Short course – Summer 2008
Biomedical Ontology in Practice

Biomedical Ontology in Practice

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

Objectives

- Learn about biomedical ontologies
  - History
  - Design principles, formalisms and tools
  - What are they?
  - What are they used for?
- Work with biomedical ontologies
  - Search
  - Analyze
  - Extend
  - Use for data integration

Agenda

<table>
<thead>
<tr>
<th>Monday, June 9</th>
<th>Introduction to Biomedical Ontologies</th>
<th>Design Principles, Formalisms and Tools for Biomedical Ontologies</th>
<th>Biomedical Ontologies - Content and Structure - Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, June 10</td>
<td>Interfaces to Biomedical Ontologies</td>
<td>Searching and Analyzing Biomedical Concepts</td>
<td>Contrasting Biomedical Ontologies</td>
</tr>
<tr>
<td>Wednesday, June 11</td>
<td>Critical Analysis of Biomedical Ontologies</td>
<td>Extending Biomedical Ontologies</td>
<td>Using Biomedical Ontologies for Data Integration</td>
</tr>
</tbody>
</table>

References Bio-ontology courses

- Barry Smith, U. Buffalo / NCBO
  - http://ontology.buffalo.edu/smith/Ontology_Course.html
- Stefan Schulz, U. Freiburg, Germany / KR-MED 2008 tutorial

Introduction to Biomedical Ontologies

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

- Historical perspective
- Introduction to biomedical terminologies through an example
- Biomedical terms as names for biomedical classes
- Terminological relations as a surrogate for ontological relations

Historical perspective

Why biomedical terminologies?

- To support a theory of diseases
- To classify diseases
- To support epidemiology
- To index and retrieve information
- To serve as a reference

To classify diseases (and plants)

- Carolus Linnaeus (1707-1778)
  - Genera Plantarum (1737)
  - Genera Morborum (1763)
- François Boissier de La Croix a.k.a. F. B. de Sauvages (1706-1767)
  - Methodus Foliorum (1751)
  - Nosologia Methodica (1763/68)
- William Cullen (1710-1790)
  - Synopsis Nosologiae Methodicae (1785)

To support a theory of diseases

- Hippocrates
  - Dismisses superstition
  - Four humors
    - Blood
    - Phlegm
    - Yellow bile
    - Black bile
- Thomas Sydenham (1624-1689)
  - Medical observations on the history and cure of acute diseases (1676)

From plants…
… to diseases

◆ Four categories (W. Cullen)
  - Fevers
  - Nervous disorders
  - Cachexias
  - Local diseases

“...the distinction of the genera of diseases, the distinction of the species of each, and often even that of the varieties, I hold to be a necessary foundation of every plan of physic, whether dogmatical or empirical.”
- William Cullen, Edinburgh, 1785

Synopsis Nosologica Methodicae (Cited by Chris Chute)

London Bills of Mortality

Limits of existing classifications

“The advantages of a uniform statistical nomenclature, however imperfect, are so obvious, that it is surprising no attention has been paid to its enforcement in Bills of Mortality. Each disease has, in many instances, been defined by terms of four or five syllables, and in such terms has been applied to as many different diseases; vague, ambiguous names have been employed, or complications have been passed over as diseases, which are in reality only symptoms. These nomenclature is of very poor importance in the physical sciences, and would be utterly useless in public affairs.

- William Farr


To index and retrieve information

◆ Biomedical literature
  - MEDLINE (15M citations from 4600 journals)
  - Manually indexed
◆ Medical Subject Headings (MeSH)
◆ Genome
  - Model organism databases (Fly, Mouse, Yeast, …)
  - Manually / semi-automatically curated
◆ Gene Ontology

MEDLINE and MeSH
Mouse Genome Database and GO

- To serve as a reference
  - Reference terminology/ontology
    - Universally needed
    - Developed independently of any purposes
    - Reusable by many applications
  - Examples
    - VA National Drug File (NDF)
    - Foundational Model of Anatomy (FMA)
    - SNOMED CT

Anatomy in Biomedicine

- Anatomy
  - Physiology
  - Clinical medicine
  - Biomedical literature
  - Biomedical research

Administrative terminologies

- Coding patient records
  - International Classification of Primary Care (ICPC)
  - SNOMED
  - Read Codes
- Reporting claims to health insurance companies
  - International Classification of Diseases (ICD-9 CM)
  - Healthcare Common Procedure Coding System (HCPCS)

History of Medical Ontologies

Biomedical ontology in PubMed
Introduction to biomedical terminologies through an example

Addison's disease

- Addison's disease is a rare endocrine disorder
- Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol
- For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism

Adrenal insufficiency Clinical variants

- Primary / Secondary
  - Primary: lesion of the adrenal glands themselves
  - Secondary: inadequate secretion of ACTH by the pituitary gland
- Acute / Chronic
- Isolated / Polyendocrine deficiency syndrome
Addison's disease: Symptoms

- Fatigue
- Weakness
- Low blood pressure
- Pigmentation of the skin (exposed and non-exposed parts of the body)
- ...

AD in medical vocabularies

- Synonyms: different terms
  - Addisonian syndrome
  - Bronzed disease
  - Addison melanoderma
  - Asthenia pigmentosa
  - Primary adrenal deficiency
  - Primary adrenal insufficiency
  - Primary adrenocortical insufficiency
  - Clinical variants

Internal Classification of Diseases

CHAPTER 4
Endocrine, nutritional and metabolic diseases (E00-E90)

Disorders of other endocrine glands (E20-E50)

927 Other disorders of adrenal gland
927.0 Addisonian syndrome
927.1 Addison melanoderma
927.2 Asthenia pigmentosa
927.3 Primary adrenal deficiency
927.4 Primary adrenal insufficiency
927.5 Chronic adrenocortical insufficiency

Medical Subject Headings

SNOMED CT

Biomedical terms as names for biomedical classes
Terms reflecting valid classes
- Pulmonary anthrax
- BRCA1 protein
- Coronary artery
- Coronary artery bypass
- ...
  - Non-insulin dependent diabetes mellitus
  - Non-Hodgkin lymphoma
  - Non-steroidal anti-inflammatory drugs
  - Non-opioid analgesics
  - Non-invasive medical procedure

Issues
- Multiple terms for a class
- Multiple classes for a term
- Presence of non-ontological features in terms
- Composite terms

Multiple terms for a class
- Synonymy
  - Left coronary artery
  - LCA
  - Arteria coronaria sinistra
- “Clinical synonymy” (vs. identity)
  - Abdominal swelling
  - Swollen abdomen
  - Posttransfusion hepatitis
  - Posttransfusion viral hepatitis

Multiple classes for a term
- Polysyndrome
  - Cold
  - Common cold
  - Cold temperature
  - Cold
  - Chronic Obstructive Airway Disease
- Truncated terms
  - Calcium
  - Ca++
  - Calcium measurement

Non-ontological features in terms
- Epistemological features
  - Gallbladder calculus without mention of cholecystitis
  - Diarrhea of presumed infectious origin
  - Replacement of unspecified heart valve
  - ...

Ontology vs. Epistemology
- Ontology
  - Invariants in reality
    - Classes (universals)
    - Relations between them
  - Theory of reality
- Epistemology
  - Knowledge about such entities
  - Perception of reality
  - Bone metastasis diagnosed by CT scan
  - Bone metastasis diagnosed by Tc99m bone scan
**Composite terms**

- **Sentence-like terms**
  - Several classes and their relations
  - May contain epistemological features
- **Tuberculosis of adrenal glands, tubercle bacilli not found (in sputum) by microscopy, but found by bacterial culture**

**More composite terms**

- Nontraffic accident involving being accidentally pushed from motor vehicle, except off-road motor vehicle, while in motion, not on public highway, driver of motor vehicle injured
- Determine whether the elder patient and caretaker have a functional social support network to assist the patient in performing activities of daily living and in obtaining health care, transportation, therapy, medications, community resource information, financial advice, and assistance with personal problems
- Telephone call by a physician to patient or for consultation or medical management or for coordinating medical management with other health care professionals (eg, nurses, therapists, social workers, nutritionists, pharmacists); complex or lengthy (eg, lengthy counseling session with anxious or distraught patient, detailed or prolonged discussion with family members regarding seriously ill patient, lengthy communication necessary to coordinate complex services of several different health professionals working on different

**Issues**

- Lack of explicit classificatory principle
- Underspecification of the relations
- Thesaurus relations
- Limited depth in hierarchies “by design”
1. Certain infectious and parasitic diseases
2. Neoplasms
3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4. Endocrine, nutritional, and metabolic diseases
5. Mental and behavioral disorders
6. Diseases of nervous system
7. Diseases of the eye and adnexa
8. Diseases of the ear and mastoid process
9. Diseases of the circulatory system
10. Diseases of respiratory system
11. Diseases of digestive system
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Diseases of the genitourinary system
15. Pregnancy, childbirth, and the puerperium
16. Certain conditions originating in the newborn (perinatal) period
17. Congenital malformations, deformations and chromosomal abnormalities
18. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
19. Injury, poisoning and certain other consequences of external causes
20. External causes of morbidity
21. Factors influencing health status and contact with health service

**Fully specified relations**

Viral meningitis in SNOMED CT

**Underspecification of the relations**

Diseases

Virus diseases

CNS viral diseases

Viral meningitis

**Thesaurus relations**

- Addison’s disease
  - Due to auto-immunity in 80% of the cases
  - Other causes include tuberculosis
Accidents in MeSH

Limited depth in hierarchies “by design”

- Term identifier (code) used to record the position in the hierarchy
  - Limited number of digits available
  - May hide part of the structure
- Terminologies: ICD, SNOMED, ...

Cystic fibrosis in ICD

Conclusions

- Biomedical terms
  - reflect some aspects of biomedical reality
  - Although the primary concern of terminology is naming, not reflecting reality
  - often convey additional features (e.g., epistemology)
- Biomedical terminology tends to offset part of the complexity
  - but often reflects utility

Conclusions 😊

- Biomedical terminologies can help populate biomedical ontologies
- Resources needed
  - Linguistic analysis of terms
  - Statistical analysis of terms in a corpus / annotation database (dependence relations)
  - Manual curation
Agenda

Monday, June 9
Introduction to Biomedical Ontologies
Design Principles, Formalisms and Tools for Biomedical Ontologies
Biomedical Ontologies - Content and structure - Function

Tuesday, June 10
Interfaces to Biomedical Ontologies
Searching and Analyzing Biomedical Concepts
Contrasting Biomedical Ontologies

Wednesday, June 11
Critical Analysis of Biomedical Ontologies
Extending Biomedical Ontologies
Using Biomedical Ontologies for Data Integration

Overview
- Definitions
  - Ontologies vs. other artifacts
  - Kinds of ontologies
- Some principles of formal ontology
  - Top-level categories
  - Categories of relationships
- Formalisms and tools

Introduction
- Concept
  - Symbol
  - Object

Definitions
- The *What* question
  - Objects in the world
    - With their properties
    - With their relations to other objects
  - Also: events, processes, and states
- Explicit specification of a conceptualization
  - Support software applications
- Domain ontology reflects
  - Underlying reality
  - Theory of the domain
Examples of use

- Natural language processing
- Access to heterogeneous sources of information (e.g., Semantic Web)
- Systems engineering
- Interoperability

Ontology vs. other artifacts

- Ontology
  - Defining types of things and their relations
- Terminology
  - Naming things in a domain
- Thesaurus
  - Organizing things for a given purpose
- Classification
  - Placing things into (arbitrary) classes
- Knowledge bases
  - Assertional knowledge

Examples of use

- Natural language processing
- Access to heterogeneous sources of information (e.g., Semantic Web)
- Systems engineering
- Interoperability

Ontology vs. other artifacts

- Ontology
  - Defining types of things and their relations
- Terminology
  - Naming things in a domain
- Thesaurus
  - Organizing things for a given purpose
- Classification
  - Placing things into (arbitrary) classes
- Knowledge bases
  - Assertional knowledge

(C)ontrolled Terminology

- Objective: naming things
- Shared understanding
  - Agreement on what terms to use
  - Utility-driven (arbitrary)

Thesaurus

- Objective: organize things for a purpose
- e.g., information retrieval
  - Organization by relatedness
- Example: Medical Subject Headings (MeSH)
- Relations used in hierarchies vs. hierarchical relations

Thesaurus vs. ontology

- Objective: placing things into classes
- Characteristics
  - Single inheritance (tree)
  - Idiosyncratic inclusion/exclusion criteria
Classification

- Characteristics (continued)
  - Everything must be classified, including
    - When there is no specific slot (NEC)
    - When there is insufficient information (NOS)

Knowledge Bases

- Objective: represent knowledge needed for a given application
- Example: drug knowledge bases
- Assertional knowledge
  - Vs. definitional knowledge in ontologies
  - Often probabilistic
- Examples of assertions
  - Indications of a drug
  - Signs and symptoms of a disease

Fuzzy borders

- Some ontologies also collect names
  - FMA
- Some terminologies also provide formal definitions
  - SNOMED CT
- Some terminologies/ontologies include both definitional and assertional knowledge
  - SNOMED CT

Types of resources

- Lexical resources
  - Collections of lexical items
  - Additional information
    - Part of speech
    - Spelling variants
  - Useful for entity recognition
- Ontological resources
  - Collections of
    - kinds of entities (substances, qualities, processes)
    - relations among them
  - Useful for relation extraction
- Terminological resources
  - Collections lexical items + identifiers
  - Useful for entity resolution
  - UMLS Metathesaurus

The Knowledge Semantics Continuum

Kinds of ontologies
Ontology-related issues

- Creation
- Merging
- Alignment
- Validation

Formal Ontological Principles

Formal ontological distinctions

- Universal vs. individual
- Continuant vs. occurrent
- Independent vs. dependent

Universal vs. Individual

- Universal = category
- Synonyms
  - Kind, Type, (Class)
- Examples
  - eyeball
  - blood pressure
  - conference

- Individual = instance
- Synonyms
  - Particular, Token
- Examples
  - my right eyeball
  - my blood pressure (132/79)
  - AMIA Annual Symposium 2003

Continuant vs. Occurrent

- Continuant = Continues to exist through time
- Synonyms
  - Substance
- Examples
  - tennis racquet
  - mitochondrion
  - insulin production

- Occurrent = Unfolds through time
- Synonyms
  - Process
- Examples
  - tennis tournament
  - metabolism
  - producing insulin

Independent vs. Dependent

- Independent = Can exist without support from other entities
- Examples
  - virus
  - molecule
  - plant

- Dependent = Require support from other entities for its existence
- Examples
  - viral infection
  - DNA binding
  - food
Formal ontology Upper level

Universals (classes) → Continuant

Independent continuant → Dependent continuant

Particulars (instances) → Occurrent

Formal ontological distinctions

- Basic distinctions in many top-level ontologies
  - Generic: BFO, DOLCE
  - Biomedical: BioTop, UMLS Semantic Network
- Condition the relations between various types of entities
  - Relations
    - Between instances (e.g., part_of [at time])
    - Between classes (e.g., isa, part_of [atemporal])
    - Between one instance and one class (instance_of)

[Smith, Genome Biology 2005]

Formal ontology in practice

- Provides foundational classes and relations
  - Upper level ontologies
  - Relation ontology
- Provides a framework for analyzing entities and relations

Examples

General ontologies

- OpenCyc
  - General ontology
  - Cycorp, Inc (D. Lenat & al.)
  - Over 1M hand-coded assertions
  - http://www.opencyc.org
- WordNet
  - Electronic lexical database
  - Princeton University (G. Miller & al.)
  - Over 100,000 synsets
  - http://wordnet.princeton.edu/

Top level in OpenCyc

- Mathematical or computational class
- Physical or biological class
- Conceptual or informational class
- Other relations

Over 1M hand-coded assertions
**Top level in WordNet**

- Abstraction
- Activity
- Entity
- Event
- Group
- Location
- Natural phenomenon
- Possession
- Psychological feature
- Shape
- State

**GALEN**

- Generalised Architecture for Languages, Encyclopaedias, and Nomenclatures in Medicine
- European Union project (A. Rector & al.)
- Over 25,000 concepts (primitives)
- [http://www.opengalen.org](http://www.opengalen.org)

**Top level in GALEN**

**UMLS Semantic Network**

- Definitional knowledge in the biomedical domain
- NLM (A. McCray & al.)
- Content
  - 135 semantic types
  - 54 types of relationship
  - 6700 semantic relations

**Top level in the Semantic Network**

**Differences between ontologies**

*Examples*
Granularity, plesionymy

UMLS

WordNet

Differing categorization

UMLS

WordNet

Ontology and Formalism

- Frames
- Description logics
  - OWL DL
- First-order logic
- OBO Format
  - Conversion to OWL DL

Tools for ontology developers

- Protégé
  - Publicly available
  - Frames and DL
  - Classifier
  - Supports many file formats (import/export)
  - Large community of users
  - Well supported
  - http://protege.stanford.edu/
- OBO-Edit
  - Specific to the biomedical/OBO community
  - Simpler than Protégé (“tool for biologists”)
  - http://oboedit.org/

Agenda

Monday, June 9
Introduction to Biomedical Ontologies
Design Principles, Formalisms and Tools for Biomedical Ontologies
Biomedical Ontologies

Tuesday, June 10
Interfaces to Biomedical Ontologies
Searching and Analyzing Biomedical Concepts
Contrasting Biomedical Ontologies

Wednesday, June 11
Critical Analysis of Biomedical Ontologies
Extending Biomedical Ontologies
Using Biomedical Ontologies for Data Integration
Overview

- Structural perspective
  - What are they (vs. what are they for)?
  - "High-impact" biomedical ontologies
    - International Classification of Diseases (ICD)
    - Logical Observation Identifiers, Names and Codes (LOINC)
    - SNOMED Clinical Terms
    - Foundational Model of Anatomy
    - Gene Ontology
    - RxNorm
    - Medical Subject Headings (MeSH)
    - NCI Thesaurus
    - Unified Medical Language System (UMLS)

ICD Characteristics (1)

- Current version: ICD-10
- Type: Classification
- Domain: Disorders
- Developer: World Health Organization (WHO)
- Funding: WHO
- Availability
  - Publicly available: No
  - Repositories: UMLS [ICD9-CM in NCBO BioPortal]
- URL: http://www.who.int/classifications/icd/en/

ICD Characteristics (2)

- Number of
  - Concepts: 12,318
  - Terms: 1 per concept (tabular)
- Major organizing principles:
  - Tree (single inheritance hierarchy)
  - No explicit classification criteria
    - Idiosyncratic inclusion/exclusion mechanism
  - .8 slots for Not elsewhere classified (NEC)
  - .9 slots for Not otherwise specified (NOS)
- Formalism: Proprietary format

ICD Top level

<table>
<thead>
<tr>
<th>chapter</th>
<th>code</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Certain infectious and parasitic diseases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neoplasms</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diseases of the blood and blood-forming organs, and certain disorders involving the immune system</td>
</tr>
<tr>
<td>01</td>
<td>010-099</td>
<td>Acute infectious and parasitic diseases</td>
</tr>
<tr>
<td>02</td>
<td>100-199</td>
<td>Chronic infectious and parasitic diseases</td>
</tr>
<tr>
<td>03</td>
<td>200-299</td>
<td>Neoplasms of the lymphatic and hematopoietic system</td>
</tr>
<tr>
<td>04</td>
<td>300-399</td>
<td>Diseases of the blood and blood-forming organs, and certain disorders involving the immune system</td>
</tr>
<tr>
<td>05</td>
<td>400-499</td>
<td>Diseases of the skin and subcutaneous tissue</td>
</tr>
<tr>
<td>06</td>
<td>500-599</td>
<td>Diseases of the musculoskeletal system and connective tissue</td>
</tr>
<tr>
<td>07</td>
<td>600-699</td>
<td>Diseases of the digestive system</td>
</tr>
<tr>
<td>08</td>
<td>700-799</td>
<td>Diseases of the endocrine system and metabolism</td>
</tr>
<tr>
<td>09</td>
<td>800-899</td>
<td>Diseases of the respiratory system</td>
</tr>
<tr>
<td>10</td>
<td>900-959</td>
<td>Diseases of the circulatory system</td>
</tr>
<tr>
<td>11</td>
<td>960-999</td>
<td>Congenital malformations, deformations, and chromosomal abnormalities</td>
</tr>
</tbody>
</table>

| 12 | 000-099 | Certain infectious and parasitic diseases |
| 13 | 100-199 | Chronic infectious and parasitic diseases |
| 14 | 200-299 | Neoplasms of the lymphatic and hematopoietic system |
| 15 | 300-399 | Diseases of the blood and blood-forming organs, and certain disorders involving the immune system |
| 16 | 400-499 | Diseases of the skin and subcutaneous tissue |
| 17 | 500-599 | Diseases of the musculoskeletal system and connective tissue |
| 18 | 600-699 | Diseases of the digestive system |
| 19 | 700-799 | Diseases of the endocrine system and metabolism |
| 20 | 800-899 | Diseases of the respiratory system |
| 21 | 900-959 | Diseases of the circulatory system |
| 22 | 960-999 | Congenital malformations, deformations, and chromosomal abnormalities |
ICD Example

- Idiosyncratic inclusion/exclusion criteria

Example:

- Not elsewhere classified (NEC)
- Not otherwise specified (NOS)

LOINC Characteristics (1)

- Current version: 2.22 (Dec. 2007)
- Type: Controlled terminology*
- Domain: Laboratory and clinical observations
- Developer: Regenstrief Institute
- Funding: NLM
- Availability
  - Publicly available: Yes
  - Repositories: UMLS
  - URL: www.regenstrief.org/loinc/loinc.htm

LOINC Example

- Sodium: SCnc -: Pt: Ser/Plas: Qn
  [the molar concentration of sodium is measured in the plasma (or serum), with quantitative result]
SNOMED Clinical Terms

SNOMED CT Characteristics (1)
- Current version: January 31, 2008 (2 annual releases)
- Type: Reference terminology / ontology
- Domain: Clinical medicine
- Developer: IHTSDO
- Funding: IHTSDO
- Availability
  - Publicly available: Yes* (in member countries)
  - Repositories: UMLS
- URL: http://www.ihtsdo.org/

SNOMED CT Characteristics (2)
- Number of
  - Concepts: 311,313 active concepts (Jan. 31, 2008)
  - Terms: 794,061 active “descriptions”
- Major organizing principles:
  - Utility for clinical medicine (e.g., assertional + definitional knowledge)
  - Model of meaning (incomplete)
  - Rich set of associative relationships
  - Small proportion of defined concepts (many primitives)
- Formalism: Description logics (KRSS)

SNOMED CT Top level

SNOMED CT Example

Foundational Model of Anatomy
FMA Characteristics (1)

- Current version: ? (no fixed release schedule)
- Type: Ontology
- Domain: Anatomy (anatomical structures)
- Developer: U. Washington, Department of Biological Structure
- Funding: NLM (grants and contract) and others
- Availability
  - Publicly available: Yes
  - Repositories: [UMLS] / OBO / NCBO BioPortal
- URL: http://fma.biostr.washington.edu/

FMA Characteristics (2)

- Number of
  - Concepts: ~72k
  - Terms: ~1.5 / concept
- Major organizing principles:
  - Explicit classificatory criteria
  - Distinct isa and part_of hierarchies
  - Additional spatial relations (e.g., adjacency)
  - Multiple levels of granularity (organism to sub-cellular)
- Formalism: Frames (Protégé)
  - Conversion to OWL Full and OWL DL available

Gene Ontology Characteristics (1)

- Current version: n/a (daily/monthly releases)
- Type: Controlled vocabulary
- Domain: Molecular biology
- Developer: GO Consortium
- Funding: NIH (grants)
- Availability
  - Publicly available: Yes
  - Repositories: UMLS / OBO / NCBO BioPortal
- URL: http://www.geneontology.org/
Gene Ontology Characteristics (2)

- Number of
  - Terms: 2.15 per concept
- Major organizing principles:
  - 3 major hierarchies
    - Molecular function
    - Cellular component
    - Biological process
- Relations (within hierarchies): isa, part_of, regulates
- No relations between concepts across hierarchies
- Formalism: OBO format

Gene Ontology Top level (MF)

- Major organizing principles:
  - 3 major hierarchies
    - Molecular function
    - Cellular component
    - Biological process
- Relations (within hierarchies): isa, part_of, regulates
- No relations between concepts across hierarchies

Gene Ontology Top level (CC)

- Major organizing principles:
  - 3 major hierarchies
    - Molecular function
    - Cellular component
    - Biological process
- Relations (within hierarchies): isa, part_of, regulates
- No relations between concepts across hierarchies

Gene Ontology Top level (BP)

- Major organizing principles:
  - 3 major hierarchies
    - Molecular function
    - Cellular component
    - Biological process
- Relations (within hierarchies): isa, part_of, regulates
- No relations between concepts across hierarchies

Gene Ontology Example

- Major organizing principles:
  - 3 major hierarchies
    - Molecular function
    - Cellular component
    - Biological process
- Relations (within hierarchies): isa, part_of, regulates
- No relations between concepts across hierarchies

RxNorm

- Major organizing principles:
  - 3 major hierarchies
    - Molecular function
    - Cellular component
    - Biological process
- Relations (within hierarchies): isa, part_of, regulates
- No relations between concepts across hierarchies
**RxNorm Characteristics (1)**

- Current version: June 2, 2007 (monthly releases)
- Type: Controlled terminology
- Domain: Drug names
- Developer: NLM
- Funding: NLM
- Availability
  - Publicly available: Yes*
  - Repositories: UMLS

**RxNorm Characteristics (2)**

- Number of
  - Concepts: 93k
  - Terms: 105k
- Major organizing principles:
  - Generic vs. brand
  - Combinations of Ingredient / Form / Dose
  - No hierarchical structure
  - Links to all major US drug information sources
  - No clinical information
- Formalism: UMLS RRF format

**RxNorm Normalized form**

<table>
<thead>
<tr>
<th>Strength</th>
<th>Ingredient</th>
<th>Dose form</th>
</tr>
</thead>
<tbody>
<tr>
<td>4mg/ml</td>
<td>Fluoxetine</td>
<td>Oral Solution</td>
</tr>
</tbody>
</table>

**Rx Norm Generic vs. Brand**

- **Generic**
  - Ingredient (IN)
  - Clinical drug form (SCDF)
  - Clinical drug component (SCDC)
  - Clinical drug (SCD)

- **Brand**
  - Brand name (BN)
  - Branded drug form (SBDF)
  - Branded drug component (SBDC)
  - Branded drug (SBD)

**RxNorm Relations among drug entities**

**Medical Subject Headings (MeSH)**
**MeSH Characteristics (1)**
- Current version: 2008 (yearly releases)
- Type: Thesaurus / Controlled vocabulary
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (Library Operations)
- Availability
  - Publicly available: Yes
  - Repositories: UMLS / NCBO BioPortal

**MeSH Characteristics (2)**
- Number of
  - Terms: 7.5 per descriptor
- Major organizing principles:
  - Descriptor + entry terms
    (also: Qualifiers, Supplementary concepts)
  - Thesaurus relations (RB/RN/RO)
- Formalism: Thesaurus / Proprietary XML DTD

**MeSH Top level**
- 1. Anatomy [A]
- 2. Organisms [B]
- 3. Genetics [G]
- 4. Chemicals and Drugs [F]
- 5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [T]
- 6. Psychiatry and Psychology [P]
- 7. Biological Sciences [K]
- 8. Health Sciences [H]
- 10. Technology, Industry, Apportionment [J]
- 11. Human Information [X]
- 12. Information Sources [I]
- 13. Shared Groups [M]
- 14. Health Care [N]
- 15. Public Health: Cautions and Precautions [V]
- 16. Geographic sites [L]

**MeSH Example (terms)**
- Polycyclic Compounds
  - Steroids
    - Pregnenes
    - Pregnenediols
    - 11-Hydroxypregnenediols
  - Hydrocortisone
- Hormones, Hormone Substitutes, and Hormone Antagonists
  - Adrenal Cortex Hormones
    - Hydrocorticosteroids

**MeSH Example (hierarchies)**

**NCI Thesaurus**
NCI thesaurus Characteristics (1)

- Current version: 08.04d (~monthly releases)
- Type: Controlled terminology / ontology
- Domain: Cancer
- Developer: NCI Center for Bioinformatics
- Funding: NCI
- Publicly available: Yes
- Repositories: UMLS / OBO / NCBO
- URL: http://nciterms.nci.nih.gov/

NCI thesaurus Characteristics (2)

- Number of:
  - Concepts: 58,868 (2007_05E)
  - Terms: 2.68 per concept
- Major organizing principles:
  - Subsumption hierarchy
  - Rich set of associative relationships
  - Small proportion of defined concepts (many primitives)
  - Links to many external resources
- Formalism: OWL Lite

NCI thesaurus Top level

- NCI Thesaurus Taxonomy:
  - Abnormal Cell
  - Anatomy
  - Anatomical Structure, System, or Substrate
  - Biological Process
  - Biochemical Reaction
  - Clinical Sign or Symptom
  - Conceptual Domain
  - Diagnostic, Therapeutic, and Research Procedure
  - Diagnostic or Diagnostic Follow-up
  - Drug, Food, Chemical or Biomedical Material
  - Experimental Organism
  - Experimental Organism Anatomical Part
  - Experimental Organism Organ
  - Experiment
  - Event
  - Etiologic Factor
  - External Source or Information
  - Gene Product
  - Molecular Abnormality
  - NCI Administrative Concept
  - Organ
  - Property or Attribute
  - Internal Concept

NCI thesaurus Example

Unified Medical Language System (UMLS)

UMLS Characteristics (1)

- Current version: 2008AA (2-3 annual releases)
- Type: Terminology integration system
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (intramural)
- Publicly available: Yes* (cost-free license required)
- Repositories: UMLS
- URL: http://umlsks.nlm.nih.gov/
**UMLS Characteristics (2)**

- **Number of**
  - Concepts (~ 1.5 M) 2008AA
  - Terms (~ 6 M)
- **Major organizing principles (Metathesaurus):**
  - Concept orientation
  - Source transparency
  - Multi-lingual through translation
- **Formalism:** Proprietary format (RRF)

---

**Addison's Disease: Concept**

- Concept: *Addison's Disease*
  - Disease or Syndrome
  - SNOMED CT
  - MeSH
  - SNOMED Intl
  - UMLS

---

**Metathesaurus Concepts (2007AB)**

- **Concept (~ 1.4 M) CUI**
  - Set of synonymous concept names
- **Term (~ 5.3 M) LUI**
  - Set of normalized names
- **String (~ 5.9 M) SUI**
  - Distinct concept name
- **Atom (~ 7.2 M) AUI**
  - Concept name in a given source

---

**Recap**

<table>
<thead>
<tr>
<th>Name</th>
<th>Scope</th>
<th>Concepts</th>
<th>Version</th>
<th>Date</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMLS</td>
<td>SNOMED CT</td>
<td>Clinical observations on the literature</td>
<td>3.0.0</td>
<td>Mar 07, 2007</td>
<td></td>
</tr>
<tr>
<td>UMLS</td>
<td>ICD-10</td>
<td>Clinical observations on the literature</td>
<td>10.0.0</td>
<td>Dec 09, 2007</td>
<td></td>
</tr>
<tr>
<td>UMLS</td>
<td>SNOMED CT</td>
<td>Clinical observations on the literature</td>
<td>3.0.0</td>
<td>Mar 07, 2007</td>
<td></td>
</tr>
<tr>
<td>UMLS</td>
<td>SNOMED CT</td>
<td>Clinical observations on the literature</td>
<td>3.0.0</td>
<td>Mar 07, 2007</td>
<td></td>
</tr>
</tbody>
</table>

---

Dr. Olivier Bodenreider

University of Utah, Biomedical Informatics
# Agenda

<table>
<thead>
<tr>
<th>Monday, June 9</th>
<th>Introduction to Biomedical Ontologies</th>
<th>Design Principles, Formalisms and Tools for Biomedical Ontologies</th>
<th>Biomedical Ontologies - Content and Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, June 10</td>
<td>Interfaces to Biomedical Ontologies</td>
<td>Searching and Analyzing Biomedical Concepts</td>
<td>Contrasting Biomedical Ontologies</td>
</tr>
<tr>
<td>Wednesday, June 11</td>
<td>Critical Analysis of Biomedical Ontologies</td>
<td>Extending Biomedical Ontologies</td>
<td>Using Biomedical Ontologies for Data Integration</td>
</tr>
</tbody>
</table>

Lester Hill National Center for Biomedical Communications

---

# Overview

- Functional perspective
  - What are they for (vs. what are they)?
  - “High-impact” biomedical ontologies
- 3 major categories of use
  - Knowledge management (indexing and retrieval of data and information, access to information, mapping among ontologies)
  - Data integration, exchange and semantic interoperability
  - Decision support and reasoning (data selection and aggregation, decision support, natural language processing applications, knowledge discovery).

[Bodenreider, YBMI 2008]

Lester Hill National Center for Biomedical Communications

---

# Knowledge management

- Annotating data and resources

- Ontology as a source of vocabulary
  - List of names for the entities in the ontology (ontology vs. terminology)
- Most ontologies have some sort of terminological component
  - Exceptions: GALEN, LOINC
- Not all surface forms represented
  - Often insufficient for NLP applications
  - Large variation in number of terms per concept across ontologies

Lester Hill National Center for Biomedical Communications
Annotating data

- Gene Ontology
  - Functional annotation of gene products in several dozen model organisms
- Various communities use the same controlled vocabularies
- Enabling comparisons across model organisms
- Annotations
  - Assigned manually by curators
  - Inferred automatically (e.g., from sequence similarity)

Gene Ontology

Functional annotation of gene products in several dozen model organisms
Various communities use the same controlled vocabularies
Enabling comparisons across model organisms
Annotations
Assigned manually by curators
Inferred automatically (e.g., from sequence similarity)

GO Annotations for Aldh2 (mouse)

<table>
<thead>
<tr>
<th>Category</th>
<th>Classification Form</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Function</td>
<td>oxidoreductase activity</td>
<td>EBI</td>
</tr>
<tr>
<td>Molecular Function</td>
<td>oxidoreductase activity</td>
<td>EBI</td>
</tr>
<tr>
<td>Molecular Function</td>
<td>oxidoreductase activity</td>
<td>EBI</td>
</tr>
<tr>
<td>Cellular Component</td>
<td>metabolism</td>
<td>EBI</td>
</tr>
<tr>
<td>Biological Process</td>
<td>metabolism</td>
<td>EBI</td>
</tr>
</tbody>
</table>

Annotations

Assigned manually by curators
Inferred automatically (e.g., from sequence similarity)

http://www.informatics.jax.org/

GO Annotations for ALDH2 (Human)

<table>
<thead>
<tr>
<th>Function</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>oxidoreductase activity</td>
<td>EBI</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>EBI</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>EBI</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>EBI</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>EBI</td>
</tr>
</tbody>
</table>

Annotations

Assigned manually by curators
Inferred automatically (e.g., from sequence similarity)

http://www.ebi.ac.uk/GOA/

Indexing the biomedical literature

- MeSH
  - Used for indexing and retrieval of the biomedical literature (MEDLINE)
- Indexing
  - performed manually by human indexers
  - With help of semi-automatic systems (suggestions)
  - Indexing Initiative at NLM
  - Automatic indexing systems

MeSH MEDLINE indexing

For cited in patents and since 1980:

- Biomedical Literature
- MeSH MEDLINE indexing
- Indexing Initiative at NLM
- Automatic indexing systems
MeSH MEDLINE indexing

- Male
- Remaining Value of This
- Treatment Studies
- Remaining Value
- Conditions Related
- By Location
- Other Related MeSH Terms
- Related Terms

Subtopics:
- Biological Terms
- Functional Terms

ICD9-CM Coding clinical data

- ICD9-CM
  - Used for coding clinical data e.g., for billing purposes
- Other uses of ICD
  - Morbidity and mortality reporting worldwide

Knowledge management

Accessing biomedical information

Resources for biomedical search engines

- Synonyms
- Hierarchical relations
- High-level categorization
- Co-occurrence information
MeSH "synonyms" MEDLINE retrieval

- MeSH entry terms
  - Used as equivalent terms for retrieval purposes
  - Not always synonymous
- Increase recall without hurting precision

<table>
<thead>
<tr>
<th>MeSH</th>
<th>Addion Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heading</td>
<td></td>
</tr>
<tr>
<td>Entry Term</td>
<td>Primary Adrenal Insufficiency</td>
</tr>
<tr>
<td>Entry Term</td>
<td>Primary Adrenocortical Insufficiency</td>
</tr>
</tbody>
</table>

MeSH hierarchies MEDLINE retrieval

- MeSH "explosion"
  - Search for a given MeSH term and all its descendants
  - A search on Adrenal insufficiency also retrieves articles indexed with Addison disease

Co-indexing

GoPubMed

http://www.goPubMed.com/

Knowledge management

Mapping across biomedical ontologies
Reusing information

- Clinical information coded with SNOMED CT
  - Mapped to ICD9-CM and CPT for billing purposes
  - Mapped to ICD-O for epidemiology purposes
- Existing mapping tables created by terminology developers as an incentive to use SNOMED CT

Reusing tools

- For noun phrases extracted from medical texts, map to UMLS concepts [Aronson & al., AMIA, 2000]
- Then, select from the MeSH vocabulary the concepts that are the most closely related to the original concepts

Terminology integration systems

- Terminology integration systems (UMLS, RxNorm) help bridge across vocabularies
- Uses
  - Information integration
  - Ontology alignment
  - Medication reconciliation

Integrating subdomains

- Biomedical literature
- Genomic annotations
- Other subdomains
- Model organisms
- Gene annotations
- Clinical repositories
- Genetic knowledge bases
- SNOMED CT
- UMLS
- MSH
- MeSH
- OMIM
- GO
- NCBO BioPortal
- NCBI Taxonomy
- FMA
- Anatomy
- Genome annotations

Trans-namespace integration

- Addison's disease (D000224)
Data integration, exchange and semantic interoperability

"Standards"

- Ontologies help standardize patients data
  - Facilitate the exchange of data across institutions
  - Help connect “islands of data” (silos)

- LOINC
  - Exchange of laboratory data
  - In conjunction with HL7 messaging

Semantic interoperability projects BRIDG

- Biomedical Research Integrated Domain Group
  - Information model for clinical research
  - Interoperability between clinical trials information systems
  - Ontologies provide value sets to the information model

Semantic interoperability projects CDA

- Clinical Document Architecture (CDA R2)
  - Formal representation of clinical statements
    - Clinical observations
    - Medication administration
    - Adverse events
  - Associate an information model (HL7 RIM) with terminologies (LOINC, SNOMED CT, RxNorm)

Semantic interoperability projects caCORE

- Cancer Common Ontologic Representation Environment
  - Infrastructure developed to support an interoperable biomedical information system for cancer research
  - Uses the NCI Thesaurus as a component
Data integration, exchange and semantic interoperability

Information and data integration

Approaches to data integration
- Warehousing
  - Sources to be integrated are transformed into a common format and converted to a common vocabulary
  - Normalization through ontologies (e.g., GO annotations)
- Mediation
  - Local schema (of the sources)
  - Global schema (in reference to which the queries are made)
  - Ontologies help define the global schema and map between local and global schemas (OntoFusion, ARIANE)

Ontologies and integration
- Terminology integration systems help bridge across terminologies and the domains they represent
- Mappings across ontologies enable the integration of namespaces in the Semantic Web

Trans-namespace integration

Data selection
- The structure of biomedical ontologies helps define groups of values from a high-level value
  - Vs. enumerating all possible values
- Useful for data selection in clinical studies
- ICD is used pervasively for this purpose
  - E.g., Study on supraventricular tachycardia (SVT), based on 2 high-level ICD codes
- Similarity with the definition of value sets for use in the information model
Data aggregation

- Ontologies help partition/aggregate data in data analysis
  - Clinical studies: Study a variable in groups of patients corresponding to the top level categories in ICD
  - Biology studies: Functional characterization of gene expression signatures with high-level concepts from the Gene Ontology
    - Recent trend: co-clustering

Decision support

- Clinical decision support
  - Ontologies help normalize the vocabulary and increase the recall of rules
  - Ontologies provide some domain knowledge and make it possible to create high-level rules (e.g., for a class of drugs rather than for each drug in the class)
- Other forms of decision support
  - Based on automatic reasoning services for OWL ontologies (e.g., grading gliomas with NCIt)

Natural language processing applications

- Ontologies provide background domain knowledge for NLP applications
  - Question answering
  - Document summarization
  - Literature-based discovery
- The UMLS is often used, but other specific resources have been developed

Knowledge discovery

- By standardizing the vocabulary in a given domain, ontologies are enabling resources for knowledge discovery through data mining
- Less frequently, the structure of the ontology is leveraged by data mining algorithms
- Example of available datasets
  - ICD-coded clinical data (in conjunction with non-clinical information, e.g., environmental data)
  - Annotation of gene products to the GO (function prediction)

Barriers to usability of biomedical ontologies

- Many ontologies are freely available
- The UMLS is freely available for research purposes
  - Cost-free license required
- Licensing issues can be tricky
  - SNOMED CT is freely available in member countries of the IHTSDO
- Being freely available
  - Is a requirement for the Open Biomedical Ontologies (OBO)
  - Is a de facto prerequisite for Semantic Web applications
Discoverability

- Ontology repositories
  - UMLS: 143 source vocabularies (biased towards healthcare applications)
  - NCBO BioPortal: ~100 ontologies (biased towards biological applications)
  - Limited overlap between the two repositories
- Need for discovery services

Ontology integration

- Post hoc integration, form the bottom up
  - UMLS approach
  - Integrates ontologies “as is”, including legacy ontologies
  - Facilitates the integration of the corresponding datasets
- Coordinated development of ontologies
  - OBO Foundry approach
  - Ensures consistency ab initio
  - Excludes legacy ontologies

Formalism

- Several major formalism
  - Web Ontology Language (OWL) – NCI Thesaurus
  - OBO format – most OBO ontologies
  - UMLS Rich Release Format (RRF) – UMLS, RxNorm
- Conversion mechanisms
  - OBO to OWL
  - LexGrid (import/export to LexGrid internal format)

Quality

- Quality assurance in ontologies is still imperfectly defined
  - Difficult to define outside a use case or application
- Several approaches to evaluating quality
  - Collaboratively, by users (Web 2.0 approach)
  - Centrally, by experts
    - OBO Foundry approach
- Important factors besides quality
  - Governance
  - Installed base / Community of practice

Agenda

<table>
<thead>
<tr>
<th>Monday, June 9</th>
<th>Introduction to Biomedical Ontologies</th>
<th>Design Principles, Formalisms and Tools for Biomedical Ontologies</th>
<th>Biomedical Ontologies - Content and Structure - Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, June 10</td>
<td>Interfaces to Biomedical Ontologies</td>
<td>Searching and Analyzing Biomedical Concepts</td>
<td>Contrasting Biomedical Ontologies</td>
</tr>
<tr>
<td>Wednesday, June 11</td>
<td>Critical Analysis of Biomedical Ontologies</td>
<td>Extending Biomedical Ontologies</td>
<td>Using Biomedical Ontologies for Data Integration</td>
</tr>
</tbody>
</table>

Interfaces to Biomedical Ontologies
Overview

- Graphical interfaces
  - UMLS Knowledge Source Server
  - NCBO BioPortal
  - NCI Thesaurus
  - MeSH browser
  - Foundational Model of Anatomy (FMA) Explorer
  - Gene Ontology AmiGO
  - ICD-10 online
  - RxNav (RxNorm)
  - [...] 
- Application Programming Interfaces

Graphical interfaces

![UMLS Knowledge Source Server](http://umlsks.nlm.nih.gov/)

![UMLS Semantic Navigator](http://mor.nlm.nih.gov/perl/semnav.pl)

![NCBO BioPortal](http://www.bioontology.org/tools/portal/bioportal.html)

![NCI Thesaurus (EVS Server)](http://nciterms.nci.nih.gov/NCIBrowser/SearchConcept.do)
MeSH Browser


Foundational Model of Anatomy

http://sig.biostr.washington.edu/projects/fm/FME/

Gene Ontology AmiGO

http://www.geneontology.org/

SNOMED CT University of Sydney

http://www.it.usyd.edu.au/~hitru/sct/A1.cgi

SNOMED CT CliniClue

www.cliniclue.com

SNOMED CT SNOB

http://snob.eggbird.eu/
ICD-10

RxNav (RxNorm)

http://www.who.int/classifications/apps/icd/icd10online/


Application Programming Interfaces

- Expose resources in such a way that they can be integrated in programs
- Programming “against” a resource
- Standard protocols for communication
- Web services (SOAP, REST)
- Standard libraries for programming
- Focus on content, not message

UMLSKS Web Service API

  - Developer's Guide > Webservice Operations
- WSDL available
- API give access to all 3 knowledge sources
- Licensing issues
  - Granting ticket and Single-use tickets

UMLSKS Web Service API Example

Example: ConceptIdGroup findCUIByNormString

ConceptIdGroup findCUIByNormString
(ConceptIdNormStringRequest request);

This class contains the arguments that better search the behavior of the call

- ConceptIdNormStringRequest
  - ConceptId
    - String
  - String
  - String
  - String
  - String
  - String
  - String
  - String

- Single-use tickets
- Granting ticket and Single-use tickets
Other APIs to terminology systems

- NCBO BioPortal
  [http://www.bioontology.org/docs/bioportal/development/web_services.html](http://www.bioontology.org/docs/bioportal/development/web_services.html)
- OLS - Ontology Lookup Service
  [http://www.ebi.ac.uk/ontology-lookup/WSDL/Documentation.do](http://www.ebi.ac.uk/ontology-lookup/WSDL/Documentation.do)
- RxNorm

Applications based on WS APIs

- UMLSKS API
  - UMLSKS
- RxNorm API
  - RxNav
  - MyMedicationList

Agenda

<table>
<thead>
<tr>
<th>Day</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday, June 9</td>
<td>Introduction to Biomedical Ontologies</td>
</tr>
<tr>
<td>Tuesday, June 10</td>
<td>Interfaces to Biomedical Ontologies</td>
</tr>
<tr>
<td>Wednesday, June 11</td>
<td>Critical Analysis of Biomedical Ontologies</td>
</tr>
</tbody>
</table>

Exercise 1

- What are the Clinical Drug Components for Zyrtec? (RxNav)

Exercise 2

- What are the parts of the Aorta? (FMA)
Exercise 3

- What are the parents of Hodgkin’s disease in SNOMED CT?
  - Try SNOMEDCTID: 118599009
- What is its associated morphology?

Exercise 4

- What are the various meanings of IL-2? (UMLS)

Exercise 5

- What are the pharmacologic actions of Zyrtec? (MeSH)

Exercise 6

- What are some synonyms for Schwannoma? (NCI Thesaurus)

Solutions
Exercise 2

◆ What are the parts of the Aorta? (FMA)

Exercise 3

◆ What are the parents of Hodgkin’s disease in SNOMED DT?
  • Try SNOMEDCTID: 118599009
◆ What is its associated morphology?

Exercise 4

◆ What are the various meanings of IL-2? (UMLS)

Exercise 5

◆ What are the pharmacologic actions of Zyrtec? (MeSH)

Exercise 6

◆ What are some synonyms for Schwannoma? (NCI Thesaurus)

What to look for

◆ Search modalities
  • Spelling correction / auto-completion / normalization
  • Word combinations (AND/OR)
◆ Visualization
  • Graph vs. (forest of) trees
◆ Navigation
  • What properties are displayed
Biomedical Ontology in Practice

Agenda

<table>
<thead>
<tr>
<th>Monday, June 9</th>
<th>Introduction to Biomedical Ontologies</th>
<th>Design Principles, Formalisms, and Tools for Biomedical Ontologies</th>
<th>Biomedical Ontology - Content and Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuesday, June 10</td>
<td>Interfaces to Biomedical Ontologies</td>
<td>Searching and Analyzing Biomedical Concepts</td>
<td>Contrasting Biomedical Ontologies</td>
</tr>
<tr>
<td>Wednesday, June 11</td>
<td>Critical Analysis of Biomedical Ontologies</td>
<td>Extending Biomedical Ontologies</td>
<td>Using Biomedical Ontologies for Data Integration</td>
</tr>
</tbody>
</table>

Exercise #1

- Hodgkin’s disease
  - NCI Thesaurus
  - SNOMED CT

Exercise #2

- Prostate
  - FMA
  - SNOMED CT

Exercise #3

- Cetirizine
  - MeSH
  - SNOMED CT

Solutions
Exercise #1

◆ Hodgkin’s disease
  - NCI Thesaurus
    - Using the NCI browser (EVS)
    - SNOMED CT
      - Using the online browser from U. Sydney

Comments on Hodgkin’s disease in NCIt (1)

- Search term: “Hodgkin’s disease”
  - Not found, although “Hodgkin’s disease” is listed as a synonym
  - Search on “hodgkin”, select “Hodgkin lymphoma”
- Parent classes
  - Common hematopoietic neoplasm
    - Not an ontological category
    - Would be better represented through an associative relation (e.g., along the lines of “has_prevalence high prevalence”)
    - Isa overloading

Comments on Hodgkin’s disease in NCIt (2)

- Associative relations
  - For cancers, anatomy and morphology are foundational relations
  - Here
    - Anatomy: Disease_Has_Primary_Anatomic_Site
      - Hematopoietic and Lymphatic System
    - Morphology: not directly represented (indirectly through Disease_Has_Normal_Cell_Origin Mature Lymphocyte)
Hodgkin’s disease in SNOMED CT (1)

CONCEPT

Concept ID                  Fully Specified Name                  Concept Status
11980009                  Hodgkin’s disease                        Current

DESCRIPTIONS

- \(11980009\) in [Hodgkin’s disease - Concept]

ATTRIBUTES

- Name: Hodgkin’s disease
- Definition: Hodgkin’s disease
- Designation: Hodgkin’s disease

Comments on Hodgkin’s disease in SNOMED CT (1)

- Search term: “Hodgkin’s disease”
  - Not found, although “Hodgkin’s disease” is listed as a synonym
  - Search results:
    - “Hodgkin lymphoma, nodular sclerosis, grade 1 (morphologic abnormality)”
    - “Hodgkin’s disease” is ambiguous
    - “Hodgkin lymphoma, Hodgkin’s disease (disorder)”
    - “Malignant lymphoma, Hodgkin’s disease (disorder)”
    - Malignant lymphoma (clinical) [OK]
  - Parent classes
    - Malignant lymphoma (clinical) [OK]

Hodgkin’s disease NCIt vs. SNOMED CT (1)

- Shared synonyms: NCIt 1/2, SNOMED CT 1/3
  - Hodgkin’s disease
- Shared relations
  - ISA
    - NCIt: Lymphoma
      - Definition: “malignant (clonal) proliferation of B-lymphocytes or T-lymphocytes which involves the lymph nodes, bone marrow and/or extranodal sites. This category includes Non-Hodgkin lymphomas and Hodgkin lymphomas.”
    - SNOMED CT: Malignant lymphoma
  - Same UMLS concept (CUI: C0024299)

Hodgkin’s disease in SNOMED CT (2)

CONCEPT

Concept ID                  Fully Specified Name                  Concept Status
11980009                  Hodgkin’s disease                        Current

DESCRIPTIONS

- \(11980009\) in [Hodgkin’s disease - Concept]

Attributes

- Name: Hodgkin’s disease
- Definition: Hodgkin’s disease
- Designation: Hodgkin’s disease

Comments on Hodgkin’s disease in SNOMED CT (2)

- Search term: “Hodgkin’s disease”
  - Not found, although “Hodgkin’s disease” is listed as a synonym
  - Search results:
    - “Hodgkin lymphoma, nodular sclerosis, grade 1 (morphologic abnormality)”
    - “Hodgkin’s disease” is ambiguous
    - “Hodgkin lymphoma, Hodgkin’s disease (disorder)”
    - “Malignant lymphoma, Hodgkin’s disease (disorder)”
    - Malignant lymphoma (clinical) [OK]
  - Parent classes
    - Malignant lymphoma (clinical) [OK]

Hodgkin’s disease NCIt vs. SNOMED CT (2)

- Shared relations: Associative relations
  - Anatomy
    - In NCIt, but not in SNOMED CT
  - Morphology
    - In SNOMED CT, but not in NCIt
    - Only indirectly, through cell type
  - Cell type
    - Only in NCIt
Exercise #2

- **Prostate**
  - FMA
    - Using the Foundational Model Explorer
  - SNOMED CT
    - Using the online browser from U. Sydney

Comments on Prostate in FMA

- No synonyms in English
- Latin and Spanish synonyms
- Hierarchies
  - *Isa*: Lobular organ
  - *Part_of*: Set of pelvic viscera
- Associative relations
  - Lymphatic drainage
- No spatial relations

Prostate in SNOMED CT (1)

- 45 results found for prostate:

<table>
<thead>
<tr>
<th>Concept ID</th>
<th>Full Specification</th>
<th>Preferred Terms and Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>31216002</td>
<td>Prostate, male</td>
<td>Prostate (372), Prostate, Male</td>
</tr>
<tr>
<td>31216011</td>
<td>Prostate, female</td>
<td>Prostate (372), Prostate, Female</td>
</tr>
<tr>
<td>31216020</td>
<td>Prostate, prostate</td>
<td>Prostate (372), Prostate, Male</td>
</tr>
</tbody>
</table>

Prostate in FMA (1)

Prostate in FMA (2)
Prostate in SNOMED CT (2)

<table>
<thead>
<tr>
<th>Concept</th>
<th>FMA Target Concept</th>
<th>Relationship Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire prostate (bone structure)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Entire prostate (bone)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Prostatic structure (bone)</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Prostate in bone</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Prostate in bone (prostate structure)</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Prostate in bone (bone)</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

Comments on Prostate in SNOMED CT

- "Ambiguous" term
  - Entire prostate
  - Prostatic structure
- Structure-Entire-Part representation of anatomical entities in SNOMED CT
  - Refinement of part_of
  - Enables mereological inference through isa hierarchy
  - Not intuitive

Structure-Entire-Part (SEP) triples

- S – The entity or any of its parts
- E – The entire anatomical entity
- P – Any parts of the anatomical entity

Mereological inference through isa

FMA mapping goes to Entire

- Shared synonym: FMA 1/1, SNOMED CT 1/2
- Prostate
- Shared relations
  - isa: no
  - FMA
    - Lobular organ
    - SNOMED CT
      - Prostatic structure
      - Male internal genital organ
      - Entire male genital organ

Prostate  FMA vs. SNOMED CT
Prostate  FMA vs. SNOMED CT

- Shared relations
  - Part of: almost
    - FMA
      - Genital system
      - Content of male pelvis
      - Set of male pelvic viscera
      - Set of pelvic viscera
    - SNOMED CT
      - Entire minor pelvis
      - Entire male genital system

Solutions

Exercise #3

- Cetirizine
  - MeSH
    - Using the MeSH browser
    - SNOMED CT
      - Using the online browser from U. Sydney

Cetirizine in MeSH (1)

- Entry Terms: [645:63-4-Chlorophenyl)methylpiperazine-1-carboxylic acid]
- Entry Terms: Atropine
- Entry Terms: Alcoholism
- Entry Terms: Alcoholic liver disease
- Entry Terms: Alcoholic liver cirrhosis
- Entry Terms: Antihistamines
- Entry Terms: Antihistamines, Non-Sedating
- Entry Terms: Anti-Allergic Agents
- Entry Terms: Antihistamines, H1-Blocking Agents
- Entry Terms: Antihistamines, H1-Antagonists, Non-Sedating
- Entry Terms: Anti-Histaminic Agents
- Entry Terms: Anti-Histaminic Agents, Non-Sedating
- Entry Terms: Histamine H1 Antagonists, Non-Sedating

Comments on Cetirizine in MeSH

- 45 entry terms
  - Various generic and brand names
  - Chemical formula
  - Code (P-071)
- Hierarchy
  - Iau: Piperazines [chemistry]
- Pharmacologic action
  - Anti-Allergic Agents
  - Histamine H1 Antagonists, Non-Sedating
Cetirizine in SNOMED CT (1)

Lister Hill National Center for Biomedical Communications

Comments on Cetirizine in SNOMED CT

- Ambiguous term
  - Cetirizine (product)
  - Cetirizine (substance)
- Hierarchy
  - Isa: Non-sedating antihistamine (substance)
- No associative relations

Cetirizine MeSH vs. SNOMED CT (1)

- Shared synonyms: MeSH 1/45, SNOMED CT 1/1
- Cetirizine
- Shared relations: none
- MeSH:
  - Isa: <chemistry>
  - Associative: <pharmacologic action>
- SNOMED CT:
  - Isa: <pharmacologic action>
  - Associative: none

Summary

- Differing representations
  - Not necessarily inconsistent
  - Consistency may be difficult to assess automatically
- Often due to idiosyncratic representation in one ontology
- Hindrance to ontology alignment and evaluation methods relying on shared relations
Agenda

Monday, June 9
- Introduction to Biomedical Ontologies
- Design Principles, Formalisms and Tools for Biomedical Ontologies
- Biomedical Ontologies - Content and Structure
- Function

Tuesday, June 10
- Interfaces to Biomedical Ontologies
- Searching and Analyzing Biomedical Concepts
- Contrasting Biomedical Ontologies

Wednesday, June 11
- Critical Analysis of Biomedical Ontologies
- Extending Biomedical Ontologies
- Using Biomedical Ontologies for Data Integration

Overview

- Corpus terminology
- Identify terms in biomedical text (in reference to the UMLS)
- Identify additional terms
- Place these terms in UMLS hierarchies

Tiny corpus  One MEDLINE abstract

[Corpus terminology]

Identify UMLS concepts with MetaMap

http://skr.nlm.nih.gov/

Interactive mode
Interactive MetaMap

Select options

Run MetaMap

Output

Suggest term candidates

- Not recognized by MetaMap at all
- Partially identified by MetaMap
- Missing terms in a concept
Suggest placement in UMLSS

- Use a browser
- Identify close parent
- Examine its children
- Assess placement by comparing with potential siblings

Possible new terms (1)

- Hypothalamopituitary axis
  - Concept exists: C0678897, but missing exact (neoclassical) synonym
    - hypothalamic pituitary axis
    - hypothalamus hypophysis axis
    - hypothalamus-pituitary axis
- American College of Chest Physicians
  - Similar to other American Colleges (e.g., American College of Physicians)
  - Integrate as a child of Professional Organization or Group (C1522486)
  - NB: instance, cannot be a child of ACP

Possible new terms (2)

- Free cortisol
  - Identified as a substance (C0443476), not a laboratory procedure / test result
    - Cortisol, free measurement (C0236401)
- Coolens method
  - Missing term / concept
  - Method for estimating (not measuring directly) the free fraction of cortisol

Possible new terms (3)

- Electrochemiluminescence immunoassay
  - Missing concept
  - Create as a child of Chemiluminescence assay (C0201709)
- Nonsurvivors
  - Survivors exists as a concept (C0206194)
  - Create as a child of Patients (C0030705)

Agenda

<table>
<thead>
<tr>
<th>Day</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday, June 9</td>
<td>Introduction to Biomedical Ontologies</td>
<td></td>
</tr>
<tr>
<td>Tuesday, June 10</td>
<td>Design Principles, Formalisms and Tools for Biomedical Ontologies</td>
<td></td>
</tr>
<tr>
<td>Wednesday, June 11</td>
<td>Critical Analysis of Biomedical Ontologies</td>
<td></td>
</tr>
</tbody>
</table>
Overview

- Motivation
- Some practical considerations and issues
  - Integration approaches
  - Concept repositories
  - Using existing mappings
  - Creating mappings through the UMLS
  - Comparing semantic descriptions
- Thinking outside the integration box

Motivation

Translational research

- "Bench to Bedside"
- Integration of clinical and research activities and results
- Supported by research programs
  - NIH Roadmap
  - Clinical and Translational Science Awards (CTSA)
- Requires the effective integration and exchange and of information between
  - Basic research
  - Clinical research

Terminology and translational research

Basic Research  Clinical Research and Practice

Cancer Basic Research  EHR Cancer Patients

NCI Thesaurus  SNOMED CT
Some practical considerations and issues
Integration approaches

Approaches to data integration

- Warehousing
  - Sources to be integrated are transformed into a common format and converted to a common vocabulary
- Normalization through ontologies (e.g., GO annotations)

- Mediation
  - Local schema (of the sources)
  - Global schema (in reference to which the queries are made)
  - Ontologies help define the global schema and map between local and global schemas (OntoFusion, ARIANE)

(Integrated) concept repositories

- Unified Medical Language System
- NCBO’s BioPortal
  http://www.bioontology.org/tools/portal/biportal.html
- Open Biomedical Ontologies (OBO)
  http://obofoundry.org/
- caDSR
  http://ncicb.nci.nih.gov/NCICB/infrastructure/cadsvsreview/cadsr

Integrating subdomains

Trans-namespace integration

Addison Disease (D000224)
Some practical considerations and issues

Mappings

- Created manually
  - UMLS
- Created automatically
  - BioPortal
- Key to enabling semantic interoperability
- Enabling resource for the Semantic Web

Quality of mappings

- Created for a purpose
- Reusability issues
- Generally unidirectional
- Mapping from ontology 1 to ontology 2
- Not necessarily reversible

Comparing semantic descriptions

Semantic descriptions Consistent

- owl:sameAs
- N:Pancreas
- N:Pancretic disorder
- U:hasCUI
- C0000286
- C0000286
- N:Endocrine pancreas disorder
- N:subClassOf
- U:hasCUI
- C0000286
Comparing formal definitions

- Relatively small proportion of relata in common between equivalent concepts from NCIt and SNOMED CT
- Large number of primitive concepts in NCIt and SNOMED CT (70-80%)
- Insufficient for effectively comparing definitions
  - Could not be used for validating the mapping provided by the UMLS

Exercises

Exercise #1

- Check the equivalence (shared relata) between these 2 concepts:
  - NCI Thesaurus: N:Endocrine pancreas disorder
  - SNOMED CT: S:Disorder of endocrine pancreas

Exercise #2

- Find a correspondence in SNOMED CT for the LOINC term: Sodium:SCnc:-Pt:Ser/Plas:Qn
  [the molar concentration of sodium is measured in the plasma (or serum), with quantitative result]

Comments on exercise #2

- Difficult in the absence of a search mechanism on the values of the relations
- Large number of underspecified descriptions in SNOMED CT
- 2 separate concepts for plasma and serum concentrations of sodium in SNOMED CT
- Property, time and scale not represented in SNOMED CT
Thinking outside the integration box

The Butte approach

Integrating genomic and clinical data


Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

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