From *terminology* integration
to *information* integration

*An example in the domain of genomics*

*Olivier Bodenreider*

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

◆ Background
  ● Terminology integration:
    *The Unified Medical Language System*
  ● Information integration:
    *Genomics as an example*

◆ Applications
  ● GenesTrace
  ● BioMeKe
Terminology integration

The Unified Medical Language System
Motivation

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

«[...] 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

- 134 source vocabularies
  - 126 contributing concept names
- 73 families of vocabularies
  - multiple translations (e.g., MeSH, ICPC, ICD-10)
  - variants (American-English equivalents, Australian extension/adaptation)
  - subsequent editions usually considered distinct families (ICD: 9-10; DSM: IIIR-IV)
- Broad coverage of biomedicine
- Common presentation
Biomedical terminologies

- **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 (cont’d)

- **Specialized vocabularies**
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART)
  - primary care (ICPC)
  - genomics (GO, OMIM, HUGO)

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

The UMLS serves as a vehicle for the regulatory standards (HIPAA, CHI)
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- GO
- SNOMED
- OMIM
- MeSH
- UWDA
- NCBI Taxonomy
- Model organisms
- Other subdomains
- Anatomy
- ...
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Anatomy
- Model organisms
- Other subdomains
UMLS: 3 components

- **Metathesaurus**
  - Concepts
  - Inter-concept relationships

- **Semantic Network**
  - Semantic types
  - Semantic network relationships

- **Lexical resources**
  - SPECIALIST Lexicon
  - Lexical tools
Addison’s Disease: Concept

Addison’s Disease

Disease or Syndrome

ADRENAL INSUFFICIENCY (ADDISON'S DISEASE)
ADRENOCORTICAL INSUFFICIENCY, PRIMARY FAILURE
Addison melanoderma
Melasma addisonii
Primary adrenal deficiency
Asthenia pigmentosa
Bronzed disease
Insufficiency, adrenal primary
Primary adrenocortical insufficiency
Addison's, disease

MALADIE D'ADDISON - French
Addison-Krankheit - German
Morbo di Addison - Italian
DOENCA DE ADDISON - Portuguese
ADDISONOVA BOLEZN' - Russian
ENFERMEDAD DE ADDISON - Spanish

A disease characterized by hypotension, weight loss, anorexia, weakness, and sometimes a bronze-like melanic 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 (2004AB)

- **Concept** (> 1M) CUI
  - Set of synonymous concept names
- **Term** (> 3.8 M) LUI
  - Set of normalized names
- **String** (> 4.3M) SUI
  - Distinct concept name
- **Atom** (> 5.1M) AUI
  - Concept name in a given source

| A0000001   | headache (source 1) | S0000001 |
| A0000002   | headache (source 2) |
| A0000003   | Headache (source 1)  | S0000002 |
| A0000004   | Headache (source 2)  |
| L0000001   |                     |
| A0000005   | Cephalgia (source 1) | S0000003 |
| L0000002   |                     |
| C0000001   |                     |
Cluster of synonymous terms

<table>
<thead>
<tr>
<th>Concept</th>
<th>C0001621</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>L0001621</th>
<th>S0011232 Adrenal Gland Diseases</th>
<th>S0011231 Adrenal Gland Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S0000441 Disease of adrenal gland</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S0481705 Disease of adrenal gland, NOS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S0220090 Disease, adrenal gland</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>S0044801 Gland Disease, Adrenal</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>L0041793</th>
<th>S0860744 Disorder of adrenal gland, unspecified</th>
<th>S0217833 Unspecified disorder of adrenal glands</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>L0161347</th>
<th>S0225481 ADRENAL DISORDER</th>
<th>S0627685 DISORDER ADRENAL (NOS)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>L0181041</th>
<th>S0632950 Disorder of adrenal gland</th>
<th>S0354509 Adrenal Gland Disorders</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>L0368399</th>
<th>S0586222 Adrenal disease</th>
<th>S0466921 ADRENAL DISEASE, NOS</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>L1279026</th>
<th>S1520972 Nebennierenkrankheiten</th>
<th>GER</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>L0162317</th>
<th>S0226798 SURRENALE, MALADIES</th>
<th>FRE</th>
</tr>
</thead>
</table>

...
Metathesaurus  Evolution over time

- Concepts never die (in principle)
  - CUIs are permanent identifiers

- What happens when they do die (in reality)?
  - Concepts can merge or split
  - Resulting in new concepts and deletions

Addison’s disease
C0001403

Addison’s disease, NOS
C0271735

Metathesaurus Relationships

- Symbolic relations: ~9 M pairs of concepts
- Statistical relations: ~7 M pairs of concepts (co-occurring concepts)
- Mapping relations: 100,000 pairs of concepts

Categorization: Relationships between concepts and semantic types from the Semantic Network
# Symbolic relations

- **Relation**
  - Pair of “atom” identifiers
  - Type
  - Attribute (if any)
  - List of sources (for type and attribute)

- **Semantics of the relationship:**
  defined by its **type** [and attribute]

Source transparency: the information is recorded at the “atom” level
## Symbolic relationships

### Type

- **Hierarchical**
  - Parent / Child
  - Broader / Narrower than

- **Derived from hierarchies**
  - Siblings (children of parents)

- **Associative**
  - Other

- **Various flavors of near-synonymy**
  - Similar
  - Source asserted synonymy
  - Possible synonymy
Symbolic relationships

- Hierarchical
  - isa (is-a-kind-of)
  - part-of

- Associative
  - location-of
  - caused-by
  - treats
  - ...

- Cross-references (mapping)
Semantic Types

- Anatomical Structure
- Embryonic Structure
- Disease or Syndrome
- Pharmacologic Substance
- Population Group
- Body Part, Organ or Organ Component

Concepts

- Heart
- Esophagus
- Mediaspinum
- Saccular Viscus
- Left Phrenic Nerve
- Heart Valves
- Fetal Heart
- Angina Pectoris
- Cardiotonic Agents
- Tissue Donors

Semantic Network

Metathesaurus
Lexical tools

- To manage lexical variation in biomedical terminologies

- Major tools
  - Normalization
  - Indexes
  - Lexical Variant Generation program (lvg)

- Based on the SPECIALIST Lexicon

- Used by noun phrase extractors, search engines
Normalization

- Remove genitive: Hodgkin’s diseases, NOS
- Remove stop words: Hodgkin diseases, NOS
- Lowercase: Hodgkin diseases,
- Strip punctuation: hodgkin diseases,
- Uninflect: hodgkin diseases
- Sort words: hodgkin disease

disease hodgkin
Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
disease, Hodgkin

normalize
disease hodgkin
Information integration

Genomics as an example
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

OMIM Genbank

External resources

Neoplastic Process
Gene or Genome
Biologically Active Substance
Amino Acid, Peptide, or Protein

Neurofibromatoses
Benign neoplasms of cranial nerves
Tumor suppressor genes
Tumor suppressor proteins

UMLS Semantic Network (Semantic Types)
Amino Acid, Peptide, or Protein

UMLS Metathesaurus
(Concepts and relations)
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)
Applications (1)

*GenesTrace™*

*Lussier Lab*
*Columbia University*
Objectives

◆ Relate diseases to genes through structured, integrated terminologies

◆ Biological Knowledge Discovery
Resources and Methods

1. Start from a disease in UMLS
2. Select related concepts
3. Map related UMLS concepts to genes and GO terms
4. Relate GO terms to genes

UMLS

Genes (LocusLink)

Annotation terms (GO)

disease

Lister Hill National Center for Biomedical Communications
Validation  Breast cancer – BRCA1 association

1. Disease = Breast neoplasms
2. 2129 related concepts
3. Several related genes and 168 related GO terms
4. 10,000 gene products associated (including BRCA1)
Limitations

◆ Noise
  - Too many non-specific GO terms associated (e.g., nucleus)
  - Too many genes associated

◆ But
  - Promising preliminary results
  - Room for refinement
Search GenesTrace for diseases and genes sharing identical processes, functions or biological structures.

Disease or Gene......

Query using............. Disease complete name in UMLS

Please select the species database from below
- Fly(FB)
- Mouse(MGI)
- Worm(WB)
- Yeast(SGD)
- Swissprot(SPTR)

Submit Query
Applications (2)

**BioMeKe**

G. Marquet & al.
*LIM, Univ. Rennes, France*
Objectives

◆ To develop a **knowledge warehouse for transcriptome analysis (liver diseases)**

◆ **Semantic interoperability**
  - Medical knowledge bases
    - *Clinical genomics*
  - Molecular biology and genetics knowledge bases
    - *Functional genomics*
Components

Core Ontology

- HUGO
- UMLS
- GO
- Annotations
- Swiss-Prot
- GenBank
- MEDLINE

Cross-referenced resources

Query Processor

- Biological search module
- Heterogeneity manager
- Medical search module
Example

- **Input:** *ferritin, heavy polypedpide 1*
- **Mapping to biological resources**
  - Not found in the Core ontology
  - Official name *Ferritin heavy chain* found through Xref
- **Biological information obtained from GOA**
- **Mapping to medical resources**
  - Not found in UMLS
  - Synonym *Ferritin H* found through Xref (Swiss-Prot)
- **Medical information obtained through co-occurrence of MeSH index terms in MEDLINE**
Results

- iron binding protein
- iron ion homeostasis
- intracellular iron ion storage
- cell proliferation
- ferritin complex

- liver
- hemochromatosis
- cataract
- ...
Limitations

- Non-formal ontologies
  - Knowledge may be inconsistently represented
  - Knowledge may be implicit (mappings)

- Partial automation
  - User input required to select databanks, reformulate queries

- Semantic integration
  - Naming issues
  - Mappings must be updated regularly
BioMeKe

Biological and Medical Knowledge Extraction system

BioMeKe has been achieved to extract and to associate Medical and Biological information such as Gene Ontology™, Unified Medical Language System®, Genew ...

BioMeKe is an ontology-based tool composed of two-part:

- Consultation of Ontology and link with Public Databank
- Biological and Medical annotation

Genomic and Biological Resources in BioMeKe:

- Gene: Lexical resources providing official gene names and their synonyms and links to multiple others databases including UniProt, LocusLink.
- Gene Ontology (GO): general "ontology" for molecular biology which provides a controlled vocabulary for annotating sequences and genes products.
- GeneOntology Annotation@EBI (GOA): provides assignments of GO terms to genes products for all organisms, including human.
Conclusions
Conclusions

- Terminology integration provides some degree of information integration
- Most terminologies and the cross-referenced databases are readily available
- Lack of consistent representation
- Additional resources/techniques needed
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
Questions

◆ What do I need to do to get the UMLS?
◆ What is an ontology?
◆ How is ontology different from
  ● Terminology? / Database? / Knowledge base?
◆ Is the UMLS an ontology?
◆ Does the UMLS use Protégé?
◆ I heard of OWL. Is that any good?
◆ What is the Semantic Web going to do for us?
References UMLS

- **UMLS**
  
  umlsinfo.nlm.nih.gov

- **UMLS browser**
  
  

  (free, but UMLS license required)

- **UMLS and information integration**
  
References

GenesTrace

- http://phene.cpmc.columbia.edu:8080/genesTrace/index.jsp

BioMeKE

- http://www.med.univ-rennes1.fr/~marquet/