Johns Hopkins University
November 29, 2005

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”

“ [...] 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 is \textit{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
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
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>
<th>MedDRA</th>
<th>10036696</th>
<th>ICD-10</th>
<th>E27.1</th>
<th>SNOMED CT</th>
<th>363732003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary adrenocortical insufficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Addison's disease (disorder)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0001403</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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
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

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

Adrenal Gland Dysfunction

Addison's Disease

Secondary hypocortisolism

Adrenal Gland Hypofunction

Adrenal Cortical hypofunction

Hypoadrenalism

Other disorders of adrenal gland

Disorders of other endocrine gland

Endocrine Diseases

Abdominal organ

Diseases

relate to other concepts

Addison's disease due to autoimmunity
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
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 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

- 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
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 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

- **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>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></td>
<td></td>
</tr>
<tr>
<td>A0000003</td>
<td>Headache (source 1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A0000004</td>
<td>Headache (source 2)</td>
<td></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>
Addison's disease
Addison Disease
Addison's Disease
Addisons Disease
Disease, Addison
ADDISON'S DISEASE
Addison's disease (disorder)
Primary Adrenocortical Insufficiency
Insufficiencies, Primary Adrenocortical
Primary Adrenal Insufficiency
adrenal; insufficiency, primary
Primary Hypoadrenalism
Hypoadrenalism, Primary
Addison-Krankheit
Maladie d'Addison
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  PAR/CHD
  - Broader / Narrower than  RB/RN

- **Derived from hierarchies**
  - Siblings (children of parents)  SIB

- **Associative**
  - Other  RO

- **Various flavors of near-synonymy**
  - Similar  RL
  - Source asserted synonymy  SY
  - Possible synonymy  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)
Associative (non-isa) relationships

- Organism
  - Anatomical Structure
    - Embryonic Structure
    - Anatomical Abnormality
      - Congenital Abnormality
      - Acquired Abnormality
    - Body System
  - Finding
    - Laboratory or Test Result
    - Sign or Symptom
    - Body Substance
    - Body Space or Junction
    - Injury or Poisoning
    - Pathologic Function
    - Physiologic Function
- Biologic Function

- Organism Attribute
  - part of
  - property of

- Evaluation of
  - part of
  - conceptual part of

- Part of
  - part of
  - contains, produces
  - disrupts
  - co-occurs with
  - adjacent to
  - location of
  - conceptual part of

- Location of
  - next to
  - close to
  - near

- Conceptual part of
  - next to
  - close to
  - near

- NLM

Page 48
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
- Body Part, Organ, or Organ Component
- Adrenal Cortex
- Adrenal Cortical hypofunction

Metathesaurus

location of

isa
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 (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

Hodgkin diseases, NOS

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 *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
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
Virus
Pox virus (Poxviridae)
Skin
Papular eruption
Cutaneous eruption
Pathologic Function
Disease or Syndrome
Neoplastic Process

causes
location of
manifestation of
Using MetaMap MMTx

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

Conclusions
Integrating subdomains

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

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

UMLS Knowledge Source Server Home Page

UMLS Knowledge Source Server (UMLSKS)

About the UMLSKS
- Home
- Overview
- Frequently Asked Questions
- Edit Views/Profile

Downloads
- UMLS Knowledge Sources
- Developer’s API

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

Resources
- NLP & Lexical Resources
- Semantic Network Resources
- Metathesaurus Resources

Quick Search
- Select UMLS Release: 2004AB
- Enter search value: Addison’s disease

Advanced Searches
- Metathesaurus Advanced Search
  - Facilitates advanced searching of the UMLS Metathesaurus, including restricting vocabularies, performing batch searches, and performing XML queries, all using a command-line type interface.

- Semantic Network Browser
  - Allows browsing of the hierarchies for the Semantic Network.

What’s New
- 2004 AB Metathesaurus now available to download and searching for those that have signed the new license agreement!
- UMLSKS Version 4.3 released on August 30, 2004 for 2004 AB download access and searching.
Knowledge Source Server
Application Programming Interface
# UMLSKS API basics

- Remote server at NLM
- Local application connected through

<table>
<thead>
<tr>
<th>Java RMI</th>
<th>TCP/IP socket</th>
</tr>
</thead>
<tbody>
<tr>
<td>◆ Java-based applications</td>
<td>◆ XML-based queries</td>
</tr>
<tr>
<td>◆ Developer’s Guide: Chapter 3</td>
<td>◆ Developer’s Guide: Chapter 5</td>
</tr>
<tr>
<td>◆ Set of Java classes</td>
<td>◆ XML schema</td>
</tr>
<tr>
<td>(part of the UMLSKS API download)</td>
<td>◆ Socket server</td>
</tr>
<tr>
<td>◆ Detailed <em>Javadoc</em> documentation online and with API download</td>
<td>⚫ Host: umlsks.nlm.nih.gov</td>
</tr>
<tr>
<td></td>
<td>⚫ Port: 8042</td>
</tr>
</tbody>
</table>
This guide describes the installation and use of the UMLS Knowledge Source Server (UMLSKS) and its software 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 executed by the server with the appropriate additional options.
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
References
References

- **UMLS**
  umlsinfo.nlm.nih.gov

- **UMLS browsers**
  (free, but UMLS license required)
  - Semantic Navigator:
  - RRF browser
    (standalone application distributed with the UMLS)
References

◆ Recent overviews


References

◆ UMLS as a research project


References

◆ Technical papers
