Unified Medical Language System

Overview

Olivier Bodenreider

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

◆ Introduction

◆ Overview through an example
  
  *Addison’s disease*

◆ The three UMLS Knowledge Sources
  
  - UMLS Metathesaurus
  
  - UMLS Semantic Network
  
  - SPECIALIST Lexicon and lexical tools
Introduction
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”
- 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.»
The UMLS in practice

- **Database**
  - Series of relational files

- **Interfaces**
  - Web interface: Knowledge Source Server (UTS)
  - Application programming interfaces (Java and web services)

- **Applications**
  - lvg (lexical programs)
  - MetamorphoSys (installation and customization)
  - RRF browser (browsing subsets)

The UMLS is *not* an end-user application
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
- Melasma addisonii
- 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>D000224</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td></td>
<td>10036696</td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td>MedDRA</td>
<td>E27.1</td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td>ICD-10</td>
<td></td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td>SNOMED CT</td>
<td>363732003</td>
</tr>
<tr>
<td>C0001403</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Addison's disease
Disease

Endocrine system diseases

Adrenal gland diseases

Adrenal Insufficiency

Addison Disease
Diseases of the endocrine system

Diseases of the adrenal glands

Addison's Disease
Disorder of endocrine system

Disorder of adrenal gland

Hypoadrenalism

Adrenal hypofunction

Disorder of adrenal cortex

Adrenal cortical hypofunction

Addison's Disease
Disorder of endocrine system

Disorder of adrenal gland

Adrenal gland hypofunction

Disorder of adrenal cortex

Adrenal cortical hypofunction

Addison's Disease
Non-neoplastic endocrine disorder

Adrenal gland disorder

Non-neoplastic adrenal gland disorder

Adrenal gland insufficiency

Adrenal cortical insufficiency

Addison’s Disease
Organize concepts

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

Disease

Endocrine system diseases

Adrenal gland diseases

Adrenal gland hypofunction

Adrenal cortex diseases

Adrenal cortical hypofunction

Addison's Disease

SNOMED CT
SNOMED Intl
MeSH
MedDRA
Disease

Endocrine system diseases

- Non-neoplastic endocrine disorder
- Non-neoplastic adrenal gland disorder

Adrenal gland diseases

- Adrenal gland hypofunction
- Adrenal cortical hypofunction

Adrenal cortex diseases

- Addison's Disease

Endocrine / nutritional / metabolic disorder

Disorders of other endocrine glands

Other disorders of adrenal gland
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
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 gland hypofunction
    - Addison’s Disease
  - Adrenal cortex diseases
  - Other disorders of adrenal gland

- Adrenal cortical hypofunction
UMLS Summary

- Synonymous terms clustered into concepts
- Unique identifier

- Finer granularity
- Broader scope
- Additional hierarchical relationships
- Semantic categorization
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 Knowledge Sources

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

- 161 source vocabularies
- 21 languages
- Broad coverage of biomedicine
  - 8.7M distinct names
  - 2.6M concepts
  - >10M relations
- Common presentation
Biomedical terminologies

◆ General vocabularies
  ● anatomy (FMA, 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 (MedDRA, WHO ART)
  ● primary care (ICPC)

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

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

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

UMLS
Integrating subdomains

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

Lister Hill National Center for Biomedical Communications
Trans-namespace integration

Addison's disease (363732003)
Other subdomains

Clinical repositories

SNOMED CT

NCBI Taxonomy

Model organisms

UMLS C0001403

Genetic knowledge bases

OMIM

UMLS

NCBI

Genome annotations

FMA

UMLS

Biomedical literature

MeSH

Clinical repositories

Anatomy

Clinical repositories

Addison Disease (D000224)

MeSH

Clinical repositories

Clinical repositories

Clinical repositories

Clinical repositories

Clinical repositories

Clinical repositories
Addison's Disease: Concept

An adrenal disease characterized by the progressive destruction of the adrenal cortex, resulting in insufficient production of aldosterone and hydrocortisone. Clinical symptoms include anorexia; nausea; weight loss; muscle weakness; and hyperpigmentation of the skin due to increase in circulating levels of ACTH precursor hormone which stimulates melanocytes.
Lister Hill National Center for Biomedical Communications

Metathesaurus Concepts (2011AA)

- **Concept** (2.6M) CUI
  - Set of synonymous concept names
- **Term** (7.9M) LUI
  - Set of normalized names
- **String** (8.9M) SUI
  - Distinct concept name
- **Atom** (10.6M) AUI
  - Concept name in a given source

<table>
<thead>
<tr>
<th>A0066000</th>
<th>Headache (MeSH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0065992</td>
<td>Headache (ICD-10)</td>
</tr>
<tr>
<td></td>
<td>S0046854</td>
</tr>
<tr>
<td>A0066007</td>
<td>Headaches (MedDRA)</td>
</tr>
<tr>
<td>A12003304</td>
<td>Headaches (OMIM)</td>
</tr>
<tr>
<td></td>
<td>S0046855</td>
</tr>
<tr>
<td>L0018681</td>
<td></td>
</tr>
<tr>
<td>A0540936</td>
<td>Cephalodynia (MeSH)</td>
</tr>
<tr>
<td></td>
<td>S0475647</td>
</tr>
<tr>
<td></td>
<td>L0380797</td>
</tr>
<tr>
<td>C0018681</td>
<td></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

![Diagram showing the evolution of concepts over time with examples of Addison's disease and its codes.]( Diagram showing the evolution of concepts over time with examples of Addison's disease and its codes. )
Metathesaurus Relations

- Symbolic relations:  ~8 M pairs of concepts
- Statistical relations:  ~6 M pairs of concepts (co-occurring concepts)
- Mapping relations:  ~150,000

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
Mapping relations

◆ Simple mappings
  ● \langle atom 1 \rangle \text{ mapped\_to } \langle atom 2 \rangle
  ● e.g.,
    ▪ SNOMED CT to ICD-9-CM

◆ Complex mappings
  ● \langle atom 1 \rangle \text{ mapped\_to } \langle boolean expression \rangle
  ● e.g.,
    ▪ ICD-9-CM to MeSH (search strategies)

NB: partially redundant with relations in MRREL
Everything else

- **Co-occurrence information (MRCOC)**
  - Co-occurrence of MeSH descriptors in MEDLINE for the most part

- **Source-specific attributes (MRSAT)**
  - Legacy identifiers, external cross-references
    - SNOMED International legacy codes (SNOMED CT)
    - RxNorm to NDC
  - Concept status in a particular source (SNOMED CT)
  - Frequency of occurrence in MEDLINE (MeSH)
  - MedlinePlus URL (MeSH)
  - …
UMLS Knowledge Sources

UMLS Semantic Network
Semantic Network

- Semantic types (133)
  - tree structure
  - 2 major hierarchies
    - Entity
      - Physical Object
      - Conceptual Entity
    - Event
      - Activity
      - Phenomenon or Process
Semantic Network

- Semantic network
  - 54 relationships
  - 603 asserted relations
  - 6101 inferred relations

- Asserted semantic network relations (603)
  - hierarchical (isa = is a kind of)
    - among types (133)
      - Animal isa Organism
      - Enzyme isa Biologically Active Substance
    - among relations (54)
      - treats isa affects
  - non-hierarchical (416)
    - Sign or Symptom diagnoses Pathologic Function
    - Pharmacologic Substance treats Pathologic Function
“Biologic Function” hierarchy (isa)
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 *may* inherit semantics
UMLS Knowledge Sources

UMLS SPECIALIST Lexicon
SPECIALIST Lexicon

◆ **Content**
  - English lexicon
  - Many words from the biomedical domain

◆ **465,000 lexical items**

◆ **Word properties**
  - morphology
  - orthography
  - syntax

◆ **Used by the lexical tools**
Morphology

◆ Inflection
  ● noun  nucleus, nuclei
  ● verb  cauterize, cauterizes, cauterized, cauterizing
  ● adjective  red, redder, reddest

◆ Derivation
  ● verb ↔ noun  cauterize -- cauterization
  ● adjective ↔ noun  red -- redness
Orthography

◆ Spelling variants

- **oe/e**
  - oesophagus - esophagus

- **ae/e**
  - anaemia - anemia

- **ise/ize**
  - cauterise - cauterize

- **genitive mark**
  - Addison's disease
  - Addison disease
  - Addisons disease
Syntax

◆ Complementation

- verbs
  - intransitive: I'll treat.
  - transitive: He treated the patient.
  - ditransitive: He treated the patient with a drug.

- nouns
  - prepositional phrase: Valve of coronary sinus

◆ Position for adjectives
SPECIALIST Lexicon record

{  
  base=hemoglobin (base form)  
  spelling_variant=haemoglobin  
  entry=E0031208 (identifier)  
  cat=noun (part of speech)  
  variants=uncount (no plural)  
  variants=reg (plural: hemoglobins, hemoglobins)  
}
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
Summary
Integrating subdomains

Clinical repositories

Genetic knowledge bases

Other subdomains

SNOMED CT

OMIM

Biomedical literature

NCBI Taxonomy

MeSH

Genome annotations

Model organisms

FMA

GO

Anatomy

UMLS
Trans-namespace integration

Addison's disease
(363732003)

Other subdomains

Clinical repositories

Snomed CT

Genetic knowledge bases

OMIM

Biomedical literature

Addison Disease
(D000224)

NCBI Taxonomy

UMLS
C0001403

FMA

GO

Genome annotations

Model organisms

Anatomy

MeSH

Clinical repositories

MeSH

Genetic knowledge bases

OMIM

Biomedical literature

Addison Disease
(D000224)
Other things you would need to know

◆ UMLLS license agreement

◆ MetamorphoSys

◆ UMLLS Terminology Services (UTS)
  (formerly, UMLLS Knowledge Source Server)
Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider
Lister Hill National Center for Biomedical Communications
Bethesda, Maryland - USA
References
References: UMLS home page

◆ UMLS home page
  ● http://www.nlm.nih.gov/research/umls/

◆ UMLS documentation
  ● Formerly known as the “Green Book”
  ● Now online documentation
  ● http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML

◆ UMLS online tutorials
References

◆ Recent overviews


References

◆ UMLS as a research project


References

◆ Technical papers
