The Unified Medical Language System

Integrating Biomedical Terminology

Olivier Bodenreider

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

Public Health Informatics Fellowship Program
October 14, 2005
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
  - 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>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>D000224</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
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

Hypoadrenalism

Adrenal Gland Hypofunction

Addison’s Disease

Adrenal Gland Hypofunction

Adrenal Gland Diseases

Endocrine Diseases

organize concepts

SNOMED
MeSH
AOD
Read Codes

UMLS
Relate to other concepts

- Additional hierarchical relations
  - link to other trees
  - make relationships explicit
- Non-hierarchical relations
- Co-occurring concepts
- Mapping relations
Categorize concepts

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

Diagram:
- Disease or Syndrome
  - Diseases
    - Endocrine Diseases
      - Adrenal Gland Diseases
        - Adrenal Gland Hypofunction
          - Addison’s Disease
UMLS Knowledge Sources
UMLS 3 components

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

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

- **Lexical resources**
  - SPECIALIST Lexicon
  - Lexical tools
UMLS Metathesaurus
Metathesaurus Basic organization

◆ Concepts
  - Synonymous terms are clustered into a concept
  - Properties are attached to concepts, e.g.,
    - Unique identifier
    - Definition

◆ Relations
  - Concepts are related to other concepts
  - Properties are attached to relations, e.g.,
    - Type of relationship
    - Source
Source Vocabularies

- 133 source vocabularies contributing concept names
- ~80 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 (file format + Unicode)
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)
  • oncology (PDQ)
  • psychiatry (DSM, APA)
  • adverse reactions (COSTART, WHO ART)
  • primary care (ICPC)

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

The UMLS serves as a vehicle for the regulatory standards (HIPAA, CHI)
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 (2005AB)

- **Concept** (~ 1.2 M) CUI
  - Set of synonymous concept names
- **Term** (~ 4.2 M) LUI
  - Set of normalized names
- **String** (~ 4.8 M) SUI
  - Distinct concept name
- **Atom** (~ 5.6 M) AUI
  - Concept name in a given source

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0000001</td>
<td>headache</td>
<td>(source 1)</td>
</tr>
<tr>
<td>A0000002</td>
<td>headache</td>
<td>(source 2)</td>
</tr>
<tr>
<td>S0000001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A0000003</td>
<td>Headache</td>
<td>(source 1)</td>
</tr>
<tr>
<td>A0000004</td>
<td>Headache</td>
<td>(source 2)</td>
</tr>
<tr>
<td>S0000002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L0000001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A0000005</td>
<td>Cephalgia</td>
<td>(source 1)</td>
</tr>
<tr>
<td>S0000003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L0000002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0000001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Cluster of synonymous terms

<table>
<thead>
<tr>
<th>Concept</th>
<th>Term</th>
<th>Term ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0001403</td>
<td>L0001403</td>
<td>S0354372</td>
<td>Addison's disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S0010792</td>
<td>Addison Disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S0010794</td>
<td>Addison's Disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S0010796</td>
<td>Addisons Disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S0033587</td>
<td>Disease, Addison</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S0352253</td>
<td>ADDISON'S DISEASE</td>
</tr>
<tr>
<td></td>
<td>L2799243</td>
<td>S3341310</td>
<td>Addison's disease (disorder)</td>
</tr>
<tr>
<td></td>
<td>L0494940</td>
<td>S5907336</td>
<td>Primary Adrenocortical Insufficiency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S5901878</td>
<td>Insufficiencies, Primary Adrenocortical</td>
</tr>
<tr>
<td></td>
<td>L0494851</td>
<td>S5907334</td>
<td>Primary Adrenal Insufficiency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4094828</td>
<td>adrenal; insufficiency, primary</td>
</tr>
<tr>
<td></td>
<td>L0585243</td>
<td>S5907343</td>
<td>Primary Hypoadrenalism</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S5901432</td>
<td>Hypoadrenalism, Primary</td>
</tr>
<tr>
<td></td>
<td>L1229627</td>
<td>S1471573</td>
<td>Addison-Krankheit</td>
</tr>
<tr>
<td></td>
<td>L5345155</td>
<td>S6107160</td>
<td>Maladie d'Addison</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FRE</td>
<td></td>
</tr>
</tbody>
</table>

Note: [...]

Source: NLM
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
C021935
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

Type
- PAR/CHD
- RB/RN
- SIB
- RO
- RL
- SY
- RQ
Symbolic relationships  Attribute

- 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

Semantics 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
    - Disease or Syndrome
      - Mental or Behavioral Dysfunction
    - Experimental Model of Disease
      - Neoplastic Process
Associative (non-isa) relationships

- Organism
  - Anatomical Structure
    - Embryonic Structure
      - Congenital Abnormality
    - Anatomical Abnormality
      - Acquired Abnormality
    - Body System
      - Body Part, Organ or Organ Component
      - Tissue
      - Cell
      - Cell Component
      - Gene or Genome

- Process of
  - Fully Formed Anatomical Structure
    - contains, produces

- Evaluation of
  - Finding
    - Laboratory or Test Result
    - Sign or Symptom
    - Injury or Poisoning
    - Physiologic Function
    - Pathologic Function

- Location of
  - Body Location or Region
  - Body Space or Junction

- Conceptual Part of
  - Gene or Genome
    - part of
  - Cell Component
    - part of
  - Cell
    - part of
  - Tissue
    - part of
  - Body Part, Organ or Organ Component
    - part of

- Disrupts
  - Body Substance
    - disrupts
    - location of
  - Biologic Function
    - disrupts
    - location of

- Co-occurs with
  - Injury or Poisoning
    - co-occurs with
    - location of

- Adjacent to
  - Body Space or Junction
    - adjacent to
  - Body Part, Organ or Organ Component
    - adjacent to
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
- Adrenal Cortex
- Adrenal Cortical hypofunction
- Disease or Syndrome
- Biologic Function
- Pathologic Function

Metathesaurus

location of

isa

location of
SPECIALIST Lexicon and lexical tools
Lexical tools

- To manage lexical variation in biomedical terminologies

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

- 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
Conclusions
Integrating subdomains

Clinical repositories

Genetic knowledge bases

Biomedical literature

Genome annotations

UMLS

MeSH

SNOmed

OMIM

Model organisms

NCBI Taxonomy

UWDA

Anatomy

Other subdomains

...
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
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
