Ontologies and Biosurveillance

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

Public Health Informatics Fellowship Program
March 12, 2010
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.
Outline

◆ Biomedical ontologies
  ● What they are
  ● What they are for

◆ Ontologies and biosurveillance
  ● Support for text mining
  ● Controlled vocabulary
  ● Data aggregation
  ● Data integration
  ● Reasoning

◆ Ontology in action in a biosurveillance system – BioCaster
Biomedical ontologies

What they are
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)

[J. Cimino, YBMI 2006]
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/](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
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Blocks</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>A00-B99</td>
<td>Certain infectious and parasitic diseases</td>
</tr>
<tr>
<td>II</td>
<td>C00-D48</td>
<td>Neoplasms</td>
</tr>
<tr>
<td>III</td>
<td>D50-D89</td>
<td>Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism</td>
</tr>
<tr>
<td>IV</td>
<td>E00-E90</td>
<td>Endocrine, nutritional and metabolic diseases</td>
</tr>
<tr>
<td>V</td>
<td>F00-F99</td>
<td>Mental and behavioural disorders</td>
</tr>
<tr>
<td>VI</td>
<td>G00-G99</td>
<td>Diseases of the nervous system</td>
</tr>
<tr>
<td>VII</td>
<td>H00-H59</td>
<td>Diseases of the eye and adnexa</td>
</tr>
<tr>
<td>VIII</td>
<td>H60-H95</td>
<td>Diseases of the ear and mastoid process</td>
</tr>
<tr>
<td>IX</td>
<td>I00-I99</td>
<td>Diseases of the circulatory system</td>
</tr>
<tr>
<td>X</td>
<td>J00-J99</td>
<td>Diseases of the respiratory system</td>
</tr>
<tr>
<td>XI</td>
<td>K00-K93</td>
<td>Diseases of the digestive system</td>
</tr>
<tr>
<td>XII</td>
<td>L00-L99</td>
<td>Diseases of the skin and subcutaneous tissue</td>
</tr>
<tr>
<td>XIII</td>
<td>M00-M99</td>
<td>Diseases of the musculoskeletal system and connective tissue</td>
</tr>
<tr>
<td>XIV</td>
<td>N00-N99</td>
<td>Diseases of the genitourinary system</td>
</tr>
<tr>
<td>XV</td>
<td>O00-O99</td>
<td>Pregnancy, childbirth and the puerperium</td>
</tr>
<tr>
<td>XVI</td>
<td>P00-P96</td>
<td>Certain conditions originating in the perinatal period</td>
</tr>
<tr>
<td>XVII</td>
<td>Q00-Q99</td>
<td>Congenital malformations, deformations and chromosomal abnormalities</td>
</tr>
<tr>
<td>XVIII</td>
<td>R00-R99</td>
<td>Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified</td>
</tr>
<tr>
<td>XIX</td>
<td>S00-T98</td>
<td>Injury, poisoning and certain other consequences of external causes</td>
</tr>
<tr>
<td>XX</td>
<td>V01-Y98</td>
<td>External causes of morbidity and mortality</td>
</tr>
<tr>
<td>XXI</td>
<td>Z00-Z99</td>
<td>Factors influencing health status and contact with health services</td>
</tr>
<tr>
<td>XXII</td>
<td>U00-U99</td>
<td>Codes for special purposes</td>
</tr>
</tbody>
</table>
ICD Example

*Idiosyncratic inclusion/exclusion criteria*

**E10**

*Insulin-dependent diabetes mellitus*

[See before E10 for subdivisions.]

**Includes:**
- brittle
- juvenile-onset
- ketosis-prone
- type I

**Excludes:**
- diabetes mellitus (in): malnutrition-related (**E12.-**)
- neonatal (**P70.2**)
- pregnancy, childbirth and the puerperium (**O24.-**)
- glycosuria:
  - NOS (**R81**)
  - renal (**E74.8**)
- impaired glucose tolerance (**R73.0**)
- postsurgical hypoinsulinaemia (**E89.1**)
# ICD Example

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>E84.0</td>
<td>Cystic fibrosis with pulmonary manifestations</td>
</tr>
<tr>
<td></td>
<td><em>Includes</em>: mucoviscidosis</td>
</tr>
<tr>
<td>E84.1</td>
<td>Cystic fibrosis with intestinal manifestations</td>
</tr>
<tr>
<td></td>
<td>Meconium ileus+ (P75*)</td>
</tr>
<tr>
<td></td>
<td><em>Excludes</em>: meconium obstruction in cases where cystic fibrosis is known not to be present (P76.0)</td>
</tr>
<tr>
<td>E84.8</td>
<td>Cystic fibrosis with other manifestations</td>
</tr>
<tr>
<td>E84.9</td>
<td>Cystic fibrosis, unspecified</td>
</tr>
</tbody>
</table>
Logical Observation Identifiers, Names and Codes (LOINC)
LOINC Characteristics (1)

- Current version: 2.27 (July 2009)
- 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
SNOMED CT Characteristics (1)

- Current version: January 31, 2009 (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 Example

Hierarchy:
- 27010001 partial excision of large intestine
- 8613002 operation on appendix
- 80146002 appendectomy
- 82730006 incidental appendectomy
- 49438003 appendectomy with drainage
- 174036004 emergency appendectomy
- 174045003 interval appendectomy
- 6025007 laparoscopic appendectomy
- 235313004 non-emergency appendectomy
- 235314005 inversion appendectomy
- 1299000 excision of appendiceal stump

Definition: Fully defined by...
- partial excision of large intestine
- operation on appendix
- method: excision - action
- procedure site: Direct appendix
- qualifier: surgical access values
- priority: priorities

Codes:
- Original SnomedId: P1-57450
- Read Code: Ctv3ld: X20Wz
RxNorm
**RxNorm Characteristics (1)**

- **Current version:** July 6, 2009 (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 (June 2008)
  ● 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
<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>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strength</td>
<td>Ingredient</td>
<td>Dose form</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RxNorm Relations among drug entities
# Recap

<table>
<thead>
<tr>
<th>Name</th>
<th>Scope</th>
<th># concepts</th>
<th>Median</th>
<th>Subs. Hier</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNOMED CT</td>
<td>Clinical medicine (patient records)</td>
<td>310,314</td>
<td>2</td>
<td>yes</td>
<td>July 31, 2007</td>
</tr>
<tr>
<td>LOINC</td>
<td>Clinical observations and laboratory tests</td>
<td>46,406</td>
<td>3</td>
<td>no</td>
<td>Version 2.21 (no “natural language” names)</td>
</tr>
<tr>
<td>FMA</td>
<td>Human anatomical structures</td>
<td>~72,000</td>
<td>?</td>
<td>yes</td>
<td>(not yet in the UMLS)</td>
</tr>
<tr>
<td>RxNorm</td>
<td>Standard names for prescription drugs</td>
<td>93,426</td>
<td>1</td>
<td>no</td>
<td>Aug. 31, 2007</td>
</tr>
<tr>
<td>NCI Thesaurus</td>
<td>Cancer research, clinical care, public information</td>
<td>58,868</td>
<td>2</td>
<td>yes</td>
<td>2007_05E</td>
</tr>
<tr>
<td>ICD-10</td>
<td>Diseases and conditions (health statistics)</td>
<td>12,318</td>
<td>1</td>
<td>no</td>
<td>1998 (tabular)</td>
</tr>
<tr>
<td>MeSH</td>
<td>Biomedicine (descriptors for indexing the literature)</td>
<td>24,767</td>
<td>5</td>
<td>no</td>
<td>Aug. 27, 2007</td>
</tr>
<tr>
<td>UMLS .</td>
<td>Terminology integration in the life sciences</td>
<td>1,4 M</td>
<td>2</td>
<td>n/a</td>
<td>2007AC (English only)</td>
</tr>
</tbody>
</table>

[Bodenreider, YBMI 2008]
Biomedical ontologies

What they are for
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]
Biomedical ontologies

Needs for biosurveillance
Needs for biosurveillance

- Support for text mining
- Controlled vocabulary
- Aggregation
- Data integration
- Reasoning
Support for text mining

◆ Lexical resources
  ● Identify mentions in text
    ◦ Lexical variants

◆ Terminological resources
  ● Identify concepts
    ◦ Synonyms

◆ Ontological resources
  ● Identify relations, Semantic interpretation
    ◦ Domain knowledge
Controlled vocabulary

- Coded information
  - Storage
  - Processing
- Standardize
  - Definitions
  - Usage
Data aggregation

◆ Granularity mismatch
  ● Data recorded
  ● Data needed for making decisions

◆ Aggregation along hierarchies
  ● Subsumption hierarchies (isa)
  ● Ad hoc linearizations (e.g., ICD for mortality / morbidity)
Data integration

- Datasets annotated in reference to multiple ontologies
- Establish correspondence between equivalent concepts across ontologies (and datasets)

Role of terminology integration systems

- UMLS, RxNorm
- NCBO ontology services
Reasoning

◆ Description Logics
◆ Reasoning services (DL classifiers)
  ● Instance classification
Ontology in action in a biosurveillance system

BioCaster

http://biocaster.org/
Global Health Monitor

H1N1 swine influenza on Twitter

[+] Latest Reports
- [Cholera] PNG struggles to contain cholera outbreak - Solomon Star
  Found on Google News (2010-03-11)
  ➤ Search for biomedical references on NCBI, HighWire, GoPubMed, Google Scholar
- [Cholera] PNG struggles to contain cholera outbreak - Radio New Zealand International
  Found on Google News (2010-03-11)
  ➤ Search for biomedical references on NCBI, HighWire, GoPubMed, Google Scholar
- [Cholera] PNG struggles to contain cholera outbreak - Solomon Star
  Found on Google News (2010-03-11)
  ➤ Search for biomedical references on NCBI, HighWire, GoPubMed, Google Scholar

KML data for Google Earth
Updated every 1 hour, 24 hours per day. Next update: 12 Mar 2010 13:34 Asia/Tokyo

http://biocaster.org/
BioCaster

“Ontology-based text mining system for detecting and tracking the distribution of infectious disease outbreaks from linguistic signals on the Web”

- In operation since 2006
- Scans 1700 RSS feeds
- 4 stages
  - topic classification
  - named entity recognition (NER)
  - disease/location detection
  - event recognition
BioCaster ontology

- Vocabulary for Named Entity Recognition
  - Bridges between layman’s terms and biomedical concepts
  - Multi-lingual (8 languages)

- Knowledge about
  - Infectious diseases (e.g., causal agent, manifestations, associated syndrome)

- Mappings to reference terminologies

<table>
<thead>
<tr>
<th><strong>Identifiers</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td>Cholera</td>
</tr>
<tr>
<td><strong>Code</strong></td>
<td>DISEASE_3</td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>Cholera is an acute infectious gastrointestinal disease, caused by consuming the water or food that is contaminated with the bacterium Vibrio cholerae.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Preferred term</strong></th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th><strong>Information about this concept</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Synonyms</strong></td>
<td></td>
</tr>
<tr>
<td>en: Red death (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>en: Cholera (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>en: Asiatic cholera (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>fr: Choléra (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>jp: コレラ (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>ko: 콜레라 (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>zh: 霍乱 (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>es: Cólera (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>th: ผิดวัณโรค (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>vi: Bệnh dịch tả (DISEASE)</td>
<td></td>
</tr>
<tr>
<td>vi: Bệnh tả (DISEASE)</td>
<td></td>
</tr>
</tbody>
</table>
# BioCaster ontology: Cholera

<table>
<thead>
<tr>
<th><strong>Causal agent</strong></th>
<th>Vibrio cholerae (BACTERIUM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Symptoms</strong></td>
<td>Vomit (GastrointestinalTractSymptom)</td>
</tr>
<tr>
<td></td>
<td>Nausea (GastrointestinalTractSymptom)</td>
</tr>
<tr>
<td></td>
<td>Diarrhea (GastrointestinalTractSymptom)</td>
</tr>
<tr>
<td></td>
<td>Circulatory collapse (HematologicalSymptom)</td>
</tr>
<tr>
<td></td>
<td>Acidosis (HematologicalSymptom)</td>
</tr>
<tr>
<td></td>
<td>Dehydration (MetabolismSymptom)</td>
</tr>
<tr>
<td></td>
<td>Apathy (PsychologicalSymptom)</td>
</tr>
<tr>
<td><strong>Similar to</strong></td>
<td>Gastrointestinal syndrome (SYNDROME)</td>
</tr>
<tr>
<td><strong>Associated syndrome</strong></td>
<td></td>
</tr>
</tbody>
</table>

### Super concepts

<table>
<thead>
<tr>
<th><strong>Subconcepts</strong></th>
</tr>
</thead>
</table>
## BioCaster ontology: Cholera

<table>
<thead>
<tr>
<th>Resource</th>
<th>Identifier/Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-10</td>
<td>Cholera: A00</td>
</tr>
<tr>
<td>ICD-9</td>
<td>Cholera: 001</td>
</tr>
<tr>
<td>LOINC</td>
<td>Cholera due to <em>Vibrio cholerae</em> (disorder): DE-11601</td>
</tr>
<tr>
<td>LOINC</td>
<td>Cholera (disorder): DE-11600</td>
</tr>
<tr>
<td>MedDRA</td>
<td>Cholera, unspecified [10008634]: 10008634</td>
</tr>
<tr>
<td>MedDRA</td>
<td><em>Cholera due to Vibrio cholerae el tor</em> [10008633]: 10008633</td>
</tr>
<tr>
<td>MedDRA</td>
<td>Cholera due to <em>Vibrio cholerae</em> [10008632]: 10008632</td>
</tr>
<tr>
<td>MedDRA</td>
<td>Cholera [10008631]: 10008631</td>
</tr>
<tr>
<td>MeSH</td>
<td>Cholera: C01.252.400.959.347</td>
</tr>
<tr>
<td>SNOMED CT</td>
<td>Cholera (disorder): 63650001</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Cholera</td>
</tr>
</tbody>
</table>
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


Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Quoc-Hung Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi

BioCaster: detecting public health rumors with a Web-based text mining system


http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2639299/