Bio-ontologies

The cream in the Semantic Web layer cake

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA
Semantic Web pastry
Semantic Web layer cake

Semantic Web layer cake

http://www.cookingwithkristina.com/uploaded_images/fudgy-702607.jpg
Semantic Web layer cake
Semantic Web layer cake
Outline

- Historical perspective
- Modern bio-ontologies
- Tools and formalisms
- Institutionalization of bio-ontologies
- Bio-ontologies and Semantic Web
Bio-ontologies: current trends and future directions

Olivier Bodenreider and Robert Stevens

Submitted: 23rd June 2006; Received (in revised form): 10th July 2006

http://bib.oxfordjournals.org/cgi/reprint/7/3/256?ijkey=1ejwW7ipyG1ASil&keytype=ref
Before we called them bio-ontologies

*brief history of 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

London Bills of Mortality

A generall Bill for this present year, ending the 19th of December 1665, according to the Report made to the KINGS most Excellent Majestye.

By the Company of Parish-Clerks of London, 1665.

A
Bills of Mortality

For this Present Year:

Beginning the 27th of December 1664, and ending the 19th of December following:

As also, The GENERAL or whole years BILL:

According to the Report made to the King’s Most Excellent Majesty,

by the Company of Parish-Clerks of London, etc.
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 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: Nf2  neurofibromatosis 2  [Mus musculus]
GeneID: 18016  Locus tag: MGI:97307

GeneOntology

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)
  - ChEBI
  - SNOMED CT
  - LOINC
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)
Modern bio-ontologies
Biomedical ontologies (and terminologies)

- **The OBO family**
  - Ontologies and terminologies
  - Gene Ontology
  - Mostly biological ontologies

- **UMLS**
  - Ontologies and terminologies
  - MeSH, SNOMED CT
  - Mostly clinical ontologies
Open Biological Ontologies

◆ Extended family of the Gene Ontology (GO)

◆ Collaborative development
  - http://obo.sourceforge.net/

◆ National Center for Biomedical Ontology
  - http://bioontology.org/

◆ OBO Foundry
  - http://obofoundry.org/
  - Promote best practices in ontology development
  - 10 inclusion criteria
Open Biological Ontologies (OBO)

OBO Ontology Browser

Browse the tree by clicking on the category names; click on an ontology name to view more information on it.

- anatomy
- animal natural history and life history
- chemical
- development
- ethology
- evidence codes
- experimental conditions
- genomic and proteomic
- metabolomics
- OBO relationship types
- phenotype
- taxonomic classification
- vocabularies

http://obo.sourceforge.net/
Integrating subdomains

<table>
<thead>
<tr>
<th>RELATION TO TIME</th>
<th>CONTINUANT</th>
<th>OCCURRENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRANULARITY</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ORGAN AND ORGANISM</th>
<th>INDEPENDENT</th>
<th>DEPENDENT</th>
<th>OCCURRENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism (NCBI Taxonomy?)</td>
<td>Anatomical Entity (FMA, CARO)</td>
<td>Organ Function (FMP, CPRO)</td>
<td>Phenotypic Quality (PaTO)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CELL AND CELLULAR COMPONENT</th>
<th>INDEPENDENT</th>
<th>DEPENDENT</th>
<th>OCCURRENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell (CL)</td>
<td>Cellular Component (FMA, GO)</td>
<td>Cellular Function (GO)</td>
<td>Biological Process (GO)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MOLECULE</th>
<th>INDEPENDENT</th>
<th>DEPENDENT</th>
<th>OCCURRENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecule (ChEBI, SO, RnaO, PrO)</td>
<td>Molecular Function (GO)</td>
<td>Molecular Process (GO)</td>
<td></td>
</tr>
</tbody>
</table>

(Barry Smith)
OBO ontologies Examples

- Gene Ontology
- Cell types
- Sequence Ontology
- ChEBI
- Foundational Model of Anatomy
- PATO – phenotypic qualities
- Relationship types
- Ontology for Biomedical Investigations
UMLS Source Vocabularies (2007AA)

◆ 139 source vocabularies
  ● 17 languages

◆ Broad coverage of biomedicine
  ● 5.5M names
  ● 1.4M concepts
  ● 16M relations

◆ Common presentation
Biomedical terminologies in UMLS

◆ General vocabularies
  • anatomy (UWDA, Neuronames)
  • drugs (RxNorm, First DataBank, Micromedex, …)
  • medical devices (UMD, SPN)

◆ Several perspectives
  • clinical terms (SNOMED CT)
  • information sciences (MeSH, CRISP)
  • administrative terminologies (ICD-9-CM, CPT-4)
  • data exchange terminologies (HL7, LOINC)
Biomedical terminologies in UMLS

- **Specialized vocabularies**
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - oncology (NCI Thesaurus, PDQ)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART, MedDRA)
  - primary care (ICPC)
  - genomics (Gene Ontology, HUGO, OMIM)

- **Terminology of knowledge bases** (AI/Rheum, DXplain, QMR)
Integrating subdomains

Clinical repositories

Genetic knowledge bases

UMLS

Biomedical literature

Model organisms

Genome annotations

NCBI Taxonomy

UWDA

Anatomy

SNOMED

OMIM

MeSH

Other subdomains

...
Tools and formalisms for bio-ontologies

Three examples
Three examples

- Foundational Model of Anatomy
  - Protégé-frames
- Gene Ontology
  - OBO-Edit
- NCI Thesaurus
  - OWL DL
- Conversions
Foundational Model of Anatomy (FMA)

- University of Washington
- Canonical anatomy
- 75,000 anatomical entities
- Synonyms
- Relationships
  - Isa
  - Part of (5 subtypes)
  - Topological, etc.
- Frame-based / Protégé


http://protege.stanford.edu/
Explicit classificatory principle

**Foundational Model of Anatomy**

- **Anatomical entity**
  - Physical anatomical entity
  - Non-physical anatomical entity

- **Material physical anatomical entity**
  - Anatomical structure
  - Body substance

- **Non-material physical anatomical entity**
  - Anat. space
  - Anat. surface
  - Anat. line
  - Anat. point

- **Spatial dimension**
  - 3D
  - 2D
  - 1D
  - 0D

- **Mass**
  - +
  - -

- **Inherent 3D shape**
  - +
  - -
**FMA Conversions**

- **OWL DL**
  - Golbreich et al., JWS 2006
- **OWL Full**
  - Noy and Rubin, SMI Tech Report 2007
- **OBO**
  - [http://obofoundry.org/cgi-bin/detail.cgi?id=fma_lite](http://obofoundry.org/cgi-bin/detail.cgi?id=fma_lite)
Gene Ontology

- GO Consortium
- Annotation of gene products (Molecular functions, Cellular components, Biological processes)
- 24,000 terms
- Synonyms
- Isa and part of relations
- OBO-Edit / OBO
- Also available in RDF and OWL DL

http://www.geneontology.org/

http://oboedit.org/
OBO format

- Used to represent many ontologies in the OBO family (Open Biological Ontologies)

http://www.godatabase.org/dev/doc/obo_format_spec.html

- Essentially a subset of OWL DL

```
[Term]
id: GO:0019563
name: glycerol catabolism
namespace: biological_process
def: "The chemical reactions and pathways resulting in the breakdown of glycerol ..."
subset: gosubset_prok
exact_synonym: "glycerol breakdown" []
exact_synonym: "glycerol degradation" []
xref_analog: MetaCyc:PWY0-381
is_a: GO:0006071 ! glycerol metabolism
is_a: GO:0046174 ! polyol catabolism
```
NCI Thesaurus

- National Cancer Institute
- Cancer research
- 54,000 concepts
- 150,000 concept names
- Relations
  - Isa
  - Associative (87 relationship types)
- OWL DL

http://nciterms.nci.nih.gov/NCIBrowser/

http://protege.stanford.edu/
Institutionalization of bio-ontologies
Bio-ontologies have become mainstream

Number of articles on "ontology/ies" in PubMed/MEDLINE

- GO
- others

Year: 1996 to 2006
Some institutions Bio-ontologies

◆ National Center for Biomedical Ontology
  ● http://bioontology.org/

◆ OBO Foundry
  ● http://obofoundry.org/
  ● Promote best practices in ontology development

◆ Other ontology centers
  ● NCOR – National Center for Ontology Research (US)
  ● ECOR – European Center for Ontology Research
Some institutions  Semantic Web

- W3C Health Care and Life Sciences Interest Group
  - [http://www.w3.org/2001/sw/hcls/](http://www.w3.org/2001/sw/hcls/)
  - BioRDF
  - BioOnt
Bio-ontologies
and Semantic Web
Use cases for a biomedical SW

- **Integration**
  - Data/Information
  - E.g., translational research

- **Hypothesis generation**

- **Knowledge discovery**

- **Clinical data**
  - Aggregation, sharing, exchange
  - Support for clinical decision
Some issues

◆ Format
  ● RDF/S, OWL, SKOS vs. OBO, RRF, etc.
  ● Converters

◆ Permanent identification of biomedical entities
  ● Syntax: URI vs. LSID
  ● Semantic: Trans-namespace identification

◆ Availability, openness

◆ Governance, trust
Summary

- Biomedical information integration is a good use case for the Semantic Web
  - Semantic Web technologies
  - Ontologies

- Ontologies
  - Identification
  - Mapping
  - Reasoning
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