantic Type](#)
- [CTCAE\\_Category](#)

content search

- [Query ConceptCode](#)
- [Query Preferred Name](#)

workflow

- [Create / Update Package](#)
- [Bug Reporting](#)

# Goal 2: Definitions, Designations and Annotations

- Derived from ISO documents
  - 704
  - 1087
  - Z39.19
  - ...
- Despite ISO advice to the contrary, seems to work best *post* (or concurrent with) ontology alignment
- Loosely structured list available in spreadsheet

# Goal 2: Definitions

## Definitions:

- Must exist
- Intensional subject, copula, predicate form
- Describe the concept, not the term
- Describe only one concept
- Appropriate to subject field
- Substitution principle
- Avoid non-negative
- Non-circular
- ...

# Goal 2: Designations

- Direct word order (issues of synonym permutations)
- Accepted by subject specialists
- Designations vs. “names” (!)
- Singular and plural usage
- Interchangeable synonyms
- Distinguish quasi synonyms
- Established usage
- Consistency
- Count vs. non-count nouns
- Other term relationships
- ...

# Goal 2: Implementation

- Assemble check-list for each area
- Refine during subsequent phases

Goal 3: Semi-automated methodologies to validate and align definitions with corresponding formal relationship structures

# Goal 3: Example

Category: NCI:Apoptosis Regulation Gene

---

Contents [\[show\]](#)

## Textual Characteristics

---

### Definitions

#### Apoptosis\_Regulation\_Gene

Apoptosis Regulation Genes encode Apoptosis Regulator proteins, which either promote or impede the initiation, progress, or rate of apoptosis.

**Synonyms:** Apoptosis Regulation Gene

### Properties

#### Other Properties

**Gene\_Encodes\_Product:** Apoptosis Regulator

**NCI\_META\_CUI:** CL026945

## Formal Characteristics

---

PRIMITIVE

**Kind:** Gene\_Kind

### Defining Roles

some **Gene\_Plays\_Role\_In\_Process** Apoptosis

9/22/2008

54

# Goal 3: Example

## Editing Category talk:Apoptosis Regulation Gene:NCI

---

You've followed a link to a page that doesn't exist yet. To create the page, start typing in the box below (see the [help page](#) for more info). If you are here by mistake, just click your browser's **back** button.



== Proposed Definition ==

A [[[:Category:NCI:Regulatory\_Gene]]

that [[NCI:Gene\_Product\_Encoded\_By\_Gene::Category:NCI:Apoptosis\_Regulator]] can encode an apoptosis regulation protein]

# Goal 3: Example

Category talk:Apoptosis Regulation Gene:NCI

---

Proposed Definition

---

A `Category:NCI:Regulatory_Gene` that can encode an apoptosis regulation protein

---

# Goal 3: Example

## Formal Characteristics

---

PRIMITIVE

Kind: Gene\_Kind

## Defining Roles

some Gene\_Plays\_Role\_In\_Process Apoptosis

Semantic\_Type Gene or Genome

Category talk:Apoptosis Regulation Gene:NCI

---

## Proposed Definition

---

A Category:NCI:Regulatory\_Gene that can encode an apoptosis regulation protein

---

Category: NCI:Gene

# Goal 3 - Implementation

Tooling is being put into place to automatically generate definitions and paraphrase assertions

# Goal 3 Prototype

The image displays a software interface for a class hierarchy. On the left, a tree view shows the following classes under 'Anatomic\_Structure\_System\_or\_Substance':

- Body\_Fluid\_or\_Substance
  - Aqueous\_Humor
  - Bile\_Salt
  - Blood
  - Cerebrospinal\_Fluid
  - Exocrine\_Gland\_Fluid\_or\_Secretion
  - Female\_Genital\_System\_Fluid\_or\_Secretion
  - Fibrin
  - Gastrointestinal\_Fluid\_or\_Secretion
  - Lymph
  - Male\_Genital\_System\_Fluid\_or\_Secretion
  - Mucus
  - Otolymph
  - Plasma
  - Respiratory\_System\_Fluid\_or\_Secretion
  - Serum
  - Skin\_Fluid\_or\_Secretion
  - Synovial\_Fluid
  - Urine
  - Vitreous\_Humor

The right pane shows the details for the 'Blood' class:

- Semantic\_Type:** "Body Substance"
- Structured\_Definition:** "Blood is a body fluid or substance that is a physical part of some hematopoietic system." Below this, a note states: "Every instance of blood is also an instance of a body fluid or substance where the body fluid or substance is a physical part of some hematopoietic system."
- UMLS\_CUI:** "C0005767"
- Unity:** "1"

The 'Class Description: Blood' section includes:

- Equivalent classes:** (empty)
- Superclasses:**
  - Body\_Fluid\_or\_Substance
  - Anatomic\_Structure\_Is\_Physical\_Part\_Of\_some\_Hematopoietic\_System

# Goal 3 Prototype

ALT\_DEFINITION

```
"<def-source>MSH2003_2003_05_12</def-source><def-definition>Disorders of the blood and blood forming tissues.</def definition><Definition_Review_Date>060127</Definition_Review_Date><Definition_Reviewer_Name>DEFAULT_Review</Definition_Reviewer_Name>"
```

Superclasses **Defining**

- Hematopoietic\_System\_Disorder
- Disease\_Has\_Normal\_Cell\_Origin\_only\_Hematopoietic\_Cell
- Disease\_Has\_Normal\_Tissue\_Origin\_only\_Hematopoietic\_and\_Lymphoid\_Tissue
- Disease\_Has\_Primary\_Anatomic\_Site\_only\_Hematopoietic\_and\_Lymphatic\_System

Structured\_Text

...

Structured\_Definition

"A hematologic and lymphocytic disorder is a hematopoietic and lymphoid system disorder that has a normal cell origin of only a hematopoietic and lymphoid cell and has a normal tissue origin of only a hematopoietic and lymphoid tissue and has a primary anatomic site of only a hematopoietic and lymphatic system.

Every instance of a hematologic and lymphocytic disorder is also an instance of a hematopoietic and lymphoid system disorder where the hematopoietic and lymphoid system disorder has a normal cell origin of only a hematopoietic and lymphoid cell and the hematopoietic and lymphoid system disorder also has a normal tissue origin of only a hematopoietic and lymphoid tissue and lymphoid cell and the hematopoietic and lymphoid system disorder also has a primary anatomic site of only a hematopoietic and lymphatic system."

# Goal 4: Adopt a meta-model for the ontological resources

# Goal 4

- Meta-model includes
  - Top (Upper) Level Ontology
  - OntoClean Methodology
  - Model for associations

# Goal 4

## Motivations

- **Quality** – The alignment process asks the sort of questions necessary to clarify and document intent (*near term*)
- **Federation/Integration** – Alignment *should* aid in integration/federation/re-use (*longer term*)
- **Politics** – Alignment is required to play in some communities

# Goal 4 Implementation

- Evaluated
  - BFO
  - Dolce
  - TopBio (Rector)
  - UMLS Semantic Net
- Recommended UMLS Semantic Net
  - *If* National Library of Medicine would update

# Goal 4 Implementation

- Ended up selecting BFO
  - Second most useful
  - More closely aligned with other biomedical resources
  - BioTop (Shulz) is emerging as viable middle layer

# Goal 5: Differentiate Thesaurus and Ontology

# Goal 5: Terminology

## (as used in this presentation)

**Thesaurus** – A **controlled vocabulary** arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators.

**Controlled Vocabulary** - A list of terms that have been enumerated explicitly. This list is controlled by and is available from a controlled vocabulary registration authority.

**Term** - One or more words designating a **concept**.

# Goal 5: Terminology

(As used in this presentation)

**Ontology(1)** - a **Concept System** that incorporates a declarative definition of a universe of discourse that models entities and the relationships that hold among them utilizing a formalized structure which incorporates standardized terminology.

**Concept System** - system of concepts set of **concepts** structured according to the relations among them

# Goal 5: Terminology

(As used in this presentation)

**Ontology(2)** - The study of the nature of being, reality, and substance.

While modern usage includes the plural form, the core characteristics of this still hold in the biomedical domain.

*An **Ontology*** is supposed to be a model of reality.

# Goal 5: Thesauri

- Represents *terminology* – the special language of a discipline, scientific field, trade, etc.
- Purpose is indexing and information retrieval
- Focus is on (a) language and (b) subject field
- Crosses multiple scientific disciplines

# Goal 5: Thesauri (continued)

- Organizational structure is pragmatic
  - Abnormal cells can be a different category than normal cells
  - “Food and Food Product” can be a single category
  - “Protein Organized By Function” makes perfect sense
- There can be many overlapping thesauri based on pragmatic issues
  - Varying by content, detail, etc.
- Thesauri *reference* ontology

# Goal 5: Ontology

- Represents some aspect of “reality”
  - Terminology, at best, serves a secondary role
- Purpose is to describe what is known.
  - Fractal
  - Noisy – unacceptable from indexing and search perspective
- Tends to be partitioned by sub-specialty and perspective (orthogonality)

# Goal 5: Thesauri / Ontology

- Both entities are necessary
  - But they *are* different!
- Thesauri and terminologies provide the human/machine/human interface
- Ontology provides
  - A formal anchor (concept system++)
  - Potential for reasoning
  - Potential for crossing fields and disciplines

# What a Thesaurus Won't Do

1. Provide an exhaustive perspective on a given scientific field
  - Too much depth or breadth just adds noise
  - Multiple fields have to be touched
2. Represent pure “reality”
  - Thesauri are descriptive, *not* revisionary
  - Nodes / classifications / generics are all tools of a Thesaurus

# What Ontology won't (or shouldn't) do

- Provide a useful (high precision and recall) set of indices for a document corpus
- Navigational aids - “drug organized by function”, “see also”, etc.
- Provide translations, sense specific usages, near synonyms, etc.
- Define the terminology of a domain or specialty
- Provide reference terms for nonsense

# One view of terminological resources



# Goal 5

- Three categories
  - Things that belong in a general dictionary
  - Things that belong in a thesaurus
    - Reference things in a general dictionary
    - Reference things in an ontology
  - Things that belong in an ontology
- *Definition* is the common characteristic
  - (which is the why of goals 1-4!)

# Goal 5 Implementation

- NCIt Preliminary Categorization complete
- Partition underway
  - Thesaurus (Navigational and Category Nodes)
  - Ontology
  - Words

# Implementation of Goal 5

- Rules of reference established
  - Ontology nodes must stand alone – may not reference navigational or word
  - Navigational nodes must be fully defined
  - Word references and tooling beginning to be deployed
    - WordNet as first entry
    - Possibility to revitalize WordNet hosting

# Implementation of Goal 5

- BiomedGT
  - Multiple namespaces
    - Ontology
    - Navigational and Category namespaces
    - Words
  - Model of external reference being produced
- Proving to be quite valuable
  - Need to do reproducibility studies

# Goal 6: Subdivide thesaurus by subject field / discipline

# Goal 6

A thesaurus represents the *terminology* – the special language - of a discipline, scientific field, trade, etc.

- Goals are to match the level of precision and detail required by the specific user community
- Well understood and unambiguous *within the discipline*

# Goal 6

- It is the thesaurus *for a discipline* that provides the sort of granularity (and terminology) that allows subsetting on ontological resources
- One big thesaurus risks becoming “nothing more than a big list of all the kinds of things that do or could exist...”

# Goal 6

Keys to accomplishment:

- Context (goal 1) – what was the purpose and community from which the term first originated
- Some sort of taxonomy or other terminological resource of communities

# Goal 6

Suggested approach – classic opportunity to use the “ground up” approach and try a self-organizing structure such as a Wiki???

Goal 7: Organize the ontology horizontally by subject field and vertically by ontological “meta type”

# Goal 7

- Ultimate goal is “orthogonal structure”
- Short term goal is the ability to reference chunks of external ontology
  - OBO resources
  - FMA
  - ...
- Thesis: Thesaurus development can be localized and regional, ontology development needs to align

# Goal 7

- Horizontal
  - Anatomy
  - Chemistry
  - Cell process
  - Physics
  - ...
- Vertical
  - Endurants / perdurants
  - (Other structures imposed by TLO *and?* MLO)

# Goal 8: Adopt faceted / dimensional classifications schemes

# Motivation

There are a relatively small number of independent structures in an ontology (or thesaurus?) (e.g. LOINC Structure)

For each structure, there can be many, many 'instances' of the structure (e.g. LOINC Tests)

It takes a whole different type of person to build the structures than the instances

Instances can be federated

# Goal 8

The image shows a screenshot of a hierarchical ontology browser. At the top, the 'Cytokine' class is selected and circled in green. Below it, the following information is displayed:

- Code: C20464
- ID: 20464
- Namespace: NCI ( Ontylog, Subscription, Read-Only, 2.0.0.0 )

The 'Cytokine' class has three main sub-branches:

- Properties** (indicated by a plus sign icon)
- Superconcepts** (indicated by a minus sign icon), containing one sub-concept: *Protein\_Organized\_by\_Function*
- Subconcepts** (indicated by a minus sign icon), containing a list of 14 sub-concepts, each represented by a blue sphere icon:
  - Cardiotrophin-1
  - Chemokine
  - Hematopoietic\_Growth\_Factor
  - Interferon
  - Interleukin
  - Leukemia\_Inhibitory\_Factor
  - Leukoregulin
  - Lymphokine
  - Oncostatin-M
  - Pre-B-Cell\_Colony-Enhancing\_Factor
  - Thymic\_Stromal\_Lymphopoietin
  - Tumor\_Necrosis\_Factor\_Family\_Protein

A large green oval encircles the 'Subconcepts' list. A mouse cursor is visible at the bottom right of the screen.

# Outline

- Purpose of this Presentation
- Brief introduction to the NCI Thesaurus
- Ceusters' critique of the Thesaurus
- Evaluation, Recommendations and New Approach
- **Lessons Learned**

# Lessons Learned

*Ontology* should be

- *Reality based - classes not concepts*
- *Designed for integration*
  - *Note that most ontology today doesn't come as OWL or CL See <http://www.cas.org>*
- *Centralized – “ontologies” is a result of perspective – not reality.*

# Lessons Learned

Ontology can be *referenced* by a variety of resources

- Thesauri
- Categorization system
- Knowledge bases
- Information models
- ...

# Lessons Learned

The value (and cost) of terminological resources is the SME

- Tooling should be designed to capture as much SME expertise as possible
- Preserve references
  - Books
  - Web resources
  - Journal articles
  - Definitions
- Use formalism to encourage detail
  - Structured definitions
  - Paraphrases
  - Uncover tacit knowledge

# Lessons Learned

- History and provenance is a central part of any terminological resource
  - What was the SME / author / coder trying to say?
  - Did they think of ...?
  - Why is this this way?
  - What were their plans

# Lessons Learned

Ontological and terminological knowledge is a community effort

- Design for community input from the bottom up
- Recognize different formats
- Enable conflicting information

# Lessons Learned

Registries and shared metamodels are crucial

- There are going to be many, many sources of ontological information
- There will be many models
  - They will vary in amount of formalism
  - They will vary in kind of formalism

# Credits

The work for this presentation was funded by the National Cancer Institute under

- GSA Contract GS-35F-0009L, *Review of NCI Thesaurus for OBO-Compliance and Training to Help NCI Achieve Compliance*

Apelon, Inc. was the primary contractor, and the work was done from Nov 2006 through March 2008.

Work is still ongoing

# References

1. Fischer, Dietrich H. *From Thesauri towards Ontologies?*  
[http://www.ipsi.fraunhofer.de/orion/pubFulltexts/Fischer\\_1998.pdf](http://www.ipsi.fraunhofer.de/orion/pubFulltexts/Fischer_1998.pdf)
2. Hajime Horiuchi, Tokyo International University. *Issues for ROR (Registry of Registries): A study on the viability of MFI standards*  
<http://metadataopenforum.org/index.php?id=34,123,0,0,1,0>
3. Sioutos, de Coronado, Haber, Hartel, Shaiu, Wright *NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information.*  
[http://www.sciencedirect.com/science?\\_ob=MIimg&\\_imagekey=B6WHD-4JGJG1C-1-7&\\_cdi=6848&\\_user=10843&\\_orig=search&\\_coverDate=03%2F15%2F2006&\\_sk=999999999&view=c&\\_alid=440260223&\\_rdoc=1&wchp=dGLzVlz-zSkWb&md5=73d754e6976625513c65038efa1a7045&ie=/sdarticle.pdf](http://www.sciencedirect.com/science?_ob=MIimg&_imagekey=B6WHD-4JGJG1C-1-7&_cdi=6848&_user=10843&_orig=search&_coverDate=03%2F15%2F2006&_sk=999999999&view=c&_alid=440260223&_rdoc=1&wchp=dGLzVlz-zSkWb&md5=73d754e6976625513c65038efa1a7045&ie=/sdarticle.pdf)
4. Ceusters, Smith, Goldberg *A Terminological and Ontological Analysis of the NCI Thesaurus*  
<http://ontology.buffalo.edu/medo/NCIT.pdf>

# References

**Berners Lee:** <http://www.consortiuminfo.org/bulletins/semanticweb.php>

**NCI Thesaurus:** <http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do>

**NCI EVS:** <http://www.cancer.gov/cancertopics/terminologyresources>

**caDSR:** <http://ncicb.nci.nih.gov/core/caDSR>

**Semantic MediaWiki:** <http://semantic-mediawiki.org/>

**BiomedGT:** [http://biomedgt.nci.nih.gov/index.php/Main\\_Page](http://biomedgt.nci.nih.gov/index.php/Main_Page)

**Basic Formal Ontology (BFO):** <http://www.ifomis.org/bfo>

**Dolce:** <http://www.loa-cnr.it/DOLCE.html>

**TopBio:** <http://www.cs.man.ac.uk/~rector/ontologies/simple-top-bio/>

**OntoClean:** <http://www.ontoclean.org/>

**BioTop:** <http://www.imbi.uni-freiburg.de/biotop/>

**UMLS Semantic Net:** <http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html>

**WordNet:** <http://wordnet.princeton.edu/>

**FMA:** <http://sig.biostr.washington.edu/projects/fm/>

**LOINC:** <http://loinc.org/>