Workshop on Foundations of Clinical Terminologies and Classifications (FCTC 2006)
Timișoara, Romania, April 8, 2006

Biomedical terminology and beyond

Ontology and terminology services

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

- Why biomedical terminologies?
- Building biomedical terminologies: Recent experiences
- Terminology vs. ontology
- Terminology services
Why biomedical terminologies?
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 Methodicae

(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
“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 organisms (Fly, Mouse, Yeast, …)
  • Manually / semi-automatically annotated
  • Gene Ontology
Black bile and psychomotor retardation: shades of melancholia in Dante's Inferno.

Widmer DA.

Memorial Sloan-Kettering Cancer Center, New York, NY 10017, USA. widmerd@mskcc.org

The history of melancholy depression is rich with images of movement retardation and mental dysfunction. The recent restoration of psychomotor symptoms to the diagnostic terminology of affective disorder is not novel to the students of medieval melancholia. The move back to the biology of this psychomotor dysfunction with the technical advances in brain imaging in recent years only echoes centuries-old writings on the centrality of movement changes in the depressive condition. The Inferno, the first cantica of Dante Alighieri's Commedia, has a wonderful abundance of allusions to the importance of psychomotor symptoms in describing the depressed individual. Slowed steps, garbled speech, frozen tears, these and many other images keep the physical manifestations of psychomotor suffering in the forefront of the reader's mind. Considering Medieval and Renaissance writings on melancholy suffering, it is fitting that Dante shows a bodily illness reflected in the hellish torments visited on the damned. From the souls of the sullen to those of the violent, the panorama of psychomotor symptoms plays a prominent role in the poem as well as in the medical and literary prose of succeeding centuries.

MeSH Terms:
- Depressive Disorder/history*
- History of Medicine, Medieval
- Human
- Italy
- Literature, Medieval/history*
- Medicine in Literature*
- Poetry/history*
- Psychomotor Disorders/history*
Mouse Genome Database and GO

Entrez Gene

1. Neurofibromatosis 2 (Mus musculus)
   GeneID: 18016  Locus tag: MGI:97307

Gene Ontology

Function
- cytoskeletal protein binding
- protein binding
- structural molecule activity

Process
- intercellular junction assembly and/or maintenance
- negative regulation of cell cycle
- negative regulation of protein kinase activity
- regulation of cell proliferation

Component
- adherens junction
- cytoplasm
- cytoskeleton
- membrane
To serve as a reference

- **Reference terminology/ontology**
  - Universally needed
  - Developed independently of any purposes
  - Reusable by many applications

- **Examples**
  - RxNorm
  - Foundational Model of Anatomy (FMA)
  - SNOMED CT
  - ChEBI
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)
Building biomedical terminologies

Recent experiences
Building biomedical terminologies

Recent experiences

- Description logics approach
- Reengineering terminologies with DL
- Reorganizing MeSH
- Gene Ontology
- UMLS SemanticNetwork
Description logics approach

- Pioneered by GALEN
  - Although GALEN itself is not a terminology

- SNOMED CT
  - Although it is distributed as a relational database (terms, relations), not in DL format

- DL is used to support the creation of terminologies
- The goal is not to have terminologies in OWL
Benefits of using a DL approach

◆ Consistent organization
  ● Equivalent classes
  ● Automatic classification
  ● Error detection through reclassification
  ● …

◆ But DL does nothing for the naming component of terminologies
  ● Inconsistent synonyms for anatomical concepts in SNOMED CT (Structure/Entire)
Reengineering terminologies with DL

- Ontologizing terminologies
  - e.g., UMLS
    - Metathesaurus
    - Semantic Network
  - [Hahn, PSB 2003], [Cornet, AMIA 2002], [Pisanelli, AMIA 1998], [Kashyap, ISWC 2003]

- Migrating to OWL
  - NCI Thesaurus
  - Gene Ontology
  - MeSH
  - FMA
  - [Golbeck, JWS 2003], [Wroe, PSB 2003], [Soualmia, KE-MED 2004], [Golbreich, OWLED 205]
Reengineering with DL  Limitations

- No trivial isomorphism
- Never purely a matter or formalism
  - Not every thesaurus relation should become *isa*
  - Necessary and sufficient conditions for anatomical structures?
- Never completely automatic
- Costly in terms of human resources

**Terminology + formalism ≠ Formal terminology**
# Reorganizing MeSH

<table>
<thead>
<tr>
<th>MeSH Heading</th>
<th>Entry Term</th>
<th>Unique ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerebrovascular Accident</td>
<td>Apoplexy</td>
<td>D020521</td>
</tr>
<tr>
<td>Cerebrovascular Accident</td>
<td>Cerebral Stroke</td>
<td></td>
</tr>
<tr>
<td>Cerebrovascular Apoplexy</td>
<td>Stroke</td>
<td></td>
</tr>
<tr>
<td>Vascular Accident, Brain</td>
<td>CVA (Cerebrovascular Accident)</td>
<td></td>
</tr>
<tr>
<td>Cerebrovascular Accident, Acute</td>
<td>Cerebrovascular Stroke</td>
<td></td>
</tr>
<tr>
<td>Stroke, Acute</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Concepts and Aggregates**

- **M0328070**
- **M0334621**
- **D020521**
Gene Ontology

- Developed by biologists in the early 2000’s
- Extremely popular
  - Genome annotation across model organism databases
- Simplistic
  - No relations across hierarchies
  - Only *isa* and *part_of* relationships

- Being reengineered/ontologized
  - OBOL (formal language for representing lexical relations)
  - National Center for Biomedical Ontology
  - Relations across hierarchies will be added
UMLS Semantic Network

◆ Weak (some-some) semantics
◆ Metathesaurus concepts linked to semantic types, but no link between MT and SN relationships

◆ Being reanalyzed from the perspective of formal ontology
  • e.g., distinction between continuants and occurrents
  • Mapping of relationships between MT and SN
Terminology vs. Ontology
Terminology vs. Ontology

◆ Types of resources
  - Lexical
  - Terminological
  - Ontological

◆ Ontology is overloaded

◆ Terminology is overloaded too

◆ Formal approaches to terminology
Lexical vs. ontological 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, SNOMED CT
Types of resources revisited

◆ Lexical and terminological resources
  ● Mostly collections of names for biomedical entities
  ● Often have some kind or hierarchical organization (e.g., relations)

◆ Ontological resources
  ● Mostly collections of relations among biomedical entities
  ● Sometimes also collect names
Unified Medical Language System

◆ SPECIALIST Lexicon
  ● 200,000 lexical items
  ● Part of speech and variant information

◆ Metathesaurus
  ● 5M names from over 100 terminologies
  ● 1M concepts
  ● 16M relations

◆ Semantic Network
  ● 135 high-level categories
  ● 7000 relations among them
Ontology is overloaded

- **Hype**
- **Not every ontology built**
  - is formal
  - has definitions
  - is consistent
  - ...
- **Not everything in OWL (resp. Protégé) is an ontology**
Terminology is overloaded too Terms

“Terms” are not necessarily named for biomedical entities

- 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, 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 aspects of the total patient care plan)
Terminology is overloaded too

- Hierarchical structures created to support a task
  e.g., information retrieval for MeSH
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
Not all “isa” relations are transitive!

Terminologies do not necessarily support reasoning
Endocrine Diseases [C19]

Adrenal Gland Diseases [C19.053]

Adrenal Gland Hypofunction [C19.053.264]

- Addison's Disease [C19.053.264.263]
- Adrenoleukodystrophy [C19.053.264.270]
- Hypoaldosteronism [C19.053.264.480]

Immunologic Diseases [C20]

Autoimmune Diseases [C20.111]

- Addison's Disease [C20.111.163]
- Anemia, Hemolytic, Autoimmune [C20.111.175]
- Anti-Glomerular Basement Membrane Disease [C20.111.190]
- Antiphospholipid Syndrome [C20.111.197]
- Arthritis, Rheumatoid [C20.111.199] +

Hierarchy

- Adrenal cortical hypofunction
- Addison's disease
- Addison's disease due to autoimmunity
- Addison's disease with adrenoleukodystrophy
- Polyglandular autoimmune syndrome, type 1
- Tuberculous Addison's disease
Housekeeping relations

◆ Obsolete terms
  - Maintained in the terminology (permanence principle)
  - Linked to special “housekeeping” concepts

[Diagram showing relationships between concepts: Special concept, Inactive concept, Duplicate concept, [D]Numbness]
Formal approaches to terminology

◆ Computational terminology
  ● Tasks
    ■ Identifying terms from text corpora automatically
    ■ Organizing terms automatically
  ● Methods
    ■ Lexicosyntactic and semantic analysis
    ■ Machine learning
    ■ Information science

◆ Limited interest in biomedicine because of the existence of comprehensive terminologies
Terminology services
Terminology services

- Defining terminology services
- Lexical issues
- Ontological issues
The GALEN terminology server

- Managing external references
- Managing internal representations
- Mapping natural language to concepts
- Mapping concepts to classification schemes
- Management of extrinsic information

[Rector, Methods 1995]
Chris Chute’s desiderata

- Word normalization
- Word completion
- Spelling correction
- Lexical matching
- Term completion
- Target terminology specification
- Semantic locality
- Term composition
- Term decomposition

[Chute, AMIA 1999]
Requirements

◆ Model of the term
  ● Lexico-syntactic level (lexical resemblance)
    ■ Supported by lexicons
      – Word properties
    ■ Edit distance for spelling correction
    ■ Rules for normalization (defining inessential features)
  ● Semantic level (semantic similarity)
    ■ Supported by ontologies
      – Concept properties
      – Relations to other concepts
    ■ Constraints for composition
Requirements (continued)

- Model of the mapping
- Model of the task (context of use)

- Other terminology services
  - Subsetting terminologies
  - Helping define value sets
  - Self-generating terminologies (from orthogonal ontologies)
  - Extending terminologies
Lexico-syntactic level

- Lots of developments in the past 15 years
- Stable for English
  - UMLS SPECIALIST Lexicon
  - Lexical tools (e.g., lvg, spelling correction module)
- Underway for other languages
  - Spanish (NLM)
  - German (Freiburg)
  - French (UMLF)

[McCray, AMIA 1994]
Normalization

1. Remove genitive: Hodgkin’s diseases, NOS
2. Remove stop words: Hodgkin diseases, NOS
3. Lowercase: Hodgkin diseases,
4. Strip punctuation: hodgkin diseases,
5. Uninflect: hodgkin diseases
6. Sort words: hodgkin disease
7. Result: disease hodgkin
Normalization: Example

- Hodgkin Disease
- HODGKINS DISEASE
- Hodgkin's Disease
- Disease, Hodgkin's
- Hodgkin's, disease
- HODGKIN'S DISEASE
- Hodgkin's disease
- Hodgkins Disease
- Hodgkin's disease NOS
- Hodgkin's disease, NOS
- Disease, Hodgkins
- Diseases, Hodgkins
- Hodgkins Diseases
- Hodgkins disease
- hodgkin's disease
- Disease, Hodgkin

normalize → disease hodgkin
Lexical issues

- Normalization was developed essentially for clinical terms
- Known issues
  - Drug names
  - Chemicals
- New issues with biological corpora
  - Gene names
Semantic level

- Limited progress in the past 15 years
- Single most important contribution: SNOMED CT
- Main source of labeled relations in the UMLS i.e., explicit classificatory criteria
- Few other vocabularies in the UMLS contribute labeled relations in large numbers
  - NDFRT
  - RxNorm
Explicit classificatory principle

Foundational Model of Anatomy

- Anatomical structure
- Body substance

Material physical anatomical entity

- Inherent 3D shape

- Anatomical structure
- Body substance

Physical anatomical entity

- Spatial dimension
  - +
  - -

- Mass
  - +
  - -

Non-material physical anatomical entity

- Non-physical anatomical entity
- Non-material physical anatomical entity

- 3D
- 2D
- 1D
- 0D

Anatomical entity

- Anatomical structure
- Body substance

Non-physical anatomical entity
No explicit classificatory principle

3. **Diseases [C]**

- Bacterial Infections and Mycoses [C01]
- Virus Diseases [C02]
- Parasitic Diseases [C03]
- Neoplasms [C04]
- Musculoskeletal Diseases [C05]
- Digestive System Diseases [C06]
- Stomatognathic Diseases [C07]
- Respiratory Tract Diseases [C08]
- Otorhinolaryngologic Diseases [C09]
- Nervous System Diseases [C10]
- Eye Diseases [C11]
- Urologic and Male Genital Diseases [C12]
- Female Genital Diseases and Pregnancy Complications [C13]
- Cardiovascular Diseases [C14]
- Hemic and Lymphatic Diseases [C15]
- Neonatal Diseases and Abnormalities [C16]
- Skin and Connective Tissue Diseases [C17]
- Nutritional and Metabolic Diseases [C18]
- Endocrine Diseases [C19]
- Immunologic Diseases [C20]
- Disorders of Environmental Origin [C21]
- Animal Diseases [C22]
- Pathological Conditions, Signs and Symptoms [C23]

agent/cause

location

stage in life
Semantic issues

- Lack of classificatory principles explicitly stated and represented in ontologies
- Lack of trans-ontological (associative) relations represented in ontologies

Result in

- Inconsistent representations
  - e.g., Prevention of X / X
- Maintenance issues
  - e.g., modification of a given term should trigger the review of dependent terms
Summary
Summary

◆ Terminology vs. ontology

◆ Terminology vs. terminology services

◆ Usefulness vs. elegance
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