From terminology integration
to information integration

Unified Medical Language System (UMLS)

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

◆ Overview through an example
◆ UMLS components
  ● Lexical resources
  ● Metathesaurus
  ● Semantic Network
◆ UMLS and information integration
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.”
Overview through an example
Addison's disease is a rare endocrine disorder. Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol. For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism.
Adrenal insufficiency Clinical variants

◆ Primary / Secondary
  - Primary: lesion of the adrenal glands themselves
  - Secondary: inadequate secretion of ACTH by the pituitary gland

◆ Acute / Chronic

◆ Isolated / Polyendocrine deficiency syndrome
Addison’s disease: Symptoms

- Fatigue
- Weakness
- Low blood pressure
- Pigmentation of the skin (exposed and non-exposed parts of the body)
- ...

...
AD in medical vocabularies

◆ **Synonyms: different terms**
  - Addisonian syndrome
  - Bronzed disease
  - Addison melanoderma
  - Asthenia pigmentosa
  - Primary adrenal deficiency
  - Primary adrenal insufficiency
  - Primary adrenocortical insufficiency
  - Chronic adrenocortical insufficiency

◆ **Contexts: different hierarchies**

[Diagram showing relationships between terms: eponym, symptoms, clinical, variants]
Organize terms

- Synonymous terms clustered into a concept
- Preferred term
- Unique identifier (CUI)

<table>
<thead>
<tr>
<th>Term</th>
<th>MeSH</th>
<th>D000224</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td>MedDRA</td>
<td>10036696</td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td>ICD-10</td>
<td>E27.1</td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td>SNOMED CT</td>
<td>363732003</td>
</tr>
</tbody>
</table>

C0001403

Addison's disease
Diseases of the endocrine system

Diseases of the Adrenal Glands

Addison’s Disease
Endocrine disorder

Adrenal disorder

Adrenal cortical disorder

Adrenal cortical hypofunction

Addison’s Disease
Read Codes

Endocrine disorder

Disorder of adrenal gland

Hypoadrenalism

Adrenal Hypofunction

Corticoadrenal insufficiency

Addison’s Disease
Primary adrenocortical insufficiency

Other disorders of adrenal gland

Disorders of other endocrine gland
Organize concepts

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

Hypoadrenalinism

Adrenal Gland Hypofunction

Adrenal cortical hypofunction

Addison’s Disease

Endocrine Diseases

Adrenal Gland Diseases

Adrenal Cortex Diseases

Hypoadrenalinism

Adrenal Gland Hypofunction

Adrenal cortical hypofunction

Addison’s Disease

SNOMED
MeSH
AOD
Read Codes

UMLS
Relate to other concepts

- Additional hierarchical relationships
  - link to other trees
  - make relationships explicit
- Non-hierarchical relationships
- Co-occurring concepts
- Mapping relationships
Adrenal Gland Diseases

Adrenal Cortex Diseases

Adrenal Cortex Dysfunction

Adrenal Gland Hypofunction

Adrenal cortical hypofunction

Hypoadrenalism

Secondary hypocortisolism

Addison’s Disease

Addison’s disease due to autoimmunity

Other disorders of adrenal gland

Disorders of other endocrine gland

Other disorders of other endocrine gland

relate to other concepts
Categorize concepts

- High-level categories (semantic types)
- Assigned by the Metathesaurus editors
- Independently of the hierarchies in which these concepts are located
How do they do that?

- Lexical knowledge
- Semantic pre-processing
- UMLS editors
Lexical knowledge

Adrenal gland diseases
Adrenal disorder
Disorder of adrenal gland
Diseases of the adrenal glands
C0001621
Semantic pre-processing

- Metadata in the source vocabularies
- Tentative categorization
- Positive (or negative) evidence for tentative synonymy relations based on lexical features
Additional knowledge: UMLS editors

Adrenal Gland Diseases

Adrenal Cortex Diseases

Adrenal Cortex Dysfunction

Hypoadrenalism

Adrenal Gland Hypofunction

Adrenal cortical hypofunction

Addison’s Disease

Other disorders of adrenal gland
UMLS: 3 components

◆ 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
UMLS Metathesaurus
Source Vocabularies

- 139 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 5.1M names
  - 1.3M concepts
  - 16M relations
- Common presentation
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 (2006AB)

- **Concept** (> 1.3M) CUI
  - Set of synonymous concept names
- **Term** (> 4.6M) LUI
  - Set of normalized names
- **String** (> 5.1M) SUI
  - Distinct concept name
- **Atom** (> 6.2M) AUI
  - Concept name in a given source

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

<table>
<thead>
<tr>
<th>Concept</th>
<th>C0001621</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term L0001621</td>
<td>S0011232 Adrenal Gland Diseases  S0011231 Adrenal Gland Disease  S0000441 Disease of adrenal gland  S0481705 Disease of adrenal gland, NOS  S0220090 Disease, adrenal gland  S0044801 Gland Disease, Adrenal</td>
</tr>
<tr>
<td>Term L0041793</td>
<td>S0860744 Disorder of adrenal gland, unspecified  S0217833 Unspecified disorder of adrenal glands</td>
</tr>
<tr>
<td>Term L0161347</td>
<td>S0225481 ADRENAL DISORDER  S0627685 DISORDER ADRENAL (NOS)</td>
</tr>
<tr>
<td>Term L0181041</td>
<td>S0632950 Disorder of adrenal gland  S0354509 Adrenal Gland Disorders</td>
</tr>
<tr>
<td>Term L0368399</td>
<td>S0586222 Adrenal disease  S0466921 ADRENAL DISEASE, NOS</td>
</tr>
<tr>
<td>Term L1279026</td>
<td>S1520972 Nebennierenkrankheiten  GER</td>
</tr>
<tr>
<td>Term L0162317</td>
<td>S0226798 SURRENALE, MALADIES  FRE</td>
</tr>
</tbody>
</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

- **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

Types:
- PAR/CHD
- RB/RN
- SIB
- RO
- RL
- SY
- RQ
Symbolic relationships

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

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

- Cross-references (mapping)
UMLS Semantic Network
Semantic Network

- Semantic types (135)
  - tree structure
  - 2 major hierarchies
    - Entity
      - Physical Object
      - Conceptual Entity
    - Event
      - Activity
      - Phenomenon or Process
Semantic Network

◆ Semantic network relationships (54)
  ● hierarchical (isa = is a kind of)
    ■ among types
      - Animal isa Organism
      - Enzyme isa Biologically Active Substance
    ■ among relations
      - treats isa affects
  ● non-hierarchical
    ■ Sign or Symptom diagnoses Pathologic Function
    ■ Pharmacologic Substance treats Pathologic Function
“Biologic Function” hierarchy (isa)

- Biologic Function
  - Physiologic Function
    - Organ Function
      - Mental Process
    - Organ or Tissue Function
    - Cell Function
      - Genetic Function
  - Molecular Function
  - Pathologic Function
    - Cell or Molecular Dysfunction
      - Mental or Behavioral Dysfunction
    - Disease or Syndrome
    - Experimental Model of Disease
      - Neoplastic Process
Why a semantic network?

- Semantic Types serve as high level categories assigned to Metathesaurus concepts, *independently of their position in a hierarchy*.

- A relationship between 2 Semantic Types (ST) is a possible link between 2 concepts that have been assigned to those STs.
  - The relationship may or may not hold at the concept level.
  - Other relationships may apply at the concept level.
Relationships can inherit semantics

Semantic Network

- Fully Formed Anatomical Structure
- Body Part, Organ, or Organ Component
- Disease or Syndrome
- Adrenal Cortex
- Adrenal Cortical hypofunction

Metathesaurus

location of

isa

Biologic Function
Pathologic Function

location of
UMLS Summary

- Synonymous terms clustered into concepts
- Unique identifier
- Finer granularity
- Broader scope
- Additional hierarchical relationships
- Semantic categorization
Integrating subdomains

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

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

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

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 Metathesaurus (Concepts and relations)

UMLS Semantic Network (Semantic Types)

Amino Acid, Peptide, or Protein

Biologically Active Substance

Neoplastic Process

Gene or Genome

UMLSemantic 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)
References

◆ UMLS
  umlsinfo.nlm.nih.gov

◆ UMLS browsers
  (free, but UMLS license required)
  ● Knowledge Source Server: umlsks.nlm.nih.gov
  ● Semantic Navigator:
  ● RRF browser
    (standalone application distributed with the UMLS)
References

◆ Recent overviews


References

◆ UMLS as a research project


References

◆ Technical papers


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