Biomedical Ontologies

What are they good for?

Olivier Bodenreider

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

- Biomedical ontologies
- Applications
- Some issues for discussion
- Current trends and future directions
Biomedical ontologies
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]
The Knowledge Semantics Continuum

**Ontology Dimensions based on McGuinness and Finin**

- **Simple**
  - Terminologies
  - Catalog
  - Terms/glossary

- **Expressive**
  - Ontologies
  - DB Schema
  - RDF(S)
  - Formal is-a Frames (Properties)
  - Formal instances
  - Value Restriction

**Medication Lists**
- DDI Lists

**Thesauri:**
- BT/NT
- Parent/Child
- Informal Is-A

**Frames:**
- Properties

**Value Restriction:**
- Disjointness, Inverse

**Domain-Specific Databases:**
- MeSH
- Gene Ontology
- UMLS Meta
- CYC
- KEA/CRD(S)
- DB Schema
- Ontylog
- TAMBIS
- BioPAX
- BioPAX
- EcoCyc

**Repositories/Interchange Formats:**
- Medication Lists
- DDI Lists

**The Knowledge Semantics Continuum**

Ontology Dimensions based on McGuinness and Finin
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
  - 5 inclusion criteria

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

- OBO Foundry
  - [http://obofoundry.org/](http://obofoundry.org/)
  - Promote best practices in ontology development
Relation ontology

- Defines 14 core relations (isa, part of, derives from, has agent, …)
  - Textual definition
  - Formal definition
  - Comments
- To be used by other OBO ontologies

[Smith et al., Genome Biology, 2005, 6:R46]
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/
OBO format

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

http://www godevelopment.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
```
What does UMLS stand for?

- Unified
- Medical
- Language
- System

UMLS®
Unified Medical Language System®
UMLS Metathesaurus®
Motivation

- Started in 1986
- National Library of Medicine
- “Long-term R&D project”

“[…] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
- The second is the distribution of useful information among many disparate databases and systems.”
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

- 138 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 5.4M names
  - 1.4M concepts
  - 16M relations
- Common presentation

(2006AD)
Integrating subdomains

Clinical repositories

Genetic knowledge bases

Other subdomains

SNOMED

OMIM

Biomedical literature

Model organisms

NCBI Taxonomy

Genome annotations

UMLs

MeSH

GO

UWDA

Anatomy

Clinical repositories

Genetic knowledge bases

Other subdomains

SNOMED

OMIM

Biomedical literature

Model organisms

NCBI Taxonomy

Genome annotations

UMLs

MeSH

GO

UWDA

Anatomy

...
Addison’s Disease: Concept

A disease characterized by hypotension, weight loss, anorexia, weakness, and sometimes a bronze-like melanotic hyperpigmentation of the skin. It is due to tuberculosis- or autoimmune-induced disease (hypofunction) of the adrenal glands that results in deficiency of aldosterone and cortisol. In the absence of replacement therapy, it is usually fatal.
Metathesaurus Concepts (2006AD)

- **Concept** (~ 1.4 M) CUI
  - Set of synonymous concept names
- **Term** (~ 4.8 M) LUI
  - Set of normalized names
- **String** (~ 5.4 M) SUI
  - Distinct concept name
- **Atom** (~ 6.5 M) AUI
  - Concept name in a given source

<table>
<thead>
<tr>
<th>CUI</th>
<th>Term</th>
<th>String</th>
<th>Atom</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0000001 headache (source 1)</td>
<td>A0000002 headache (source 2)</td>
<td>S00000001</td>
<td></td>
</tr>
<tr>
<td>A0000003 Headache (source 1)</td>
<td>A0000004 Headache (source 2)</td>
<td>S00000002</td>
<td></td>
</tr>
<tr>
<td>L00000001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A0000005 Cephalgia (source 1)</td>
<td>S00000003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L00000002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C00000001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Organize concepts

- Inter-concept relationships: hierarchies from the source vocabularies
- Redundancy: multiple paths
- One graph instead of multiple trees (multiple inheritance)
Semantic Network

Metathesaurus

Semantic Types

- Fully Formed Anatomical Structure
- Embryonic Structure
- Disease or Syndrome
- Pharmacologic Substance
- Population Group

Concepts

- Esophagus
- Left Phrenic Nerve
- Heart Valves
- Heart
- Fetal Heart

- Medias-tinum
- Saccular Viscus

Pharmacologic Substance

- Angina Pectoris
- Cardiotonic Agents
- Tissue Donors

Population Group

Angina Pectoris

Cardiotonic Agents

Tissue Donors
Relationships can inherit semantics

Semantic Network

Fully Formed Anatomical Structure

Body Part, Organ, or Organ Component

Disease or Syndrome

Adrenal Cortex

Adrenal Cortical hypofunction

Biologic Function

Pathologic Function

location of

isa

Metathesaurus

LHNCBC
Terminology vs. ontology

- **Terminological resources**
  - Collections of terms (e.g., controlled vocabularies)
  - Useful for indexing and annotation
  - MeSH, GO

- **Ontological resources**
  - Collections of kinds of entities (substances, qualities, processes)
  - relations among them
  - Useful for reasoning
  - UMLS Semantic Network, SNOMED CT

Ontological gradient
Applications
Applications of ontologies

- Information integration
- Support for natural language processing
- Reasoning, automated classification
- Support for knowledge discovery
- Support for clinical decision
- ...
Information integration
**NF2 Gene, protein, and disease**

*Neurofibromatosis 2* is an autosomal dominant disease characterized by tumors called schwannomas involving the acoustic nerve, as well as other features. The disorder is caused by mutations of the *NF2 gene* resulting in absence or inactivation of the protein product. The protein product of NF2 is commonly called *merlin* (but also neurofibromin 2 and schwannomin) and functions as a tumor suppressor.
Schwannoma (acoustic neuroma)

http://www.mayoclinic.com
NF2 gene

http://staff.washington.edu/timk/cyto/human/

Merlin

- **Synonyms**
  - Neurofibromin 2
  - Schwannomin
  - Schwannomerlin
  - Neurofibromatosis-2

- **10 isoforms**

- **Annotations**
  - Negative regulation of cell proliferation
  - Cytoskeleton
  - Plasma membrane
Neurofibromatosis 2 (Type II neurofibromatosis, Bilateral acoustic neurofibromatosis)  
*C0027832*

NF2 (Neurofibromin 2 gene)  
*C0085114*

Merlin (Schwannomin, Neurofibromin 2)  
*C0254123*

Drosophila melanogaster merlin (Dmerlin) mRNA, complete cds.  
*U49724*

External resources

OMIM

Genbank
Limitations

- Genes not systematically represented
  - Most gene products and diseases are
- Gene/Gene product-Disease relations
  - Not systematically represented
  - Not explicitly represented (e.g., co-occurrence)
- Cross-references not systematically represented
- Naming conventions (genes)
Support for natural language processing
Semantic interpretation

- Source: text corpus / terminology
- Correspondence between
  - Linguistic phenomena
  - Semantic relations
- Semantic constraints provided by ontologies

[Navigli & al., TKE, 2002]
[Romacker, AIME, 2001]
[Rindflesch & al., JBI, 2003]
Semantic interpretation

Hemofiltration in digoxin overdose

Hemofiltration treats Digoxin overdose

Therapeutic or Preventive Procedure treats Disease or Syndrome

Antibiotic treats Disease or Syndrome
Medical Device treats Injury or Poisoning
Pharmacologic Substance treats Congenital Abnormality
Ther. or Prev. Procedure treats Disease or Syndrome
Language & Computing

- Maria van Gurp, Manuel Decoene, Marnix Holvoet, Mariana Casella dos Santos
- LinKBase®, a Philosophically-Inspired Ontology for NLP/NLU Applications
- Proceedings of KR-MED 2006, pp. 67-75
DL reasoning, automated classification
DL reasoning, automated classification

- Ontologies represent knowledge
- Automated reasoners infer conclusions from the given knowledge
  - Make implicit knowledge explicit
  - Help validate the ontology (e.g., consistency checking and automatic classification in DL)
- Need for more expressive logic
  - Inference rules
OWL reasoners

- For OWL DL, not OWL Full
- Reasoners
  - Fact++
  - Pellet
  - RacerPro
- Functions
  - Consistency checking
  - Automatic classification
Simple inference

Disease_Excludes_Primary_Anatomic_Site

Intracranial neoplasm

Intracranial meningioma

Benign intracranial meningioma

Spinal chord

Spinal chord

Spinal chord
Complex inference

- **Clinical decision support**
  - If patient is treated by aminoglycosides and patient has impaired renal function then reduce dose (or frequency of administration) of aminoglycosides

- **Not directly supported by DL reasoners**

- **Require rule languages**
  - RuleML
  - SWRL (Semantic Web Rule Language)

http://www.ruleml.org/

http://www.w3.org/Submission/SWRL/
Support for knowledge discovery
Motivation

◆ Biomedical information is growing at an increasingly faster pace
  ● High-throughput approach to knowledge processing
◆ Information retrieval is the starting point, not the end of the journey for the researcher
  ● Towards “computable” knowledge
◆ Integration between literature and other resources is insufficient
  ● Adequate for navigation purposes
  ● Insufficient for knowledge processing
Advanced Library Services

Summary

Source selection (PubMed, annotations)

Biomedical Knowledge Repository

- Biomedical Literature
- Terminological Knowledge
- Structured Knowl. Bases
- Contributed Knowledge

Document Summarization
- Question Answering
- Knowledge Discovery
- Information Retrieval

MEDLINE
CT.gov
UMLS
Entrez Gene
GO
Some issues for discussion
Issues

◆ What is the right amount of semantics?
  ● “A little semantics goes a long way” – Does it?

◆ What approach to building ontologies?
  ● Top-down vs. bottom-up

◆ What formalism?
  ● XML, RDF/S, OWL Lite/DL/Full, OBO?

◆ Semantic Web for Health Care and Life Sciences
Current trends and future directions
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
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
UMLS References

- UMLS
  umlsinfo.nlm.nih.gov

- UMLS browsers
  (free, but UMLS license required)
  - Semantic Navigator:
  - RRF browser
    (standalone application distributed with the UMLS)
UMLS References

◆ Gentle introduction

◆ Seminal paper
Semantic Web References

- World Wide Web Consortium (W3C)
  - http://www.w3.org/
- W3C Health Care and Life Sciences Interest Group
  - http://www.w3.org/2001/sw/hcls/