Public Health Informatics Fellowship Program
February 23, 2007

Public Health Ontology
vs. Ontology for Public Health

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

- Formal representation of a domain modeling the things in that domain and the relationships between those things

- A set of logical axioms designed to account for the intended meaning of a vocabulary

[Guarino, FOIS 1998]
Introduction

Ontology spectrum

Ontology Dimensions based on McGuinness and Finin

Simple Terminologies

Expressive Ontologies

Medication Lists
DDI Lists

BT/NT,
Parent/Child,
Informal Is-A

Catalog

Terms/glossary

MeSH,
Gene Ontology,
UMLS Meta

Thesauri

KEGG

DB Schema

RDF(S)

Ontylog

OWL

IEEE SUO

CYC

Disjointness,
Inverse

Formal is-a
Frames
(Properties)

Formal instances

Value
Restriction

SNOMED CT

EcoCyc

TAMBIS

BioPAX

General logical
constraints

NLM

Ontology Dimensions based on McGuinness and Finin
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
Source Vocabularies

- **139 source vocabularies**
  - 17 languages
- **Broad coverage of biomedicine**
  - 5.5M names
  - 1.4M concepts
  - 16M relations
- **Common presentation**
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- MeSH
- SNOMED
- OMIM
- NCBI Taxonomy
- GWDA
- Go
- Anatomy
- Genome annotations
- Other subdomains
- Model organisms
Overview

- Why biomedical ontologies?
- What is the difference between cardiology and public health?
- Approaches to defining public health ontology
  - Bottom-up approach
  - Top-down approach
- Value sets for public health
  - Sources
  - Selection criteria
Why biomedical ontologies?
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

London Bills of Mortality

A general Bill for this present year, ending the 19 of December 1664, according to the Report made to the King's Most Excellent Majesty, by the Company of Parish Clerks of London, etc.

The Diseases and Casualties this Year:

- Abortive and Stillborn: 6,778
- Executed: 3
- Plague: 684

- Aged: 1,643
- Found dead in ditches, fields, etc.: 653

- Apparated and Suddenly: 1,643
- Plague: 6

- Eater: 15
- Scurvy: 1

- Inflamed: 10
- Gout and Sciatica: 27

- Laced: 10
- Quinsy: 19

- Bleeding: 10
- Rickets: 1

- Bloody Flux, Screaming & Flux: 10
- Ringing in the Ears: 112

- Burnt and Scalded: 8
- Hung and Strangled themselves: 7

- Cuttage: 1
- Appendicitis & Muelhefalen: 14

- Cancer, Consumpt and Tubal: 1
- Leprosy: 1

- Canker, and Throat: 1
- Spotted Fever and Pustules: 13

- Cold and Cough: 86
- Styes: 17

- Collick and Wind: 134
- Ulcers: 36

- Consumpt, and Typhus: 14
- Sore Throat and Swelling: 3

- Convulsion and Obert: 25
- Ulcers and Swelling: 1

- Cholera: 1
- Teeth and Worms: 20

- Childbed and Swollen: 2
- Warming: 9

- Drowned: 1
- Overdose & Starved: 4

- Catarrh of Throat: 481
- Maligns: 1

- Bored and Starved: 481
- Of the Plague: 681

- Infections in the Bubba in the 130 Parishes and in the Pest-House: 730
- Infections of the Plague in the 130 Parishes and in the Pest-House: 681
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 (18M citations from 5000 journals)
  - Manually indexed
  - Medical Subject Headings (MeSH)

- Genome
  - Model organism databases (Fly, Mouse, Yeast, …)
  - Manually / semi-automatically curated
  - Gene Ontology
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
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)
Contemporary public health perspective

- The automated exchange of data between public health partners
- The use of electronic clinical data for event detection
- Specimen and lab result information management and exchange
- Analysis of public health data
- Public health information dissemination and alerting

http://www.cdc.gov/phin/vocabulary/PHIN_Vocabulary_Brochure_v4.1.ppt
What is the difference between cardiology and public health?
Anatomy-based medical disciplines

- A limited number of relatively obvious dimensions of description
- Cardiology
  - Anatomy: Cardiovascular system
  - Diseases: Cardiovascular diseases
  - Procedures: Procedure on the cardiovascular system
  - Drugs: Cardiovascular agents
  - ...

Anatomy-based medical disciplines

- Concepts in each dimension can be represented in hierarchies
- Associative relations between concepts across dimensions

Anatomy
- CV system
  - Heart
  - Blood vessels
    - Artery
    - Vein
    - Capillary

Diseases
- CV diseases
  - Heart disease
  - Disorder of artery
  - Disorder of vein

Procedures
- Procedure on the CV system
  - Procedure on heart
  - CV surgical procedure

Drugs
- CV agents
  - Cardiotonic agents
  - Antihypertensive agents
“Transversal” disciplines

- Oncology (pathologic process)
- Infectiology (agent)
- Radiology (technique)
- Pediatrics (age group)
- Emergency medicine
- Public health
Cardiology and public health in the Medical Subject Headings (MeSH)
Health Occupations [G02]
  Medicine [G02.403]
    Specialties, Medical [G02.403.776]
      Aerospace Medicine [G02.403.776.014]
      Allergy and Immunology [G02.403.776.030]
      Anesthesiology [G02.403.776.050]
      Dermatology [G02.403.776.185]
      Emergency Medicine [G02.403.776.200]
      Family Practice [G02.403.776.230]
      Forensic Medicine [G02.403.776.240] +
      Hospitalists [G02.403.776.324]
      Internal Medicine [G02.403.776.409]
        Cardiology [G02.403.776.409.163]
        Endocrinology [G02.403.776.409.323]
        Gastroenterology [G02.403.776.409.405]
        Hematology [G02.403.776.409.445]
        Medical Oncology [G02.403.776.409.515] +
        Nephrology [G02.403.776.409.580]
        Pulmonary Disease (Specialty) [G02.403.776.409.675]
        Rheumatology [G02.403.776.409.730]
    Neurology [G02.403.776.550]
    Pathology [G02.403.776.600] +
    Pediatrics [G02.403.776.610] +
    Physical Medicine [G02.403.776.620] +
    Preventive Medicine [G02.403.776.630] +
    Psychiatry [G02.403.776.640] +
      Public Health [G02.403.776.670]
    Radiology [G02.403.776.700] +
    Reproductive Medicine [G02.403.776.710]
    Venereology [G02.403.776.880]

Public health
in other biomedical terminologies
Public health in the UMLS

- C0034019

Sources
- AOD  Alcohol and Other Drug Thesaurus
- CSP  CRISP Thesaurus
- LCH  Library of Congress Subject Headings
- MSH  Medical Subject Headings
- NCI  National Cancer Institute Thesaurus
- PSY  Thesaurus of Psychological Index Terms
<table>
<thead>
<tr>
<th>Identifiers:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>Public_Health</td>
</tr>
<tr>
<td>code</td>
<td>C17039</td>
</tr>
</tbody>
</table>

**Information about this concept:**

- **ALT_DEFINITION**
  MSH2001|Branch of medicine concerned with the prevention and control of disease and disability, and the promotion of physical and mental health of the population on the international, national, state, or municipal level.

- **DEFINITION**
  NCI|The science and practice of protecting and improving the health of a community, as by preventive medicine, health education, control of communicable diseases, application of sanitary measures, and monitoring of environmental hazards.|Disability History Museum glossary; http://www.disabilitymuseum.org/glossary.php

- **Synonym with source data**
  Public Health|PT|NCI

- **Preferred_Name**
  Public Health

- **Semantic_Type**
  Biomedical Occupation or Discipline

- **Synonym**
  Public Health

- **Unified Medical Language System Concept Identifier**
  C0034019

**Superconcepts**

- Medical Specialty

**Subconcepts**

- Arctic Research
- Community Health
- Occupational Health
- Patterns of Care
- Women's Health

Associative relations  Symbolic

- Concepts & Ideas
  - public service announcement
- Disorders
  - Zoonoses
- Living Beings
  - Public Health Podiatrist
- Occupations
  - preventive medicine specialty
  - Social Medicine
- Organizations
  - United States Public Health Service
- Procedures
  - Communicable Disease Control
  - Education, Public Health Professional
  - Health Promotion
  - Public Health Administration
  - Public health service
  - Screening procedure
  - Study of epidemiology

A bibliometric analysis in the fields of preventive medicine, occupational and environmental medicine, epidemiology, and public health

Soteriades ES, Falagas ME.

[...] CONCLUSION: USA researchers maintain a leadership position in the production of scientific articles in the fields of Preventive Medicine, Occupational/Environmental Medicine and Epidemiology, at a level similar to other scientific disciplines, while USA contribution to science in the field of Public Health is by all means outstanding. Less developed regions would need to support their researchers in the above fields in order to improve scientific production and advancement of knowledge in their countries.
Associative relations

- Bibliometrics*
- Environmental Medicine /statistics & numerical data
- Epidemiology /statistics & numerical data*
- Geography
- Humans
- Occupational Medicine /statistics & numerical data
- Periodicals /statistics & numerical data*
- Preventive Medicine /statistics & numerical data*
- Public Health /statistics & numerical data*
- Research /statistics & numerical data
- World Health
Public health Co-occurrences

- Health Policy 475
- Health Promotion 321
- World Health 228
- Environmental Health 187
- Environmental Exposure 186
- Delivery of Health Care 179
- Disease Outbreaks 168
- Epidemiology 166
- Smoking 164
- Health Status 163
- Public Policy 160
- Bioterrorism 158
- Public Health Administration 144
- HIV Infections 138
- International Cooperation 136

- Health Care Reform 133
- Developing Countries 133
- Health Care Reform 133
- Communicable Disease Control 131
- Politics 125
- Research 122
- Health education 117
- Violence 110
- Disaster Planning 104
- National Health Programs 102
- [...] 102

2,453 co-occurring concepts (2006AA)
[538 concepts cover 79%]
Public health vs. Cardiology  Summary

◆ Transversal discipline
  ● Many dimensions of description
    ■ MeSH
    ■ Co-occurring concepts in MEDLINE

◆ Represented in several terminologies
  ● Mostly “information sciences” terminologies
  ● Often a leaf node
Approaches to defining public health ontology
Bottom-up approach
Controlled Health Thesaurus (CHT)

- Under the umbrella of PHIN
- Developed for indexing / retrieval purposes (CDC website)
- Borrows from the UMLS Metathesaurus
- 40,000+ terms
- Polyhierarchic structure
CHT Bottom-up approach

- Original data: CDC public website
- Term extraction: Metaphrase (Apelon)
  - Mapping to the UMLS Metathesaurus
- Manual curation (Kevric)
  - Frequent terms added when absent from the UMLS Metathesaurus
With Ed Bunker (summer 2005)

42,639 terms

- 12,000 having no associated UMLS CUI
  - 7,257 with semantic type
    - Geographic area or Partner organization (not analyzed)
  - 4,743 terms analyzed
    - 843 terms mapped to UMLS
      - 408 semantically valid mappings (resynchronization)
    - 3,900 terms with no mapping to UMLS

Examples of terms not in UMLS

- All Other Legal Services
- Beet Sugar Manufacturing
- Environmental respiratory disease
- Greeting Card Publishers
- Latin American
- Mycobacteriology
- Office of the Director
- Quarantine facilities
- Septic Tank Servicers and Sewer Pipe Cleaners
- Temporary Shelters
- Waist circumference measurement
- Youth smoking rate
Top-down approach
# Domains of interest for public health

- **Disease / Clinical History / Observation (Finding)**
- **Anatomy (Body Sites)**
- **Procedures (Non-Laboratory)**
- **Laboratory Tests / Results**
- **Medications (Drugs)**
- **Immunization (Vaccine)**
- **Clinical Encounters (ADT)**
- **Demographics**
- **Modifier / Qualifier (Coded Answers / Results)**
- **Organism (including Animals)**
- **Public or Population Health**
- **Substance / Devices / Object**
- **Units of Measure**
<table>
<thead>
<tr>
<th><strong>Property Name</strong></th>
<th><strong>Property Value</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Child Domains</strong></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
</tr>
<tr>
<td>Clinical Encounters (ADT)</td>
<td></td>
</tr>
<tr>
<td>Laboratory Tests / Results</td>
<td></td>
</tr>
<tr>
<td>Organism (Including Animals)</td>
<td></td>
</tr>
<tr>
<td>Units of Measure</td>
<td></td>
</tr>
<tr>
<td>Anatomy (Body Sites)</td>
<td></td>
</tr>
<tr>
<td>Modifier / Qualifier (Coded Answers / Results)</td>
<td></td>
</tr>
<tr>
<td>Disease / Clinical History / Observation (Finding)</td>
<td></td>
</tr>
<tr>
<td>Procedures (Non-Laboratory)</td>
<td></td>
</tr>
<tr>
<td>Medications (Drugs)</td>
<td></td>
</tr>
<tr>
<td>Substance / Devices / Object</td>
<td></td>
</tr>
<tr>
<td>Public or Population Health</td>
<td></td>
</tr>
<tr>
<td>Immunization (Vaccine)</td>
<td></td>
</tr>
<tr>
<td><strong>Date Created</strong></td>
<td>8/9/2006</td>
</tr>
<tr>
<td><strong>Date Revised</strong></td>
<td>8/9/2006</td>
</tr>
<tr>
<td><strong>Definition</strong></td>
<td>PHIN Vocabulary Domains / Categories are based on PHIN Vocabulary domains and are used to group the value sets. Usage of PHIN Vocabulary Domains / Categories can be used in PHIN VADS Advanced Search efficiently to retrieve the value sets in groups. (2) PHIN VADS would allow users to browse the PHIN Vocabulary Domains and its metadata. PHIN Vocabulary Domain metadata would include the following: (a) Description of the PHIN Vocabulary Domain, (b) Other related PHIN Vocabulary Domain, (c) Associated Value Sets. Note: For detailed hierarchy of PHIN Vocabulary Domains and its associated code system, please download the PHIN Vocabulary Domain Finder from the PHIN Vocabulary Services website at <a href="http://www.cdc.gov/phin/vocabulary/index.html">www.cdc.gov/phin/vocabulary/index.html</a></td>
</tr>
<tr>
<td><strong>Parent Domain</strong></td>
<td>PHIN Vocabulary Domains</td>
</tr>
<tr>
<td><strong>Status</strong></td>
<td>Active</td>
</tr>
</tbody>
</table>
Populating the domains
Ontology approach

- **Identify classificatory principles**
  - Typically “jointly exhaustive and pairwise disjoint”
  - Organize concepts along these dimensions
  - E.g., anatomy (FMA tree)
    - Spatial dimension
    - Mass
    - Inherent 3D shape
    - ...
  - E.g., cancer (polyhierarchy)
    - By topography
    - By morphology
Explicit classificatory principle

Anatomical entity

Physical anatomical entity

Non-physical anatomical entity

Material physical anatomical entity

Non-material physical anatomical entity

Anatomical structure

Body substance

Anat. space

Anat. surface

Anat. line

Anat. point

Spatial dimension

Mass

Inherent 3D shape

3D 2D 1D 0D
Populating the domains
Information model approach

◆ **Identify value sets**
  
  ● For each (sub)domain
    - List of *diseases*
    - *Anatomy* vocabulary
    - Values for *EKG observations*
  
  ● For each data element
    - Allowable values for *gender*
    - List of *occupations*
    - List of *geographical locations*
Value sets for public health
Value sets Sources

- **Biomedical vocabularies**
  (terminologies, ontologies)
  - UMLS
  - Others (ICD-O3, CTCAE, ChEBI, RadLex,...)

- **Metadata registries**
  - Cancer Data Standards Repository (caDSR)

- **Already identified value sets**
  - PHIN
  - HL7
Value sets Selection criteria

◆ “Technical”
  ● Coverage
  ● Formalism for representation (standards)
  ● ...

◆ “Political”
  ● Intellectual property restrictions
  ● Mandated use
    ■ Consolidated Health Informatics (CHI)
  ● Governance
    ■ SDO for SNOMED CT

http://www.hhs.gov/healthit/chi.html
Related efforts  Health

- **HL7**

- **National Center for Biomedical Ontology**
  - [http://bioontology.org/](http://bioontology.org/)

- **caBIG**

- **Clinical Data Interchange Standards Consortium**
  - [http://www.cdisc.org/](http://www.cdisc.org/)
Related efforts  Others

- Federal Enterprise Architecture Reference Model Ontology (FEA-RMO)
  - [http://www.web-services.gov/fea-rmo.html](http://www.web-services.gov/fea-rmo.html)
- W3C Health Care and Life Sciences Interest Group
  - [http://www.w3.org/2001/sw/hcls/](http://www.w3.org/2001/sw/hcls/)
Conclusions

- Public health is a “transversal” discipline
  - No specific ontology / terminology
  - Borrows from multiple domains
    - Both within and outside biomedicine

- Interoperability of health-related systems
  - Information models
  - Standard vocabularies / value sets
  - Technical and political issues
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
References  UMLS

◆ UMLS
  umlsinfo.nlm.nih.gov

◆ UMLS browsers
  (free, but UMLS license required)
  ● Knowledge Source Server: umlsks.nlm.nih.gov
  ● Semantic Navigator:
  ● RRF browser
    (standalone application distributed with the UMLS)
References

Recent overviews
References