Introduction to the Unified Medical Language System

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

- Introduction
- Overview through an example
- The three UMLS Knowledge Sources
  - UMLS Metathesaurus
  - UMLS Semantic Network
  - SPECIALIST Lexicon and lexical tools
- UMLS in action: *MetaMap*
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 (UMLSKS)
  - Application programming interfaces (Java and XML-based)

- **Applications**
  - lvg (lexical programs)
  - MetamorphoSys (installation and customization)

The UMLS is *not* an end-user application
Overview through an example
Addison’s disease

- 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

ACTH

Pituitary gland
Parathyroid gland
Thyroid gland
Thymus gland
Adrenal gland
Pancreas
Ovary (female)
Testis (male)
Addison’s disease: Symptoms

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

10
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>Source</th>
<th>Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adrenal gland diseases</td>
<td>MeSH</td>
<td>D000307</td>
</tr>
<tr>
<td>Adrenal disorder</td>
<td>AOD</td>
<td>0000005418</td>
</tr>
<tr>
<td>Disorder of adrenal gland</td>
<td>Read</td>
<td>C15z.</td>
</tr>
<tr>
<td>Diseases of the adrenal glands</td>
<td>SNOMED</td>
<td>DB-70000</td>
</tr>
</tbody>
</table>

C0001621

Adrenal Gland Diseases
Endocrine Diseases

Adrenal Gland Diseases

Adrenal Gland Hypofunction

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

ICD-10
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

Adrenal Cortical Hypofunction

Addison’s Disease

Endocrine Diseases

Adrenal Gland Diseases

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
Endocrine Diseases

Adrenal Gland Diseases

Adrenal Cortex Diseases

Hypoadrenalism

Adrenal Gland Hypofunction

Adrenal cortical hypofunction

Secondary hypocortisolism

Addison’s Disease

Addison’s disease due to autoimmunity

relate to other concepts

Endocrine System

Abdominal organ

Diseases

Endocrine Diseases

Adrenal Dysfunction

Disorders of other endocrine gland

Other disorders of adrenal gland
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 Cortex Diseases

Adrenal Cortex Dysfunction

Hypoadrenalism

Adrenal Gland Hypofunction

Adrenal cortical hypofunction

Other disorders of adrenal gland

Addison’s Disease
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 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

- 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)
  - 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)
Addison's Disease: Concept

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

- **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

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

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

GER

FRE
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

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

- **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
    - Organism Function
      - Mental Process
    - Organ or Tissue Function
    - Cell Function
    - Molecular Function
      - Genetic Function
  - Pathologic Function
    - Cell or Molecular Dysfunction
    - Disease or Syndrome
    - Experimental Model of Disease
      - Mental or Behavioral Dysfunction
      - 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
  - Organism Attribute
  - Finding
    - Laboratory or Test Result
    - Sign or Symptom
  - Biologic Function
    - Physiologic Function
    - Pathologic Function
  - Fully Formed Anatomical Structure
    - contains, produces
    - conceptual part of
      - Body Substance
      - Body Space or Junction
        - location of
      - Body Location or Region
        - location of
        - conceptually part of
          - Pathologic Function
            - disrupts
              - Injury or Poisoning
                - co-occurs with
      - Body Part, Organ or Organ Component
        - part of
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
- Biologic Function
- Pathologic Function
- Disease or Syndrome
- Adrenal Cortical hypofunction
- Adrenal Cortex
- Body Part, Organ, or Organ Component
- Metathesaurus

Relationships:
- isa
- location of
SPECIALIST Lexicon
and lexical tools
SPECIALIST Lexicon

- **Content**
  - English lexicon
  - Many words from the biomedical domain
- **200,000+ 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
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
- Final result: 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
Normalization Applications

◆ Model for lexical resemblance
◆ Help find lexical variants for a term
  ● Terms that normalize the same usually share the same LUI
◆ Help find candidates to synonymy among terms
◆ Help map input terms to UMLS concepts
Indexes

- **Word index**
  - word to Metathesaurus strings
  - one word index per language

- **Normalized word index**
  - normalized word to Metathesaurus strings
  - English only

- **Normalized string index**
  - normalized term to Metathesaurus strings
  - English only
Lexical Variant Generation program

- Tool for specialists (linguists)
- Performs atomic lexical transformations
  - generating inflectional variants
  - lowercase
  - ...
- Performs sequences of atomic transformations
  - a specialized sequence of transformations provides the normalized form of a term (the \textit{norm} program)
UMLS in action

MetaMap
MetaMap  

Motivation

Term extraction
- Identifying UMLS concepts from text

Usage
- Information indexing and retrieval
- Knowledge extraction / discovery
- Semantic interpretation

Characteristics
- Linguistic approach
- Based on UMLS knowledge sources

[Aronson, AMIA, 2001]
MetaMap Methods

- Parsing
  - Shallow syntactic analysis
  - SPECIALIST lexicon
  - Xerox part-of-speech tagger

- Variant generation

- Candidate retrieval
  - Retrieve candidate terms containing at least one variant

- Candidate evaluation
  - Rank candidate terms with respect to closeness to input text (centrality, variation, coverage, and cohesiveness)
Molluscum contagiosum is a disease caused by a poxvirus of the Molluscipox virus genus that produces a benign self-limited papular eruption of multiple umbilicated cutaneous tumors.
Molluscum Contagiosum Disease
Cutaneous eruption
Multiple tumors
Cutaneous tumor
Skin
Papular eruption
Virus
Pox virus (Poxviridae)
Pathologic Function
Disease or Syndrome
Neoplastic Process
Molluscum Contagiosum
Multiple tumors
Cutaneous tumor
Location of
causes
manifestation of
Finding
Body Part, Organ, or Organ Component
Virus
Skin
Papular eruption
Cutaneous eruption
Metathesaurus
Using MetaMap MMTx

- Requires UMLS license
- Local implementation (Java-based)
- Provides
  - Stand-alone application
  - API for integrating in other applications

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
Appendix
Knowledge Source Server

Web Interface

Knowledge Source Server
Application Programming Interface
UMLSKS API basics

- Remote server at NLM
- Local application connected through

### Java RMI
- Java-based applications
- Developer’s Guide: Chapter 3
- Set of Java classes (part of the UMLSKS API download)
- Detailed Javadoc documentation online and with API download

### TCP/IP socket
- XML-based queries
- Developer’s Guide: Chapter 5
- XML schema
- Socket server
  - Host: umlsks.nlm.nih.gov
  - Port: 8042
Documentation

- User’s Guide
- Developer’s Guide
  1. Introduction
  2. Installing the UMLSKS
  3. Building UMLSKS Software Applications
  4. Using the XML Query Facility
  5. Using the UMLSKS Socket Server
- UMLS Documentation Set

This guide describes the installation of the Knowledge Source Server (UMLSKS) applications.

Audience
The audience for this guide is developers of UMLSKS applications using the UMLSKS API.

Release Notes
Please refer to the Release Bulletin for a detailed list of features, bug fixes, and known problems with this version of the UMLSKS.

How to Use This Guide
This manual contains the following chapters:

- **Chapter 1 – Introduction** describes the basic features and architecture of the UMLSKS.
- **Chapter 2 – Installing the UMLSKS** provides administrators instructions on installing and tailoring a UMLSKS installation.
- **Chapter 3 – Building UMLSKS Software Applications** describes the functions available to developers wanting to interface to the UMLSKS through another Java program.
- **Chapter 4 – Using the XML Query Facility** describes how to use the querying facility of the UMLSKS wherein users build XML queries to be executed.
- **Chapter 5 – Using the UMLSKS Socket Server** describes how to use the socket server to pass XML formatted commands or command-line type queries (e.g. ks -meta -c aids) that are to be submitted to the server with the server used to be

About the UMLS

- Home
- Overview
- Frequently Asked Questions
- Edit Views/Profile

Downloads
- UMLS Knowledge Sources
- Developer’s API

Documentation
- User’s Guide
- Developer’s Guide
- UMLS Documentation Set

Resources
- NLP & Lexical Resources
- Semantic Network Resources
- Meta-omics Resources
MetamorphoSys
What is MetamorphoSys?

◆ Tool distributed with the UMLS

◆ Multi-platform Java software

◆ The UMLS installation and customization wizard
  ● Installs Knowledge Sources to local storage
  ● Subsets and customizes a local Metathesaurus
Why use MetamorphoSys?

**Customize the Metathesaurus**

- To remove terminology that is unhelpful, or even harmful, to your needs and purposes
- To comply with terms of license agreement

**Changing Default Settings**

- To alter the preferred name
- To alter suppressibility of specific source term types
Bibliography
UMLS documentation and support

  - with links to all other UMLS information

  - with links to the User’s and Developer’s guides

- **Email address for support**  custserv@nlm.nih.gov
References

◆ UMLS as a research project


References

◆ Technical papers


◆ Comprehensive bibliography 1986-96