Terminological systems in biomedicine

*From terminology integration to information integration*

Olivier Bodenreider

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

- Information integration in biomedicine
  - Use case: Oncology
  - Some issues: naming, normalization, mapping
- Terminology integration in biomedicine
  - Bottom-up
    - Unified Medical Language System
  - Top-down
    - OBO Foundry ontologies
- Applications
  - Biomedical Semantic Web
Information integration in oncology
Multiple terminologies for oncology

- International Classification of Diseases-Oncology (ICD-O-3)
  - Cancer registries
  - Epidemiology, Public health
- SNOMED CT
  - Patient records
  - Clinical care
- NCI Thesaurus
  - Annotation of research data
SNOMED CT

http://www.clinical-info.co.uk/
Prostate Adenocarcinoma

Identifiers:
- name: Prostate_Adenocarcinoma
- code: C2919

Relationships to other concepts:
- Disease_Has_Finding: Invasive Lesion
- Disease_Has_Abnormal_Cell: Adenocarcinoma Cell
- Disease_Has_Normal_Tissue_Origin: Prostatic Epithelium
- Disease_May_Have_Finding: Serum Prostate Specific Antigen Increased
- Disease_Has_Finding: Carcinomatous Component Present
- Disease_Excludes_Abnormal_Cell: Neoplastic Smooth Muscle Cell
- Disease_Excludes_Abnormal_Cell: Malignant Squamous Cell
- Disease_Has_Parent_Anomal_Cell: Prostate Gland
- Disease_Has_Associated_Anomal_Cell: Male Reproductive System
- Disease_Excludes_Abnormal_Cell: Malignant Stromal Cell
- Disease_Has_Associated_Anomal_Cell: Prostate Gland
- Disease_Has_Normal_Cell_Origin: Epithelial Cell

Supercategories:
- Adenocarcinoma
- Common Carcinoma
- Invasive Prostate Carcinoma

Subcategories:
- Adenocarcinoma of Prostate
- Metastatic Prostatic Adenocarcinoma
- Moderately Differentiated Prostate Adenocarcinoma
- Poorly Differentiated Prostate Adenocarcinoma
- Prostate Adenocarcinoma with Focal Neuroendocrine Differentiation
- Prostate Ductal Adenocarcinoma
- Stage III Prostate Adenocarcinoma
- Stage II Prostate Adenocarcinoma
- Stage I Prostate Adenocarcinoma
- Well Differentiated Prostate Adenocarcinoma
ICD-O-3

◆ Morphology
  - [...]  
  - 814-838  Adenomas and adenocarcinomas  
    - 8140/3  Adenocarcinoma, NOS

◆ Anatomy
  - [...]  
  - C60-C63  Male genital organs  
    - C61  Prostate gland  
      - C61.9  Prostate, NOS  
      Prostate gland

Adenocarcinoma of prostate
Integrating terminologies

Cancer registries
Research data
Clinical repositories

Adenocarcinoma of prostate (399490008)
NCI Thesaurus
SNOMED CT
Prostate adenocarcinoma (C2919)

Adenocarcinoma of prostate (8140/3 + C61.9)
ICD-O
Cancer registries
Integrating subdomains

Clinical repositories

SNOMED CT

Research data

Cancer registries

NCI Thesaurus

ICD-O
Information integration in biomedicine

Some issues: naming, normalization, mapping
Many biomedical entities have several names (synonymy)

- Drug names
- Gene names
- Disease names
- ...

A given name may refer to several different entities (polysemy)

- Nail (body part)
- Nail (medical device)
## Brand names for paracetamol (acetaminophen)

- [Brand names](http://en.wikipedia.org/wiki/List_of_paracetamol_brand_names)

<table>
<thead>
<tr>
<th>Brand name</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acamol</td>
<td>Israel</td>
</tr>
<tr>
<td>Atamel</td>
<td>Venezuela</td>
</tr>
<tr>
<td>Adol</td>
<td>Oman</td>
</tr>
<tr>
<td>Aldolor</td>
<td>Israel</td>
</tr>
<tr>
<td>Alvedon</td>
<td>Sweden</td>
</tr>
<tr>
<td>APAP</td>
<td>Poland</td>
</tr>
<tr>
<td>Benuron</td>
<td>Portugal, Germany</td>
</tr>
<tr>
<td>Biogesic</td>
<td>Philippines</td>
</tr>
<tr>
<td>Buscapina</td>
<td>Argentina</td>
</tr>
<tr>
<td>Cemol</td>
<td>Thailand</td>
</tr>
<tr>
<td>Crocin</td>
<td>India</td>
</tr>
<tr>
<td>Dafalgan</td>
<td>Belgium, France, Portugal, Russia, Ukraine</td>
</tr>
<tr>
<td>Doleron</td>
<td>Slovenia</td>
</tr>
<tr>
<td>Depon</td>
<td>Greece</td>
</tr>
<tr>
<td>Doxamol</td>
<td>Israel</td>
</tr>
<tr>
<td>Dolex</td>
<td>Colombia</td>
</tr>
<tr>
<td>Dolprane</td>
<td>France, Portugal, Russia, Ukraine</td>
</tr>
<tr>
<td>Efferalgan</td>
<td>France, Italy, Portugal, Russia, Spain, Ukraine</td>
</tr>
<tr>
<td>FeverAll</td>
<td>United States</td>
</tr>
<tr>
<td>Gelocatil</td>
<td>Spain</td>
</tr>
<tr>
<td>Gripin</td>
<td>Turkey</td>
</tr>
<tr>
<td>Lekadol</td>
<td>Croatia, Slovenia</td>
</tr>
<tr>
<td>Metacin</td>
<td>India</td>
</tr>
<tr>
<td>Pamol</td>
<td>Denmark, Finland, France</td>
</tr>
<tr>
<td>Panado</td>
<td>South Africa</td>
</tr>
<tr>
<td>Panadol</td>
<td>Australia, Azerbaijan, Central America, Egypt, Finland, Greece, Hong Kong, Hungary, Indonesia, Ireland, Kenya, Lebanon, Macedonia, Malaysia, Malta, Netherlands, New Zealand, Nigeria, Pakistan, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Sri Lanka, Switzerland, Taiwan, Ukraine, Estonia, United Kingdom</td>
</tr>
<tr>
<td>Panamax</td>
<td>Australia, United Kingdom</td>
</tr>
<tr>
<td>Panodil</td>
<td>Denmark, Iceland, Sweden</td>
</tr>
<tr>
<td>Paracet</td>
<td>Norway</td>
</tr>
<tr>
<td>Paralen</td>
<td>Czech Republic, Slovakia</td>
</tr>
<tr>
<td>Paramed</td>
<td>Botswana, South Africa, Zimbabwe</td>
</tr>
<tr>
<td>Paramol</td>
<td>Israel, Taiwan</td>
</tr>
<tr>
<td>Perdolan</td>
<td>Belgium</td>
</tr>
<tr>
<td>Perfalgan</td>
<td>Germany</td>
</tr>
<tr>
<td>Pinex</td>
<td>Denmark, Iceland, Norway</td>
</tr>
<tr>
<td>Plicet</td>
<td>Croatia</td>
</tr>
<tr>
<td>Reliv</td>
<td>Sweden</td>
</tr>
<tr>
<td>Rokamol</td>
<td>Israel</td>
</tr>
<tr>
<td>Sara</td>
<td>Thailand</td>
</tr>
<tr>
<td>Tachiprina</td>
<td>Italy</td>
</tr>
<tr>
<td>Tylenol</td>
<td>Brazil, Canada, Japan, South Korea, Thailand, United States</td>
</tr>
<tr>
<td>Tempra</td>
<td>Philippines</td>
</tr>
</tbody>
</table>
Names for dystrophin

Dystrophin (muscular dystrophy, Duchenne and Becker types) [Homo sapiens]

Official Symbol: DMD
Other Aliases: GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272
Other Designations: Duchenne muscular dystrophy protein
Chromosome: X, Location: Xp21.2
Annotation: Chromosome X, NC_000023.9 (33267646..31047265, complement)
MIM: 300377
GeneID: 1756

Names for renal cell carcinoma

http://www.clinical-info.co.uk/
Entity recognition

◆ Identifying biomedical entities in text
  ● Names entity recognition
  ● Tagging “mentions”
  ● Semantic annotation

◆ Supported by terminology
  ● Collects the names used in the domain
  ● Often incompletely

◆ Example: BioCreative
  ● 1A – Gene name identification
  ● 2GM – Gene mention tagging
Normalization

- Biomedical entities are identified by unique identifiers in various terminology systems.
- Resolve names into identifiers (in a given namespace).
- Supported (in part) by terminology resources.
- Example: BioCreative
  - 1B and 2GN – Gene Normalization.
**Identifier for paracetamol (acetaminophen)**

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master Drug Data Base. Medi-Span</td>
<td>5005</td>
<td>Acetaminophen</td>
</tr>
<tr>
<td>FDA National Drug Code Directory</td>
<td>50612</td>
<td>PARACETAMOL</td>
</tr>
<tr>
<td>FDA Structured Product Labels</td>
<td>362O9ITL9D</td>
<td>ACETAMINOPHEN</td>
</tr>
<tr>
<td>First DataBank NDDF Plus</td>
<td>001605</td>
<td>Acetaminophen</td>
</tr>
<tr>
<td>SNOMED Clinical Terms</td>
<td>90332006</td>
<td>Acetaminophen (product)</td>
</tr>
<tr>
<td>SNOMED Clinical Terms</td>
<td>387517004</td>
<td>Acetaminophen (substance)</td>
</tr>
<tr>
<td>VA National Drug File</td>
<td>4017513</td>
<td>ACETAMINOPHEN</td>
</tr>
</tbody>
</table>

Source: RxNorm database (5/3/2007)
Identifier for dystrophin

Entrez Gene


**DMD**

Official Symbol DMD and Name: dystrophin (muscular dystrophy, Duchenne and Becker types) [*Homo sapiens*]

Other Aliases: GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272

Other Designations: Duchenne muscular dystrophy protein: dystrophin

Chromosome: X, Location: Xp21.2

Annotation: Chromosome X, NC_000023.9 (33267646..31047265, complement)

MIM: 300377

GeneID: 1756
Identifier for renal cell carcinoma

Details of 'clear cell carcinoma of kidney'

Concept Status: Current

Descriptions:
- clear cell carcinoma of kidney (disorder)
- clear cell carcinoma of kidney
- adenocarcinoma of kidney
- carcinoma of kidney
- Grawitz tumor
- renal cell adenocarcinoma
- renal cell carcinoma

Fully defined by...
- Is a:
  - malignant tumor of kidney parenchyma
  - primary malignant neoplasm of kidney
  - primary malignant neoplasm of retroperitoneum

Group:
- Associated morphology:
  - clear cell adenocarcinoma
- Finding site:
  - structure of parenchyma of kidney

Laterality:
- Left
- Right

Qualifiers:

Legacy codes:
- SNOMED: D7-F011C
- CTV3ID: X78Yx

http://www.clinical-info.co.uk/
Mapping / Integration

- Identify equivalent entities across systems (across namespaces)
  - Shared identifiers
  - Existing mappings (e.g., SNOMED CT to ICD-9-CM)
  - Ontology alignment techniques (lexical + structural)
- Align equivalent entities
  - Pairwise: mapping
  - More broadly: integration
- Forms the basis for information integration in the Semantic Web (mashups)
## Identifier for paracetamol (acetaminophen)

<table>
<thead>
<tr>
<th>Database/Source</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master Drug Data Base. Medi-Span</td>
<td>5005</td>
<td>Acetaminophen</td>
</tr>
<tr>
<td>FDA National Drug Code Directory</td>
<td>50612</td>
<td>PARACETAMOL</td>
</tr>
<tr>
<td>FDA Structured Product Labels</td>
<td>362091TL9D</td>
<td>ACETAMINOPHEN</td>
</tr>
<tr>
<td>First DataBank NDDF Plus</td>
<td>001605</td>
<td>Acetaminophen</td>
</tr>
<tr>
<td>SNOMED Clinical Terms</td>
<td>90332006</td>
<td>Acetaminophen (product)</td>
</tr>
<tr>
<td>SNOMED Clinical Terms</td>
<td>387517004</td>
<td>Acetaminophen (substance)</td>
</tr>
<tr>
<td>VA National Drug File</td>
<td>4017513</td>
<td>ACETAMINOPHEN</td>
</tr>
<tr>
<td>RxNorm</td>
<td>161</td>
<td>Acetaminophen</td>
</tr>
</tbody>
</table>
### Identifier for dystrophin

**Entrez Gene**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DMD</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Official Symbol:** DMD and Name: dystrophin (muscular dystrophy, Duchenne and Becker types) [Homo sapiens]

**OtherAliases:** GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272

**OtherDesignations:** Duchenne muscular dystrophy protein: dystrophin

**Chromosome:** X, **Location:** Xp21.2

**Annotation:** Chromosome X, NC_000023.9 (33267646..31047265, complement)

**MIM:** 300377

**GeneID:** 1756
Identifier for renal cell carcinoma

http://www.clinical-info.co.uk/
Bottom-up terminology integration

Unified Medical Language System
**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.»
Source Vocabularies

◆ 141 source vocabularies
  - 17 languages
◆ Broad coverage of biomedicine
  - 6.1M names
  - 1.5M concepts
  - 8M 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
- SNOMED CT
- OMIM
- SNOED CT
- MeSH
- Biomedical literature
- Other subdomains
- Model organisms
- NCBI Taxonomy
- FMA
- GO
- Anatomy
- Genome annotations
Trans-namespaces integration

Addison's disease (363732003)

Other subdomains

NCBI Taxonomy

Model organisms

FMA

Anatomy

Genome annotations

Biomedical literature

Genetic knowledge bases

OMIM

Other subdomains

Addison Disease (D000224)

SNOEMED CT
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Anatomy
- Model organisms
- Other subdomains

Biomedical literature
Top-down terminology integration

*OBO Foundry 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)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Prefix</th>
<th>File</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological imaging methods</td>
<td>FBBi</td>
<td>image.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Biological process</td>
<td>GO</td>
<td>gene_ontology.obo</td>
<td>obo</td>
</tr>
<tr>
<td>BRENDA tissue / enzyme source</td>
<td>BTO</td>
<td>BrendaTissue.obo</td>
<td>obo</td>
</tr>
<tr>
<td>C. elegans development</td>
<td>WBIs</td>
<td>worm_development.obo</td>
<td>obo</td>
</tr>
<tr>
<td>C. elegans gross anatomy</td>
<td>WBbt</td>
<td>phenotype_ontology_obo.cgi</td>
<td>obo</td>
</tr>
<tr>
<td>C. elegans phenotype</td>
<td>WBPhenotype</td>
<td>cell.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Cell type</td>
<td>CL</td>
<td>cereals_development.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Cellular component</td>
<td>GO</td>
<td>plant_trait.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Cereal plant development</td>
<td>GRO</td>
<td>chebi.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Cereal plant trait</td>
<td>TO</td>
<td>caro.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Chemical entities of biological interest</td>
<td>CHEBI</td>
<td>dictvostelium_anatomy.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Common Anatomy Reference Ontology</td>
<td>CARO</td>
<td>fly_development.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Dicytostelium discoideum anatomy</td>
<td>DDANAT</td>
<td>fly_anatomy.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Drosophila development</td>
<td>FBdv</td>
<td>envo.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Drosophila gross anatomy</td>
<td>FBbt</td>
<td>envo.obo</td>
<td>obo</td>
</tr>
<tr>
<td>Environment Ontology</td>
<td>ENVO</td>
<td>envo.obo</td>
<td>obo</td>
</tr>
</tbody>
</table>

# 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>
<tr>
<td>Organ and Organism</td>
<td><strong>Organism (NCBI Taxonomy?)</strong></td>
<td><strong>Anatomical Entity (FMA, CARO)</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Phenotypic Quality (PaTO)</strong></td>
<td><strong>Biological Process (GO)</strong></td>
</tr>
<tr>
<td>Cell and Cellular Component</td>
<td><strong>Cell (CL)</strong></td>
<td><strong>Cellular Component (FMA, GO)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Molecule</td>
<td><strong>Molecule (ChEBI, SO, RnaO, PrO)</strong></td>
<td><strong>Molecular Function (GO)</strong></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
Applications

Biomedical Semantic Web
W3C Health Care and Life Sciences IG

W3C Semantic Web Health Care and Life Sciences Interest Group

The Semantic Web Health Care and Life Sciences Interest Group (HCISIG) is chartered to develop and support the use of Semantic Web technologies and practices to improve collaboration, research and development, and innovation adoption in the Health Care and Life Sciences domains. Success in these domains depends on a foundation of semantically rich system, process and information interoperability. [more]

Introduction

Both Life Science Research and Health Care are areas undergoing phenomenal growth, holding much promise for our future as long as we can manage and apply the new knowledge gained without driving up costs. Key to their success is the implementation of new informatics models that will unite many forms of biological and medical information across all institutions, through the encoding of meaning into the data and their interpretations. By focusing on the semantics of information, researchers will have more access to the knowledge required to effectively find cures to diseases, while doctors will have better tools for individualized clinical management of patients.

Mission and Scope

The Semantic Web for Health Care and Life Sciences Interest Group (HCISIG) is chartered to develop and support the use of Semantic Web technologies and practices to improve collaboration, research and development, and innovation adoption in the Health Care and Life Sciences domains. Success in these domains depends on a foundation of semantically rich system, process and information interoperability. [more]

News and Events

- Last Call: SPARQL Query Language for RDF 2007-03-27. Comments are due by 10 April (Permalink)
- HCIS demo, planned for WWW2007 in Banff. To help participate in the demo, please contact Alan Ruttenberg
- FIRST INTERNATIONAL WORKSHOP ON HEALTH CARE AND LIFE SCIENCES DATA INTEGRATION FOR THE SEMANTIC WEB, May 8, WWW2007 in Banff
- Eric Prud'hommeaux, new W3C staff contact for HCIS
- SPARQL links Microformats and Semantic Web: Working Draft, online: [http://www.w3.org/2000/svg](http://www.w3.org/2000/svg) (Permalink)

http://www.w3.org/2001/sw/hcls/
Biomedical Semantic Web

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

- **Hypothesis generation**

- **Knowledge discovery**

- **Clinical data**
  - Aggregation, sharing, exchange
  - Support for clinical decision
HCLS mashup of biomedical sources

Gene Ontology
Reactome
NeuronDB
BAMS
BrainPharm
MeSH
PubChem
Mammalian Phenotype
Homologene
AlzGene
Swan
NC Annotations
Antibodies
Entrez Gene
PDSPki
BrainPharm
Publications

http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo
Some unresolved 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

◆ Terminologies/Ontologies provide
  • Lists of entities
  • Names for entities
  • Identifiers for entities

◆ Additionally
  • Information model for integration
  • Trans-namespace resolution
  • Support for inference
Future directions

- Information integration
  - Knowledge extracted from text
  - Knowledge in structured knowledge bases
- Ontologies for relations
  - In complement to ontologies for entities
  - To support 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