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.»

(Integrated Academic Information Management 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)

Addison Disease | MeSH | D000224
---|---|---
Primary hypoadrenalism | MedDRA | 10036696
Primary adrenocortical insufficiency | ICD-10 | E27.1
Addison's disease (disorder) | SNOMED CT | 363732003

C0001403

Addison's disease
Endocrine system diseases

Adrenal gland diseases

Adrenal Insufficiency

Addison Disease
Endocrine disorders

Adrenal gland disorders

Adrenal cortical hypofunctions

Addison's 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
NCI Thesaurus

- Endocrine Disorder
  - Non-neoplastic endocrine disorder
  - Adrenal gland disorder
    - Non-neoplastic adrenal gland disorder
      - Adrenal gland insufficiency
        - Adrenal cortical insufficiency
          - Addison's Disease
ICD-10

Other disorders of adrenal gland

Disorders of other endocrine glands

Endocrine, nutritional and metabolic diseases

Primary adrenocortical insufficiency
Organize concepts

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

SNOMED CT
SNOMED Intl
MeSH
MedDRA
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
  - Adrenal cortical hypofunction
    - Addison’s Disease
  - Adrenal cortex diseases
  - 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 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
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
- GO
- FMA
- Anatomy
- NCBI Taxonomy
- SNOMED CT
- OMIM
- MeSH
- Other subdomains
- Model organisms

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)

Clinical repositories

Other subdomains

SNOMED CT

Genetic knowledge bases

OMIM

Biomedical literature

Addison Disease (D000224)

NCBI Taxonomy

Model organisms

UMLS C0001403

FMA

Anatomy

Genome annotations

GO

Addison's disease (363732003)

Other subdomains

SNOMED CT

Genetic knowledge bases

OMIM

Biomedical literature

Addison Disease (D000224)

NCBI Taxonomy

Model organisms

UMLS C0001403

FMA

Anatomy

Genome annotations

GO
Addison’s Disease: Concept

Addison’s Disease

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.
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

| A0066000 | Headache (MeSH) |
| A0065992 | Headache (ICD-10) |
|          | S0046854       |
| A0066007 | Headaches (MedDRA) |
| A12003304| Headaches (OMIM)  |
|          | S0046855       |
|          | L0018681       |
| A0540936 | Cephalodynia (MeSH) |
|          | S0475647       |
|          | L0380797       |
|          | C0018681       |
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, NOS
C0211735

Addison's disease
C0001403

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
  - <atom 1> mapped_to <atom 2>
  - e.g.,
    - SNOMED CT to ICD-9-CM

◆ Complex mappings
  - <atom 1> mapped_to <boolean expression>
  - 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)

Biologic Function

- Physiologic Function
  - Organ Function
    - Organ or Tissue Function
      - Mental Process
  - Cell Function
  - Molecular Function
    - Genetic Function

- Pathologic Function
  - Cell or Molecular Dysfunction
  - Disease or Syndrome
    - Mental or Behavioral Dysfunction
  - Experimental Model of Disease
    - Neoplastic Process
Associative (non-isa) relationships
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

Semantic Network

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

location of

isa

Biologic Function

Pathologic Function
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
    - transitive
    - ditransitive

  I'll treat.
  He treated the patient.
  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
- Biomedical literature
- Genome annotations
- Other subdomains
- SNOMED CT
- OMIM
- MeSH
- Model organisms
- FMA
- GO
- Anatomy
- NCBI Taxonomy
- UMLS
Trans-namespance integration

Addison's disease (363732003)

Clinical repositories

SNOMED CT

OMIM

Genetic knowledge bases

Biomedical literature

Addison Disease (D000224)

Other subdomains

Model organisms

NCBI Taxonomy

FMA

GO

Anatomy

Genome annotations

MeSH

Clinical repositories

MeSH

Biomedical literature
Other things you would need to know

◆ UMLLS license agreement

◆ MetamorphoSys

◆ UMLS Terminology Services (UTS)
  (formerly, UMLS Knowledge Source Server)
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
References: UMLS home page

◆ UMLS home page

◆ UMLS documentation
  - Formerly know as the “Green Book”
  - Now online documentation

◆ UMLS online tutorials
References

◆ Recent overviews


UMLS as a research project

References

◆ Technical papers
