Short course – Summer 2010
Clinical Ontology in Practice
June 15-17, 2010

Clinical Ontology in Practice

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

Objectives

- Learn about clinical ontologies
  - History
  - Design principles, formalisms and tools
  - What are they?
  - What are they used for?
- Work with clinical ontologies
  - Search, browse, navigate, query with application programming interfaces
  - Analyze, compare
  - Specific clinical uses (e.g., decisions support, natural language processing, medication reconciliation, e-prescription)
  - Specific issues (e.g., mapping across ontologies, ontologies and information models)

Agenda

Tuesday, June 15 (lecture)
Introduction to Biomedical Ontologies
Design Principles, Formations and Tools for Biomedical Ontologies
Biomedical Ontologies - Content and structure - Function

Wednesday, June 16 (hands-on)
UMLS
SNOMED CT
LOINC
RxNorm
NDF-RT

Thursday, June 17 (discussion)
Decision support
Medication reconciliation
E-prescribing
Natural language processing
Mapping across ontologies
Value sets

References


Bio-ontology courses

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

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

Medical Ontology Research

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

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

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 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)

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)
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 Nosologia Methodica
(Cited by Chris Chute)

To support epidemiology
- John Graunt (1620-1674)
  - Analyzes the vital statistics of the citizens of London
- William Farr (1807-1883)
  - Medical statistician
  - Improves Cullen’s classification
  - Contributes to creating ICD
- Jacques Berthillon (1851-1922)
  - Chief of the statistical services (Paris)
  - Classification of causes of death (161 rubrics)

London Bills of Mortality
- John Graunt (1620-1674)
  - Analyzes the vital statistics of the citizens of London
- William Farr (1807-1883)
  - Medical statistician
  - Improves Cullen’s classification
  - Contributes to creating ICD

Limitations 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 denoted by three or four terms, and each term has been applied to as many different diseases; vague, inconvenient names have been employed, or complications have been registered instead of primary diseases. The nomenclature is of as much importance in this department of inquiry as weights and measures in the physical sciences, and should be settled without delay.”
– William Farr
First annual report.

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**

- 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

**Mouse Genome Database and GO**

- Entrez Gene
  - Gene/Ontology
  - GO

**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)
Biomedical ontology in PubMed

- Number of articles in PubMed/MEDLINE on Ontology vs. Controlled vocabulary

**Introduction to biomedical terminologies through an example**

**Guy's Hospital, London**

**Thomas Addison (1795-1860)**

- 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
  - Chronic adrenocortical insufficiency
- Contexts: different hierarchies

Internal Classification of Diseases

Medical Subject Headings

SNOMED CT

Dr. Olivier Bodenreider
National Library of Medicine
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
  - Addison's disease
  - Primary adrenocortical insufficiency
  - “Clinical synonymy” (vs. identity)
  - Abdominal swelling
  - Swollen abdomen
  - Posttransfusion hepatitis
  - Posttransfusion viral hepatitis
  - Addison's disease
  - Primary adrenocortical insufficiency vs. Waterhouse-Friderichsen Syndrome

Multiple classes for a term
- Polysynmy
- Truncated terms
  - Calcium
  - Ca^++
  - Calcitonin gene-related peptide
  - 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

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, medication, 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, physicians, 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

Terminological relations as a surrogate for ontological relations

Issues

- Lack of explicit classificatory principle
- Underspecification of the relations
- Thesaurus relations
- Limited depth in hierarchies “by design”
No explicit classificatory principle

- Attribute
- Body structure
- Clinical finding
- Context-dependent categories
- Environments and geographical locations
- Events
- Observable entity
- Organism
- Pharmaceutical / biologic product
- Physical force
- Physical object
- Procedure
- Qualifier value
- Social context
- Special concept
- Specimen
- Staging and scales
- Substance

Fully specified relations

Viral meningitis in SNOMED CT

Underspecification of the relations

Thesaurus relations

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

Relations used to create hierarchical structures vs. hierarchical relations
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, …

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

Overview

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

Introduction

- 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

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

Thesaurus

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

Thesaurus vs. ontology

(Controlled) Terminology

- Objective: naming things
- Shared understanding
  - Agreement on what terms to use
  - Utility-driven (arbitrary)
  - Telephone call by a physician to patient or for consultation or medical management or for coordinating medical management with other health care professionals (e.g., nurses, therapists, social workers, nutritionists, physicians, pharmacists), complex or lengthy (e.g., 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

Thesaurus

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

Classification

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

Dr. Olivier Bodenreider
National Library of Medicine
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
    - UMLS SPECIALIST
    - Lexicon, WordNet
- Ontological resources
  - Collections of
    - kinds of entities (substances, qualities, processes)
    - relations among them
  - Useful for relation extraction
    - UMLS Semantic Network, BioTop
- Terminological resources
  - Collections lexical items + identifiers
    - 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)
  - Thing
  - Continuant
  - Occurrent
  - Independent continuant
  - Dependent continuant

Particulars (instances)

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

Dr. Olivier Bodenreider
National Library of Medicine
Top level in WordNet

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

Top level in 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

Top level in the Semantic Network

UMLS Semantic Network

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

Differences between ontologies

Examples

Dr. Olivier Bodenreider
National Library of Medicine
Granularity, plesionymy

Differing categorization

Ontology and Formalism

Formalisms and Tools

Tools for ontology developers
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)

International Classification of Diseases

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

ICD Example

◆ Idiosyncratic inclusion/exclusion criteria

Dr. Olivier Bodenreider
National Library of Medicine
ICD Example

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E21.9</td>
<td>Cystic fibrosis</td>
</tr>
<tr>
<td>E21.90</td>
<td>Cystic fibrosis with pulmonary manifestations</td>
</tr>
<tr>
<td>E21.91</td>
<td>Cystic fibrosis with bronchodilatory manifestations</td>
</tr>
<tr>
<td>E21.912</td>
<td>(Cystic fibrosis with lung disease)</td>
</tr>
<tr>
<td>E21.919</td>
<td>Cystic fibrosis with other manifestations</td>
</tr>
<tr>
<td>E21.99</td>
<td>Cystic fibrosis, unspecified</td>
</tr>
</tbody>
</table>

Logical Observation Identifiers, Names and Codes (LOINC)

LOINC Characteristics (1)

- Current version: 2.30 (Feb. 2010)
- 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](http://www.regenstrief.org/loinc/loinc.htm)

LOINC Characteristics (2)

- Number of
  - Concepts: 50k active codes (2.18)
  - Terms: n/a*
- Major organizing principles:
  - No hierarchical structure among the main codes
  - 6 axes
    - Component (analyte [+ challenge] [+ adjustments])
    - Property
    - Timing
    - System
    - Scale
    - [Method]
- Formalism: “DL-like”

LOINC Example

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

<table>
<thead>
<tr>
<th>Axis</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component</td>
<td>Sodium</td>
</tr>
<tr>
<td>Property</td>
<td>SCnc - Substance Concentration (per volume)</td>
</tr>
<tr>
<td>Timing</td>
<td>Pt - Point in time (Random)</td>
</tr>
<tr>
<td>System</td>
<td>Ser/Plas - Serum or Plasma</td>
</tr>
<tr>
<td>Scale</td>
<td>Qn - Quantitative</td>
</tr>
<tr>
<td>Method</td>
<td></td>
</tr>
</tbody>
</table>

SNOMED Clinical Terms

Dr. Olivier Bodenreider
National Library of Medicine
SNOMED CT Characteristics (1)

- Current version: January 31, 2010 (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: ~310,000 active concepts (Jan. 31, 2010)
  - Terms: ~800,000 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

- 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

**FMA Top level**

(Courtesy of C. Rosse)

**Gene Ontology**

(Courtesy of C. Rosse)

**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

**Gene Ontology Characteristics (2)**

- **Number of**
  - Concepts: 27,800 (July 22, 2009)
  - 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
RxNorm Characteristics (2)

- Number of concepts: 166k
- Terms: ~1 term per concept
- 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>

RxNorm 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

MeSH Characteristics (1)

- Current version: 2010 (yearly releases)
- Type: Thesaurus / Controlled vocabulary
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (Library Operations)
- Availability:
  - Publicly available: Yes
  - Repositories: UMLS / NCBO BioPortal
- URL: http://www.nlm.nih.gov/mesh/
MeSH Characteristics (2)

- Number of
  - Terms: 7.5 per descriptor

- Major organizing principles:
  - Descriptor + entry terms
  - Thesaurus relations (RB/BN/RO)
- Formalism: Thesaurus / Proprietary XML DTD

MeSH Top level

1. Anatomy [A]
2. Organisms [O]
3. Literature [L]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [F]
6. Psychiatry and Psychology [P]
7. Biological Sciences [G]
8. Chemical Substances [C]
10. Technology, Industry, Agriculture [T]
11. Humanities [H]
12. Information Science [I]
13. Named Groups [N]
14. Health Care [B]
15. Publication Characteristics [F]
16. Geographical [Z]

MeSH Example (terms)

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>MeSH Major heading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polycyclic Compounds</td>
<td>Steroids</td>
</tr>
<tr>
<td>Steroids</td>
<td>Hormones, Hormone Substitutes, and Hormone Antagonists</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>Adrenal Cortex Hormones</td>
</tr>
<tr>
<td>Adrenal Cortex Hormones</td>
<td>11-Hydroxycorticosteroids</td>
</tr>
<tr>
<td>11-Hydroxycorticosteroids</td>
<td>11-Hydroxycorticosteroids</td>
</tr>
<tr>
<td>11-Hydroxycorticosteroids</td>
<td>Hydrocortisone</td>
</tr>
</tbody>
</table>

MeSH Example (hierarchies)

NCI Thesaurus

NCI Thesaurus Characteristics (1)

- Current version: 10.05d (~monthly releases)
- Type: Controlled terminology / ontology
- Domain: Cancer
- Developer: NCI Center for Bioinformatics
- Funding: NCI
- Availability
  - Publicly available: Yes
  - Repositories: UMLS / OBO / NCBO BioPortal
- URL: http://nciterms.nci.nih.gov/
NCI thesaurus  Characteristics (2)

- Number of
  - Concepts: ~60,000
  - 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  Characteristics (1)

- Current version: 2010AA (2 annual releases)
- Type: Terminology integration system
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (intramural)
- Availability
  - Publicly available: Yes* (cost-free license required)
  - Repositories: UMLS
- URL: http://umlsks.nlm.nih.gov/

NCI thesaurus  Top level

NCI thesaurus Taxonomy

- Abnormal Cell
- Anatomy
- Biochemistry, Metabolism, or Substance
- Biological Process
- Chemical Therapeutic Agent Classification
- Conceptual Entity
- Diagnostic, Therapeutic, and Research Equipment
- Diagnostic or Therapeutic Device
- Drug, Drug Class, or Biologic Material
- Pharmacologic or Chemotherapeutic Concept
- Experimental Organism, Anatomical Concept
- Organ
- Procedure
- Substance

Unified Medical Language System (UMLS)

UMLS  Characteristics (1)

- Type: Terminology integration system
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (intramural)
- Availability
  - Publicly available: Yes* (cost-free license required)
  - Repositories: UMLS
- URL: http://umlsks.nlm.nih.gov/

UMLS  Characteristics (2)

- Number of
  - Concepts: 2.2M (2010AA)
  - Terms: ~10M
- Major organizing principles (Metathesaurus):
  - Concept orientation
  - Source transparency
  - Multi-lingual through translation
- Formalism: Proprietary format (RRF)
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).

Knowledge management

Annotating data and resources

Terminology in ontology

- 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

Annotations

- 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
  - Inferrable automatically (e.g., from sequence similarity)

GO Annotations for Aldh2 (mouse)

<table>
<thead>
<tr>
<th>Category</th>
<th>Classification Terms</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Function</td>
<td>Aldh2_MOUSE</td>
<td>VIA</td>
</tr>
<tr>
<td>Molecular Function</td>
<td>Aldh2_MOUSE</td>
<td>VIA</td>
</tr>
<tr>
<td>Molecular Function</td>
<td>Aldh2_MOUSE</td>
<td>VIA</td>
</tr>
<tr>
<td>Molecular Function</td>
<td>Aldh2_MOUSE</td>
<td>VIA</td>
</tr>
<tr>
<td>Cellular Component</td>
<td>Aldh2_MOUSE</td>
<td>VIA</td>
</tr>
<tr>
<td>Biological Process</td>
<td>Aldh2_MOUSE</td>
<td>VIA</td>
</tr>
<tr>
<td>Biological Process</td>
<td>Aldh2_MOUSE</td>
<td>VIA</td>
</tr>
</tbody>
</table>

http://www.informatics.jax.org/
GO ALD4 in Yeast

GO Annotations
- ALD4 is involved in: [View Compositional GO annotations for ALD4]
- Molecular Function: Alcohol dehydrogenase activity (OA, Imp, 55)
- Biological Process: Ethanol metabolic process (IMP)
- Cellular Component: Interactome protein (OA, Imp, 55)
- Interactome (OA, Imp, 55)

http://db.yeastgenome.org/

GO Annotations for ALDH2 (Human)

- GO:0006320: alcohol dehydrogenase (EC 1.1.1.1) (ASD)
- GO:0005608: electron carrier activity (EC 1.1.1.1)

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)
  - Automatic indexing systems

MeSH MEDLINE indexing

- MeSH Terms
  - Alcohol Dehydrogenase [Enzyme Activity]
  - Alcohol Dehydrogenase Activity
  - Alcohol Dehydrogenase [EC 1.1.1.1]
  - Alcohol Oxidoreductase [EC 1.1.1.1]
  - Alcohol Oxidoreductase (Class I)
  - Alcohol Oxidoreductase (Class II)
  - B1 Alcohol Dehydrogenase
  - Diabetic Nephropathy
  - Glucosamine
  - Glucose-6-Phosphate Dehydrogenase
  - Glucose Oxidase
  - Glucose Oxidation
  - Glucose Oxidation (EC 1.1.1.1)
  - Glucose Oxidation (Class I)
  - Glucose Oxidation (Class II)
  - MeSH Terms
    - Alcohol Dehydrogenase
    - Alcohol Dehydrogenase [Activity]
    - Alcohol Dehydrogenase [EC 1.1.1.1]
    - Alcohol Oxidoreductase [EC 1.1.1.1]
    - Alcohol Oxidoreductase (Class I)
    - Alcohol Oxidoreductase (Class II)
    - B1 Alcohol Dehydrogenase
    - Diabetic Nephropathy
    - Glucosamine
    - Glucose-6-Phosphate Dehydrogenase
    - Glucose Oxidase
    - Glucose Oxidation
    - Glucose Oxidation (EC 1.1.1.1)
    - Glucose Oxidation (Class I)
    - Glucose Oxidation (Class II)

MeSH MEDLINE indexing

- MeSH Terms
  - Alcohol Dehydrogenase [Enzyme Activity]
  - Alcohol Dehydrogenase Activity
  - Alcohol Dehydrogenase [EC 1.1.1.1]
  - Alcohol Oxidoreductase [EC 1.1.1.1]
  - Alcohol Oxidoreductase (Class I)
  - Alcohol Oxidoreductase (Class II)
  - B1 Alcohol Dehydrogenase
  - Diabetic Nephropathy
  - Glucosamine
  - Glucose-6-Phosphate Dehydrogenase
  - Glucose Oxidase
  - Glucose Oxidation
  - Glucose Oxidation (EC 1.1.1.1)
  - Glucose Oxidation (Class I)
  - Glucose Oxidation (Class II)

- Indexing
  - Performed manually by human indexers
  - With help of semi-automatic systems (suggestions)
  - Automatic indexing systems

- MEDLINE indexing
  - Performed manually by human indexers
  - With help of semi-automatic systems (suggestions)
  - Automatic indexing systems
MeSH MEDLINE indexing

- MeSH Terms
  - Addis Disease
  - Addis Disease, Blood
  - Addis Disease, Blood Coagulation Factor
- MeSH Entry Terms
  - Addis Disease
  - Addis Disease, Blood
  - Addis Disease, Blood Coagulation Factor

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
- Translation

MeSH “synonyms” MEDLINE retrieval

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

Knowledge management
Mapping across biomedical ontologies

Reusing information

◆ Clinical information coded with SNOMED CT
  ● Mapped to ICD-9-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 (MetaMap)  [Aronson et al., JAMA, 2010]
  ◆ 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**

- Clinical repositories
- Genetic knowledge bases
- Other subdomains
- MeSH
- Biomedical literature
- NCB Termaining
- FMA
- GO
- Anatomy
- Genome annotations

**Trans-namespace integration**

- Clinical repositories
- Genetic knowledge bases
- Other subdomains
- MeSH
- Biomedical literature
- NCB Terminology
- FMA
- GO
- Anatomy
- Genome annotations

**Data integration, exchange and semantic interoperability**

*Information 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 Oncologic Representation Environment
  - Infrastructure developed to support an interoperable biomedical information system for cancer research
  - Uses the NCI Thesaurus as a component

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)

Data integration, exchange and semantic interoperability

Information and data integration
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.

Decision support and reasoning

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 NCI)

Dr. Olivier Bodenreider
National Library of Medicine
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

Availability

- 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: 156 source vocabularies (biased towards healthcare applications)
  - NCBO BioPortal: ~200 ontologies (biased towards biological applications)
  - Some overlap between the two repositories
- Need for discovery services

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)
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
  - Current harmonization efforts (e.g., IHTSDO)
- Coordinated development of ontologies
  - OBO Foundry approach
  - Ensures consistency ab initio
  - Excludes legacy ontologies

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)
    - Marginal notes enabled by BioPortal
  - Centrally, by experts
    - OBO Foundry approach
- Important factors besides quality
  - Governance
  - Installed base / Community of practice

UMLS UMLSKS

- **UMLSKS** (Knowledge Source Server)
- Search by term: appendectomy (C0003611)
  - (default) RRF view (atom-centric)
  - Lexical View (normalized strings / lexical units)
  - Relations
  - Co-occurrence Info
  - Contexts (paths to root)
- Search by code
  - R73.0 (Postprocedural hypoinsulinaemia)

UMLS Semantic Navigator

- Available under UMLSKS
  (bottom of left-hand side pane)
- Search by term:
  - appendectomy (C0003611)
  - Addison’s disease (C0001403)
- Concept-centric vs. atom-centric
- Selection of hierarchical relations (and co-occurrences)
- Transitive reduction on/off

Notes

- Ambiguity: appendectomy, heart, calcium
- Several kinds of lexical matches (exact, normalized, approximate)
UMLSKS API

- Authentication vs. UMLS KS services
- SOAP-based (examples and documentation mostly for Java, but usable with other environments, e.g., Perl, .NET)

SNOMED CT

- Multiple web-based browsers available
  - U. Sydney browser (specific to SNOMED CT)
    http://loinc.org/relma
  - Virginia Tech browser (specific to SNOMED CT)
    http://terminology.vetmed.vt.edu/SCT/menu.cfm
  - The SNOMED CT Browser © (specific to SNOMED CT)
    http://www.medicalclassifications.com/SNOMEDBrowser/
  - BioPortal
    http://www.bioontology.org/BioPortal
  - NCI Term Browser
    http://nciters.nci.nih.gov/

SNOMED CT

- Search concepts
  - Appendectomy (80146002)
  - Simvastatin (387584000)
  - Addison's disease (363732003)
- Notes
  - No post-coordination services in standard browsers
  - Some standalone browsers offer additional services
    (CliniCue, SNOB)
  - Search on Addison's disease in The SNOMED CT Browser © does not return any results

LOINC

- Multiple web-based browsers available
  - RELMA (specific to LOINC)
    Web version of a standalone application
    http://loinc.org/relma
  - NB: Citrix ICA Client required
  - BioPortal (LOINC 2.26)
    http://www.bioontology.org/BioPortal
  - NCI Term Browser (LOINC 2.24)
    http://nciters.nci.nih.gov/

LOINC BioPortal

- BioPortal
  - Graphical interface
    - Search for Lithium, then navigate down the tree
  - Web services
    http://www.bioontology.org/wiki/index.php/NCBO_REST_services
    - Ontology ID: 1350
    - Get ID for latest version
      - http://rest.bioontology.org/bioportal/virtual/ontology/1350
    - Returns: 40400
    - Get the “first” 50 terms
      - http://rest.bioontology.org/bioportal/concepts/40400/all?pageSize=50&program=1

LOINC NCI Term Browser

- NCI Term Browser
  - Search for Lithium, then navigate through the Relationships tab
  - Search by code
  - Search concept
    - Substance concentration of lithium in urine (quantitative)
    - Lithium/Substance Concentration:Point in time:Urine:Quantitative
    - 25463-1

Dr. Olivier Bodenreider
National Library of Medicine
RxNorm RxNav

- RxNav
    - [launch the browser]
- Search by string (default): zytec, clopidogrel
  - Restrict the graph to one particular clinical drug: double-click on Cetirizine 10 MG Oral Tablet
  - RxCUI is displayed in the information bar in the bottom when clicking on a drug entity (e.g., RxCUI for Cetirizine 10 MG Oral Tablet = 309130)
  - Right-click on Cetirizine 10 MG Oral Tablet
  - View NDCs to open a window with the list of NDCs for this drug
  - View Drug Label → link out to DailyMed

Send feedback

Documentation

Functions

- getAllRelatedInfo(309130)
- findRxcuiById(00904582941)
- getIdTypes()
- getRxNormVersion()

NDC, with search string 1039008
- Returns: 103|C0000618||6—Mercaptopurine

Other example: cetirizine
- Double-click on cetirizine 75 MG Oral Tablet
- Click on the NDF-RT tab
- Explore the relations to the different categories of entities (Drug, Disease, Dose form, …)

RxNorm SOAP API

- RxNorm SOAP API (demo client)
- Functions
  - getRxNormVersion()
  - getRxcuiById(00904582941, 309130) → 309130
  - getAllRelatedInfo(309130)
- Documentation

Packs: Search for Packs: Search for zpak, , by" "menu
- Search by ID (select ID in the drop-down “Search by” menu
  - NDC, with search string 00781168401 (one of the NDC from the list obtained from Cetirizine 10 MG Oral Tablet)
  - SNOMED ID, with search string 1039008
    - Pack: Search for z-pak
      - Packs displayed with double diamonds in the clinical drug / generic pack and branded drug / branded pack

RxNorm REST API

- Test resources
- Documentation

NDF-RT

- RxNav (pilot version integrating NDF-RT)
- Search for clopidogrel (RxNorm tab)
  - other example: cetirizine
    - Double-click on cetirizine 75 MG Oral Tablet
    - Click on the NDF-RT tab
    - Explore the relations to the different categories of entities (Drug, Disease, Dose form, …)
Topics

- NLP / indexing
- PHR / consumer health information
- Decision support (drugs)
- Decision support (other)
- Medication reconciliation
- E-prescribing
- CPOE
- Problem list
- Terminology services
- Value sets
- Terminology management (versioning)
- Mapping / integration
- Meaningful use
- Health information exchange
- Clinical documentation

Questions

- What are some of the issues and challenges related to this topic?
- Do ontologies contribute to the solution? Which ones? Which features?
- Have you learned anything that is applicable towards this issue?

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