The Unified Medical Language System
What is it and how to use it?

Part I
What is the UMLS?

(1) Introduction

What does UMLS stand for?
- Unified
- Medical
- Language
- System

Outline
- What is the UMLS?
  - Introduction
  - Overview through an example
  - The three UMLS Knowledge Sources
- How to use the UMLS?
  - Obtaining a license
  - Remote access
  - Local installation and customization
  - A UMLS-based algorithm
  - Benefits and limitations

Part I: What is the UMLS?
- Introduction
- Overview through an example
- The three UMLS Knowledge Sources
  - UMLS Metathesaurus
  - UMLS Semantic Network
  - SPECIALIST Lexicon and lexical tools
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)
  - RRF browser (browsing subsets)

Part I

What is the UMLS?

(2) 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
  - Melasma addisonii
  - 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>
</tr>
</thead>
<tbody>
<tr>
<td>Addison Disease</td>
<td>MeSH</td>
<td>D000224</td>
</tr>
<tr>
<td>Primary hypoadrenalism</td>
<td>MeDRA</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>
<tr>
<td>C0001403</td>
<td>Addison's disease</td>
<td></td>
</tr>
</tbody>
</table>
Disorder of endocrine system
Disorder of adrenal gland
Adrenal gland hypofunction
Disorder of adrenal cortex
Adrenal cortical hypofunction
Addison's Disease

SNOMED CT (UMLS view)

Endocrine, nutritional, and metabolic diseases
Disorders of other endocrine glands
Other disorders of adrenal gland
Primary adrenal cortical insufficiency

ICD-10

SNOMED Intl
MeSH
MedDRA

Endocrine system diseases
Adrenal gland diseases
Adrenal gland hypofunction
Adrenal cortical diseases
Adrenal cortical hypofunction
Addison's Disease

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

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

UMLS view

Endocrine system diseases
Non-neoplastic endocrine disorder
Adrenal gland diseases
Adrenal gland hypofunction
Adrenal cortical diseases
Adrenal cortical hypofunction
Addison's Disease

Organize concepts
UMLS Tutorial - O. Bodenreider (NLM)

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

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

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

UMLS Summary
- Synonymous terms clustered into concepts
- Unique identifier
- Finer granularity
- Broader scope
- Additional hierarchical relationships
- Semantic categorization

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

- 143 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 5.9M names
  - 1.4M concepts
  - 8M relations
- 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)

Integrating subdomains

Trans-namespace integration
Addison's Disease: Concept

- ADRENAL INSUFFICIENCY (ADDISON'S DISEASE)
- ADRENOCORTICAL INSUFFICIENCY, PRIMARY FAILURE
- Hypoadrenalism, Primary
- Asthenia pigmentosa
- Bronzed disease
- Primary adrenocortical insufficiency
- Addison's, disease
- Maladie d'Addison - French
- Addison-Krankheit - German
- Morbo di Addison - Italian
- Doença de Addison - Portuguese
- АДДИСОНОВА БОЛЕЗНЬ - Russian
- アジソン病 - Japanese

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 increased circulating levels of ACTH precursor hormone which stimulates melanocytes.

Beyond concepts - Descriptor level

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

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
### Metathesaurus Relationships
- 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

### Symbolic relationships

#### Type
- Hierarchical
  - Parent / Child
  - Broader / Narrower than
  - Derived from hierarchies (children of parents)
  - Siblings (children of parents)
- Associative
  - Other
- Various flavors of near-synonymy
  - Similar
  - Source asserted synonymy
  - Possible synonymy

#### Attribute
- Hierarchical
  - isa (is-a-kind-of)
  - part-of
- Associative
  - location-of
  - caused-by
  - treats
  - …
- Cross-references (mapping)

### 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)
  - …
Semantic Types

- Semantic types (135)
  - tree structure
  - 2 major hierarchies
    - Entity
      - Physical Object
      - Conceptual Entity
    - Event
      - Activity
      - Phenomenon or Process

"Biologic Function" hierarchy (isa)

UMLS 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

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 can inherit semantics

SPECIALIST Lexicon

- Content
  - English lexicon
  - Many words from the biomedical domain
- 360,000 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's 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

- Hodgkin's diseases, NOS
  - Remove genitive: Hodgkin diseases, NOS
  - Remove stop words: Hodgkin diseases
  - Lowercase: hodgkin diseases
  - Strip punctuation: hodgkin disease
  - Uninflect: disease hodgkin

Normalization: Example

- Hodgkin Disease
  - Hodgkin's Disease
  - disease, Hodgkin
  - Hodgkin's disease
  - Hodgkin's disease NOS

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)

Outline

- Part II: How to use the UMLS?
  - Obtaining a license
  - Remote access
    - Knowledge Source Server (UMLSKS)
    - UMLSKS Application programming interface (API)
  - Local installation and customization (MetamorphoSys)
  - A UMLS-based algorithm: Restrict to MeSH
  - Benefits and limitations

First step License agreement

- Online Web-based license:
- Read license
- Read appendix 1 and 2
- Print a copy for your records
- Complete the Web form
- Verify:
  - receive e-mail from NLM; go to Web site within 72 hours and enter first and last name
  - NLM official will countersign (turn-around time of a few days)
  - Receive 2nd e-mail from NLM with new license number

Part II

How to use the UMLS?

(1) Obtaining a license

UMLS Tutorial - O. Bodenreider (NLM)

1. The UMLS is a registry of concepts that is maintained and updated by the National Library of Medicine (NLM). The UMLS contains a vocabulary of terms used in the biomedical sciences, including diseases, symptoms, medications, and procedures. It is designed to facilitate the exchange of information between different databases and to support the integration of heterogeneous data sources.

2. UMLS is a public database containing a vocabulary of concepts used in biomedical and health-related research. It is maintained by NLM and is available free of charge. The UMLS contains several components, including the Metathesaurus, which is a catalog of biomedical concepts, and the Semantic Network, which provides metadata about the relationships between concepts.

3. The UMLS Metathesaurus contains more than 10 million concept names, each of which is linked to one or more concept descriptions. The Metathesaurus is updated quarterly, and new concepts are added to it as they are discovered or identified. It is available for download in various formats, including XML and NLM's proprietary UML format.

4. A UMLS concept is a term or phrase that represents a specific concept in the biomedical domain. It is assigned a unique identifier, known as a UMLS concept identifier (CUI), which is used to identify the concept in the UMLS. Concepts can be linked to other concepts through relationships such as synonyms, definitions, and semantic types.

5. The UMLS Semantic Network (SNOMED CT) is a comprehensive, clinical terminology that covers a wide range of concepts used in healthcare and biomedical research. It is maintained by the American College of Medical Informatics (ACMI) and is supported by a large community of users.

6. SNOMED CT is a comprehensive, clinical terminology that covers a wide range of concepts used in healthcare and biomedical research. It is maintained by the American College of Medical Informatics (ACMI) and is supported by a large community of users.

7. The UMLS is a registry of concepts that is maintained and updated by the National Library of Medicine (NLM). The UMLS contains a vocabulary of terms used in the biomedical sciences, including diseases, symptoms, medications, and procedures. It is designed to facilitate the exchange of information between different databases and to support the integration of heterogeneous data sources.

8. The UMLS is a registry of concepts that is maintained and updated by the National Library of Medicine (NLM). The UMLS contains a vocabulary of terms used in the biomedical sciences, including diseases, symptoms, medications, and procedures. It is designed to facilitate the exchange of information between different databases and to support the integration of heterogeneous data sources.

9. The UMLS is a registry of concepts that is maintained and updated by the National Library of Medicine (NLM). The UMLS contains a vocabulary of terms used in the biomedical sciences, including diseases, symptoms, medications, and procedures. It is designed to facilitate the exchange of information between different databases and to support the integration of heterogeneous data sources.

10. The UMLS is a registry of concepts that is maintained and updated by the National Library of Medicine (NLM). The UMLS contains a vocabulary of terms used in the biomedical sciences, including diseases, symptoms, medications, and procedures. It is designed to facilitate the exchange of information between different databases and to support the integration of heterogeneous data sources.
License Restriction Levels 0-4  

- Level 0  (79.3%)  
  - unrestricted  
- Level 1  (5.0%)  
  - negotiate to translate  
- Level 2  (0.5%)  
  - negotiate to use in health data creation  
- Level 3  (25.4%)  
  - negotiate to use in production  
  - explicitly prohibited to provide Internet access  
- Level 4  (21.8%)  
  - SNOMED CT (unrestricted in member countries)  

There may be additional restrictions, or separate license fees, associated with usage of specific vocabularies. Read the UMLS License, including the Appendix!

Part II
How to use the UMLS?

(2) Remote access

Remote Access

- UMLS Knowledge Source Server:  
- Web search interface
- Application Programming Interface (API)
- Coming soon: web services

Knowledge Source Server
Web search interface
UMLSKS Web search interface

- Logging in
- Basic searching
- Advanced searching

UMLSKS Web search interface log in

- Returning users log in
- New users create account

UMLS Knowledge Source Server Home Page

- Tabs across top access basic searching of 3 Knowledge Sources
- Advanced searching options on right-hand side

UMLS Knowledge Source Server Home Page

Metathesaurus Basic Search

Addison’s disease

- UMLS Release
- Search Term
- UMLS Knowledge Source

Concept Report

Addison’s disease

- Concept Name /CUI
- Semantic Type(s)
- Definition(s)
- Synonyms
"Display" shows results for selected options

"Display All" shows results for all available options

- Specify:
  - UMLS Release
  - Search term

- Algorithm:
  - Search Normalized String
  - Search Normalized Word
  - Suggest Spelling

- Concept Name/CUI
- Semantic Type(s)
- Definition(s)
- Synonyms, including foreign languages
- Relations (broader, narrower, etc.)
- Co-occurrence data

Synonyms

Sources

Hierarchies
Concept Report Display All (continued)

Relations

Co-occurrence data

Metathesaurus Advanced Search Options

- Focused Search
- Raw Relational Records

Metathesaurus Advanced Search Feature
Focused Search

- UMLS Release
- Search Term
- Source Vocabularies
- String Criteria
  - Exact Match
  - Normalized string & word
  - Word
  - Truncation (left/right)
  - Approximate Match
- Language

Restricted Source Concept Report
Addison’s Disease

- UMLS Release: 2004AB
- Search Term: addison’s disease
- Source Vocabulary: SNOMED CT
- String Criteria: Normalized string
- Language: English

Addison’s disease in SNOMED CT
Preferred Term and Code

- TTY: Term Type
- ID: Source Code Descriptor
Metathesaurus Advanced Search Feature
Relational Record Request

- UMLS Release
- Search Term
- UMLS Relational Table

Relational Records MRCONSO.RRF

- Semantic Network Searching
  - Select Tab along top
  - Quick search
  - Advanced Search on right-hand side

- Semantic Network Search
  - Enter search string
  - Select semantic type
  - Select semantic relation

Semantic Type Clinical Drug

- Browse ST hierarchy
- View Concepts with ST
- View Relations valid for the ST
- View Raw Relational Records

Show Relations Between Types

- Validates whether a selected Semantic Relationship (SR) holds between two selected Semantic Types (ST)
SPECIALIST Lexicon Searching

- Select Tab along top
- Quick search

SPECIALIST Lexicon Search

UMLS Resources

- NLP & Lexical Resources
  - MetaMap Transfer (MMTs)
  - Word Sense Disambiguation (WSD) Test Collection
- Semantic Network
  - Semantic Navigator
  - Semantic Groups
- Metathesaurus
  - String Properties

The “new” UMLSKS (coming soon)
**UMLS Semantic Navigator**
- Web-based
- Concept- and relation-centric
- Displays contexts graphically
- Displays all relations simultaneously
- Excludes hierarchical cycles in the UMLS graph
- Search
  - By CUI
  - By word

**RRF Browser**
- Distributed with the UMLS
  - Along with MetamorphoSys
- Standalone
- Can browse particular subsets of the Metathesaurus
- Search
  - By code
  - By CUI
  - By word

**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

Developer’s Guide

Documentation Java API

Documentation Javadoc

Sample XML query (1) Current version

```xml
<?xml version="1.0"?>
<getCurrentUMLSVersion version="1.0"/>
<br>2007AB</CurrentUMLSYear>
```

Sample XML query (2) Concepts by string

```xml
<?xml version="1.0"?>
<findCUI version="1.0">
<conceptName>appendicectomy</conceptName>
<language>ENG</language>
<exact/>
<noSuppressibles/>
</findCUI>
```
Sample XML query (3) Concepts properties

```xml
<?xml version="1.0"?>
<getSemanticType version="1.0">
  <cui>C0033572</cui>
</getSemanticType>
```

Sample XML query (4) Relationships

```xml
<?xml version="1.0"?>
<getRelations version="1.0">
  <cui>C0033572</cui>
  <rel>RO</rel>
</getRelations>
```

Sample XML query (5) All semantic type IDs

```xml
<?xml version="1.0"?>
<listSemTypeIds version="1.0">
</listSemTypeIds>
```

Performing XML queries from UMLSKS

```xml
<?xml version="1.0"?>
<getRelations version="1.0">
  <cui>C0033572</cui>
  <rel>RO</rel>
</getRelations>
```

Part II
How to use the UMLS?

(3) Installing the UMLS locally and Customizing the Metathesaurus using 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

Using MetamorphoSys

- Simple to use
- Screens and tabs lead you through process
- Installs NLM data format files to local storage

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

Why use MetamorphoSys?

**Changing Default Settings**
- To alter the preferred name
- To alter suppressibility of specific source term types

Customization is Critical

- Requires a clear understanding of:
  - Characteristics of source vocabularies
  - License arrangements
  - User’s functional requirements
  - User’s purpose and perspective
  - Technical expertise

... and requires a multidisciplinary technical team

Machine Requirements

- A fast CPU – 1 GHz or higher
- 1 GB RAM recommended (512 MB min.)
- 6x (or better) DVD drive
- 22 GB minimum free disk space
- Runs on Sun Solaris 8 & 9, Windows XP, NT, and 2000, Linux, and Mac
- 1-10 hours run time on platforms tested
Download from UMLSKS ...

- High speed Internet connection required
- Read the README file for the release

...or DVD?

- Order at: umls_support@nlm.nih.gov
- Include your license number
- Run MetamorphoSys from DVD
  - Windows
    - Autorun; or go to root directory and click on “windows_mmsys.bat”
  - Linux, Solaris, Macintosh
    - open a terminal window, change to the root directory and type appropriate command: ./linux_mmsys.sh, ./solaris_mmsys.sh, ./macintosh_mmsys.sh

Welcome Screen

Install UMLS

UMLS License Notice

Installation progress monitor
Select a default subset

Level 0 → no separate additional license agreements

Level 0 + SNOMEDCT → Users from non-IHTSDO member countries must have separate license agreements

Input Options Tab

Output Options Tab

Source List Tab

Highlighted rows are excluded from the subset.

Precedence Tab

• Ranks names by types of terms within sources
• Highest ranking name determines the Preferred Name
Suppressibility Tab

Highlighted source term types will be marked as suppressible.

File menu

Edit menu

Options menu

Reset menu

Done – Begin Subset

- Returns all filters to default selections
- Default selections in “mmsys.prop.default file” in config folder
- mmsys.prop.default contains properties in last run
How MetamorphoSys Works

- Removes all information from relational files in excluded vocabularies
  - atoms, strings, relationships, attributes, mappings, etc.
- Applies additional options selected by user
  - such as adding source term suppressibility or altering precedence
- Produces a full set of Metathesaurus files
  - relational files with customized data
  - reflecting other user criteria

MetamorphoSys log

Output directory contents

Part II
How to use the UMLS?

(4) A UMLS-based algorithm
Indexing Initiative  

- For noun phrases extracted from medical texts, map to UMLS concepts
- Then, select from the MeSH vocabulary the concepts that are the most closely related to the original concepts

Restrict to MeSH  

- Based on the principle of semantic locality
- Use different components of the UMLS
- 4 techniques of increasing aggressiveness
  - Use Synonymy MRCONSO
  - Use Associated expressions (ATXs) MRATX + MRREL
  - Explore the Ancestors MRREL + SN
  - Explore the Other related concepts MRREL + SN

Restrict to MeSH  Synonymy  

- Term mapped to Source concept
- For this concept, is there a synonym term that comes from MeSH? (MRCONSO)

Restrict to MeSH  Assoc. expressions  

- If not,
- Is there an associated expression (ATX) that describes this concept using a combination of MeSH descriptors? (MRATX/MRMAP + MRREL)

Restrict to MeSH  Ancestors  

- If not, let us build the graph of the ancestors of this concept
  - using parents and broader concepts (MRREL)
  - all the way to the top
  - excluding ancestors whose semantic types are not compatible with those of the source concept (MRSTY)
- From the graph, select the concepts that come from MeSH (MRCONSO)
- Remove those that are ancestors of another concept coming from MeSH

Restrict to MeSH  Other related concepts  

- If not, explore the other related concepts (MRREL) whose semantic types are compatible with those of the source concept (MRSTY)
- From those, select the concepts that come from MeSH (MRCONSO)
Restrict to MeSH Example

Vein of neck, NOS

There is a MeSH term in the synonyms of SC.

SC is described by a combination of MeSH terms (ATX).

The ancestors of SC contain MeSH terms.

MeSH terms from non-hierarchically related concepts.

Vein + Neck

Restrict to MeSH Example

Restrict to MeSH Example

Vein of neck, NOS

Vein of head and neck, NOS

Neck

Blood Vessels

Vascular structure

Restrict to MeSH Example

Restrict to MeSH Example

175

176

177

178

Restrict to MeSH Quantitative results

86% of UMLS concepts mapped to MeSH (2007)

Other related concepts

Synonymy

12%

23%

11%

54%

Built-in mappings

Restrict to MeSH Qualitative results

Qualitative evaluation

1,036 concepts extracted from 200 MEDLINE citations

Manual review of every mapping or failure

61% Relevant

Subtotal Gastrectomy ➔ Gastrectomy

Encephalopathy, NOS ➔ Brain Diseases

28% More or less relevant

Vitamin A measurement ➔ Laboratory Procedure

Swelling, NOS ➔ Symptoms

11% Non relevant

Restrict to MeSH Qualitative results

Restrict to MeSH Qualitative results

Qualitative evaluation

1,036 concepts extracted from 200 MEDLINE citations

Manual review of every mapping or failure

61% Relevant

Subtotal Gastrectomy ➔ Gastrectomy

Encephalopathy, NOS ➔ Brain Diseases

28% More or less relevant

Vitamin A measurement ➔ Laboratory Procedure

Swelling, NOS ➔ Symptoms

11% Non relevant

Restrict to MeSH Qualitative results

Restrict to MeSH Qualitative results

Qualitative evaluation

1,036 concepts extracted from 200 MEDLINE citations

Manual review of every mapping or failure

61% Relevant

Subtotal Gastrectomy ➔ Gastrectomy

Encephalopathy, NOS ➔ Brain Diseases

28% More or less relevant

Vitamin A measurement ➔ Laboratory Procedure

Swelling, NOS ➔ Symptoms

11% Non relevant

Restrict to MeSH Qualitative results

Restrict to MeSH Qualitative results

Qualitative evaluation

1,036 concepts extracted from 200 MEDLINE citations

Manual review of every mapping or failure

61% Relevant

Subtotal Gastrectomy ➔ Gastrectomy

Encephalopathy, NOS ➔ Brain Diseases

28% More or less relevant

Vitamin A measurement ➔ Laboratory Procedure

Swelling, NOS ➔ Symptoms

11% Non relevant

Part II
How to use the UMLS?

(5) Benefits and Limitations

Benefits
UMLS compared to individual vocabularies

- Broader scope
- Extended coverage
- Finer granularity
- Unique identifier
- Synonymous terms clustered into concepts
- Additional synonyms
- Additional hierarchical relationships
- Semantic categorization

Direct benefits

- Concept categorization
- Information retrieval
  - Synonyms
  - Cross-language features
- Information extraction
  - MetaMap
  - Normalization
- Information visualization
  - Knowledge Source Server
  - Semantic Navigator
  - RRF browser

UMLS as an enabling resource

- Examples
  - Mapping across vocabularies
  - Semantics of statistical associations
  - Redundancy in hierarchical relations

Limitations

- Structural inconsistency
  - Cycles in the graph of hierarchical relations
- Semantic inconsistency
  - Between Metathesaurus and Semantic Network
- Underspecified relationships
- Missing relations
  - Synonymy
  - Hierarchical relations (missing or underspecified)

Structural inconsistency  From trees to graph

- Multiple tree structures combined into a graph structure
- Expected: Directed acyclic graph (DAG)
Structural inconsistency  Cycles in the UMLS graph

- Anti-infective Agents
  - Disinfectants and Cleaners
    - Disinfectants
      - Germicidal soap

Structural inconsistency  Issues

- Theoretical
  - Violate the antisymmetry property of partial ordering relations
- Practical
  - Loops in graph traversal
  - Impossible to perform transitive reduction

[Badreodeer, AMA 2001]

Acyclicity

- "back edge" from a child concept to a parent concept

- Reflexive 13,000
- Direct 1800
- Indirect 120

Semantic inconsistency  A two-level structure

Semantic inconsistency  A limited study

- 6894 interconcept relationships
  - among the 3764 concepts in the semantic neighborhood of "Heart"
  - ICR and SNR not compatible
  - ICR not specified and SNR compatible and multiple
  - ICR not specified and SNR compatible and unique

Underspecified relationships

- Relationship “attribute” not always present
- Relations used to create hierarchies vs. hierarchial relations

Missing relations Example

- Missing relations Example

Missing relations A limited study

- 28,851 pairs of terms
  - Original SNOMED term
  - Demodified term (found in UMLS)
  - Corresponding relationship in the Metathesaurus
    - Hierarchical in 50% of the cases
    - « Sibling » in 25% of the cases
    - Missing in 25% of the cases

[Bodenreider & al., TIA, 2001]

Compensation mechanisms

- Examples
  - Removing cycles from hierarchical relations
  - Using redundancy (number of sources asserting the relation)
  - Using terminological knowledge (e.g., NEC)
  - Lexically-suggested hyponymic relations
    - Properties of adjectival modification

More limitations

- Semantics of hierarchical relations
- Some missing / wrong relations are hard to detect
- Some relations are present but hard to find
Immune system diseases
Other disorders of adrenal gland
Disorders of other endocrine glands
Endocrine system diseases
Adrenal gland diseases
Adrenal cortex diseases
Adrenal gland hypofunction
Addison's disease
Addison's disease due to autoimmunity
Addison's disease due to autoimmunity
Tuberculous Addison's disease
Autoimmune Diseases
Addison's disease
Addison's disease due to autoimmunity
Tuberculous Addison's disease

Semantics of hierarchical links

Autoimmune Diseases
Addison's disease
Addison's disease due to autoimmunity
Tuberculous Addison's disease
Addison's disease due to autoimmunity

Relations Missing and difficult to detect
chronic hypertensive renal failure
hypertensive renal failure
chronic hypertensive renal failure

Relations Existing but difficult to find
carrier protein
non-binding protein
ferritin
iron ion transport
"transport" relationship
reified "transport" relationship

How to address these limitations?
- Description logics
- Natural Language Processing (semantic interpretation of the terms)
- Comparing knowledge sources (alignment, inference)

Summary
UMLS Summary

- UMLS = 3 Knowledge Sources
  - Metathesaurus
  - Semantic Network
  - SPECIALIST Lexicon and Lexical Tools
- MetamorphoSys
  - installs
  - customizes
- UMLSKS
  - remote access
  - resources and documentation

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

Bibliography

References: UMLS home page
- UMLS home page
- UMLS documentation
  - Formerly known as the “Green Book”
  - Now online documentation

References

- Short presentation
- UMLS as a research project

References

- Technical papers
- Comprehensive bibliography 1986-96
Documentation and Support

UMLS documentation and support

- UMLS homepage: links to various UMLS resources
- UMLS home page: links to the User’s and Developer’s guides
- UMLS mailing list: UMLSUSERS-L@LIST.NIH.GOV

Email address for support:
- custserv@nlm.nih.gov

Appendix

UMLS files in Rich Release Format

Appendix - Metathesaurus relational files (RRF)

MRCONSO (sample rows 1..5)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LAT</th>
<th>SCI</th>
<th>SUI</th>
<th>AUI</th>
<th>SAUI</th>
<th>SCUI</th>
<th>SDUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MRCONSO (sample rows 6..10)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LAT</th>
<th>SCI</th>
<th>SUI</th>
<th>AUI</th>
<th>SAUI</th>
<th>SCUI</th>
<th>SDUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MRCONSO (sample rows 11-13)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LAT</th>
<th>SCI</th>
<th>SUI</th>
<th>AUI</th>
<th>SAUI</th>
<th>SCUI</th>
<th>SDUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix - Metathesaurus relational files (RRF)

MRCONSO (sample rows 11-13)

<table>
<thead>
<tr>
<th>CUI</th>
<th>LAT</th>
<th>SCI</th>
<th>SUI</th>
<th>AUI</th>
<th>SAUI</th>
<th>SCUI</th>
<th>SDUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D000224</td>
<td>ENG</td>
<td>10070140</td>
<td>1</td>
<td>363732003485624014A2922421NS0354372PFL0001403PENGC0001403</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix - Metathesaurus relational files (RRF)
MRHIER (sample rows) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>AU1</th>
<th>CXN</th>
<th>PAST</th>
<th>SAB</th>
<th>RELA</th>
</tr>
</thead>
<tbody>
<tr>
<td>C000400</td>
<td>A0019740</td>
<td>1</td>
<td>A0020270</td>
<td>MHI</td>
<td></td>
</tr>
<tr>
<td>C000400</td>
<td>A0019740</td>
<td>2</td>
<td>A0020270</td>
<td>MHI</td>
<td></td>
</tr>
<tr>
<td>C000400</td>
<td>A0019740</td>
<td>3</td>
<td>A1998776</td>
<td>PSY</td>
<td>number of classes</td>
</tr>
<tr>
<td>C000400</td>
<td>A2922421</td>
<td>1</td>
<td>A2922421</td>
<td>MHI</td>
<td></td>
</tr>
<tr>
<td>C000400</td>
<td>A2922421</td>
<td>2</td>
<td>A2922421</td>
<td>MHI</td>
<td></td>
</tr>
</tbody>
</table>

MRREL (sample rows) (2004AB)

<table>
<thead>
<tr>
<th>CUI</th>
<th>AU1</th>
<th>STYPE1</th>
<th>REL</th>
<th>CU2</th>
<th>AU2</th>
<th>STYPE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>C000400</td>
<td>A0019740</td>
<td>A1988358</td>
<td>SCUI</td>
<td>C000400</td>
<td>A2922421</td>
<td>SCUI</td>
</tr>
<tr>
<td>C000400</td>
<td>A0019740</td>
<td>A2922421</td>
<td>SCUI</td>
<td>C000400</td>
<td>A2922421</td>
<td>SCUI</td>
</tr>
</tbody>
</table>

MRDEF (2004AB)

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.

MRSAT (sample rows) (2004AB)

<table>
<thead>
<tr>
<th>CUU</th>
<th>LUI</th>
<th>SUL</th>
<th>METAUI</th>
<th>STYPE</th>
<th>CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C000400</td>
<td>A0019740</td>
<td>A1988358</td>
<td>SCUI</td>
<td>A1988358</td>
<td>A2922421</td>
</tr>
<tr>
<td>C000400</td>
<td>A0019740</td>
<td>A2922421</td>
<td>SCUI</td>
<td>A2922421</td>
<td>A2922421</td>
</tr>
</tbody>
</table>

MRHIST (sample rows) (2004AB)

<table>
<thead>
<tr>
<th>CUU</th>
<th>TUI</th>
<th>SYN</th>
<th>STY</th>
<th>ATUI</th>
<th>CVF</th>
</tr>
</thead>
<tbody>
<tr>
<td>C000400</td>
<td>20020731</td>
<td>SNOMEDCT</td>
<td>363732003</td>
<td>C0001403</td>
<td>4321</td>
</tr>
<tr>
<td>C000400</td>
<td>20020731</td>
<td>SNOMEDCT</td>
<td>1490869013</td>
<td>C0001403</td>
<td>4321</td>
</tr>
</tbody>
</table>

MRSTY (2004AB)

<table>
<thead>
<tr>
<th>CUU</th>
<th>TUI</th>
<th>SYN</th>
<th>STY</th>
<th>ATUI</th>
<th>CVF</th>
</tr>
</thead>
<tbody>
<tr>
<td>C000400</td>
<td>20020731</td>
<td>SNOMEDCT</td>
<td>363732003</td>
<td>C0001403</td>
<td>4321</td>
</tr>
<tr>
<td>C000400</td>
<td>20020731</td>
<td>SNOMEDCT</td>
<td>1490869013</td>
<td>C0001403</td>
<td>4321</td>
</tr>
</tbody>
</table>

Appendix - Metathesaurus relational files (RRF) 217

Appendix - Metathesaurus relational files (RRF) 218

Appendix - Metathesaurus relational files (RRF) 219

Appendix - Metathesaurus relational files (RRF) 220

Appendix - Metathesaurus relational files (RRF) 221

Appendix - Metathesaurus relational files (RRF) 222