School of Information and Library Science
University of North Carolina at Chapel Hill
October 27, 2003

From *terminology* integration

to *information* integration

*An example in the domain of genetics*

*Olivier Bodenreider*

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

◆ Background
  ● Terminology integration:
    *The Unified Medical Language System*
  ● Information integration:
    *Genetics 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

- 117 “sources”
- ~60 families of vocabularies
  - multiple translations (e.g., MeSH, ICPC, ICD-10)
  - variants (American-English equivalents, Australian extension/adaptation)
  - subsequent versions usually considered distinct families (ICD: 9-10; DSM: IIIR-IV)
- Broad coverage of biomedicine
- Common presentation
Biomedical terminologies

◆ Core vocabularies
  ○ anatomy (UWDA, Neuronames)
  ○ drugs (First DataBank, Micromedex)
  ○ medical devices (UMD, SPN)

◆ Several perspectives
  ○ clinical terms (SNOMED, CTV3)
  ○ information sciences (MeSH, CRISP)
  ○ administrative terminologies (ICD-9-CM, CPT-4)
  ○ standards (HL7, LOINC)
Biomedical terminologies (cont’d)

◆ Specialized vocabularies
  ● nursing (NIC, NOC, NANDA, Omaha, PCDS)
  ● dentistry (CDT)
  ● oncology (PDQ)
  ● psychiatry (DSM, APA)
  ● adverse reactions (COSTART, WHO ART)
  ● primary care (ICPC)

◆ Knowledge bases (AI/Rheum, DXplain, QMR)
Integrating subdomains

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

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.

SNOMED
MeSH
AOD
Read Codes
...
Metathesaurus Concepts (2003AA)

- **Concept**: Cluster of synonymous terms
  - ~875,000 concepts
  - identified by a **CUI**

- **Term**: Set of lexical variants
  - ~1.8 M terms
  - identified by a **LUI**

- **String**: Concept name
  - ~2.1 M strings
  - identified by a **SUI**
### Cluster of synonymous terms

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

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

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

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

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

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

<table>
<thead>
<tr>
<th>Term</th>
<th>L0162317</th>
<th>S0226798</th>
<th>SURRENALE, MALADIES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>FRE</td>
</tr>
</tbody>
</table>
Metathesaurus Relationships

◆ Symbolic relations: ~5 M pairs of concepts

◆ Statistical relations: ~6.5 M pairs of concepts
  (co-occurring concepts)

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

- Relation
  - Pair of concept identifiers
  - Type
  - Attribute (if any)
  - List of sources (for type and attribute)

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

- **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
- Associateive
  - location-of
  - caused-by
  - treats
  - ...
- Cross-references (mapping)
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

Genetics 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)  
\[C0027832\]

Merlin (Schwannomin, Neurofibromin 2) \[C0254123\]

NF2 (Neurofibromin 2 gene) \[C0085114\]

UMLS Metathesaurus (Concepts and relations)

Merlin, Drosophila

NEUROFIBROMATOSIS, TYPE II; NF2  
#101000  
OMIM

Drosophila melanogaster merlin (Dmerlin) mRNA, complete cds. \[U49724\]

External resources

Amino Acid, Peptide, or Protein

Biologically Active Substance

Gene or Genome

Neoplastic Process

Tumor suppressor proteins

Tumor suppressor genes

Benign neoplasms of cranial nerves

Genome or Gene

Neurofibromatoses

UMLS Semantic Network (Semantic Types)
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™

I.N. Sarkar & al.
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
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
References

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
  - Molecular biology and genetics knowledge bases
Example

- **Input**: *ferritin, heavy polypeptide 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

**FTH1** -> BioMeKe

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

- liver
- hemochromatosis
- cataract
- …

**Biological annotations**

**Medical annotations**
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
References

Conclusions
Conclusions

- Terminology integration provides some degree of information integration
- Most terminologies and the cross-referenced databases are readily available
- Lack of consistent representation
References

◆ UMLS

[umlsinfo.nlm.nih.gov](umlsinfo.nlm.nih.gov)

◆ UMLS browser

  - Semantic Navigator: KSS, under UMLSKS resources
  - (free, but UMLS license required)

◆ UMLS and information integration

Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: etbsun2.nlm.nih.gov:8000

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