Medinfo 2007
Tutorial S009
August 19, 2007

MEDINFO 2007
BRISBANE AUSTRALIA

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

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

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

2

Part I

What is the UMLS?

Outline

- 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

4

Part I
What is the UMLS?

(1) Introduction

What does UMLS stand for?

- Unified
- Medical
- Language
- System

UMLS®
Unified Medical Language System®
UMLS Metathesaurus®

6

Motivation

- Started in 1986
- National Library of Medicine
- "Long-term R&D project"
- Complementary to IAIMS (Integrated Academic Information Management Systems)

«[…] 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.»

NLM 7

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)

The UMLS is *not* an end-user application

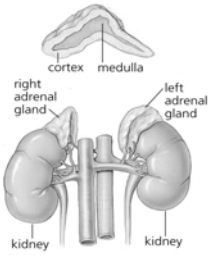
NLM 8

Part I
What is the UMLS?

(2) Overview through an example

Addison's disease

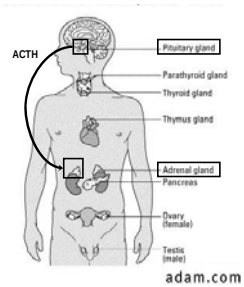
- Addison's disease is a rare endocrine disorder
- Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol
- For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism



NLM 10

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



NLM 11

Addison's disease: Symptoms

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

NLM 12

AD in medical vocabularies

- Synonyms: different terms
  - Addisonian syndrome — eponym
  - Bronzed disease — symptoms
  - Melasma addisonii — symptoms
  - Asthenia pigmentosa — symptoms
  - Primary adrenal deficiency — clinical variants
  - Primary adrenal insufficiency — clinical variants
  - Primary adrenocortical insufficiency — clinical variants
  - Chronic adrenocortical insufficiency — clinical variants
- Contexts: different hierarchies

NLM 13

Organize terms

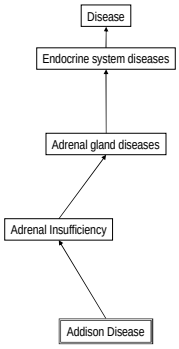
- Synonymous terms clustered into a concept
- Preferred term
- Unique identifier (CUI)

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

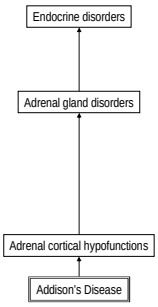
C0001403 → Addison's disease

NLM 14

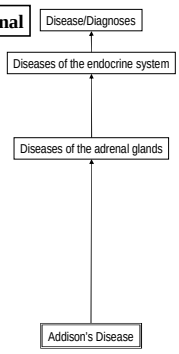
MeSH



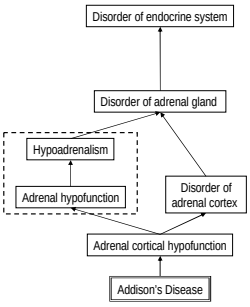
MedDRA



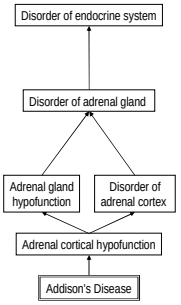
SNOMED International



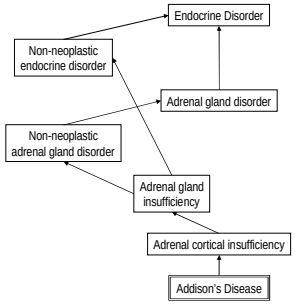
SNOMED CT (native)



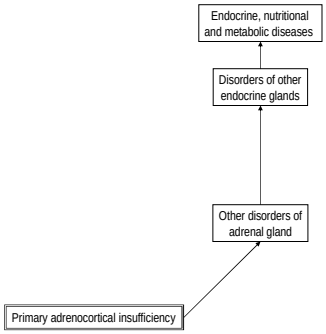
SNOMED CT (UMLS view)



NCI Thesaurus

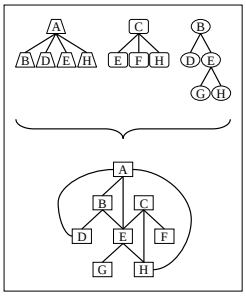


ICD-10



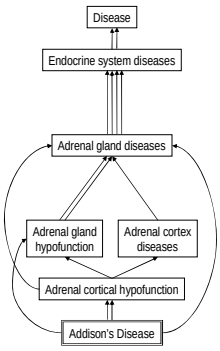
Organize concepts

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



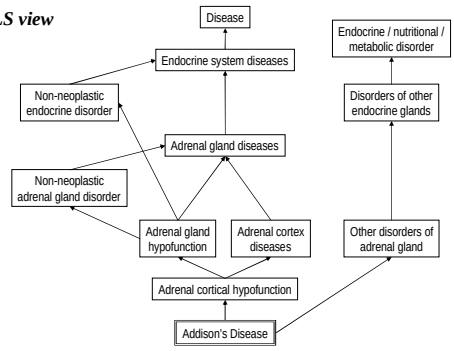
22

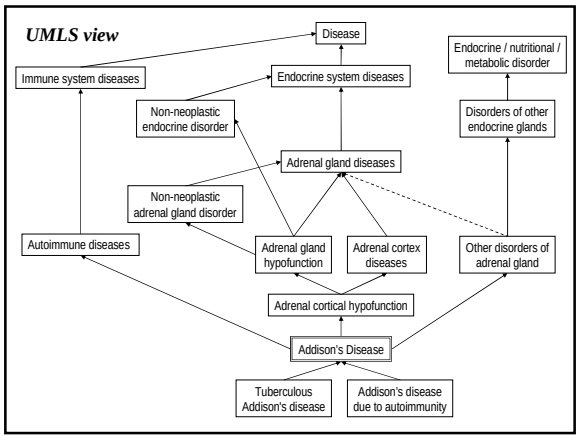
organize concepts



SNOMED CT
SNOMED Intl
MeSH
MedDRA

UMLS view



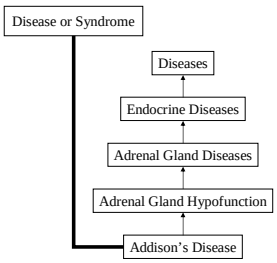


Relate to other concepts

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

Categorize concepts

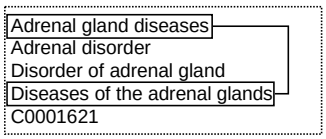
- High-level categories (semantic types)
- Assigned by the Metathesaurus editors
- Independently of the hierarchies in which these concepts are located



How do they do that?

- Lexical knowledge
- Semantic pre-processing
- UMLS editors

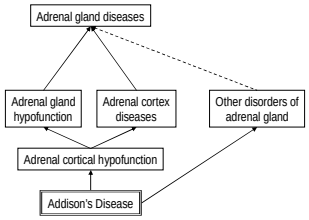
Lexical knowledge



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



31

UMLS Summary

- Synonymous terms clustered into concepts
- Unique identifier

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

32

Part I
What is the UMLS?

(3) UMLS Knowledge Sources

Unified Medical Language System

- SPECIALIST Lexicon
  - 360,000 lexical items
  - Part of speech and variant information
- Metathesaurus
  - 6M names from over 100 terminologies
  - 1.5M concepts
  - 8M relations
- Semantic Network
  - 135 high-level categories
  - 7000 relations among them

Lexical resources
Terminological resources
Ontological resources

34

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

36

Source Vocabularies (2007AB)

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

37

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)

38

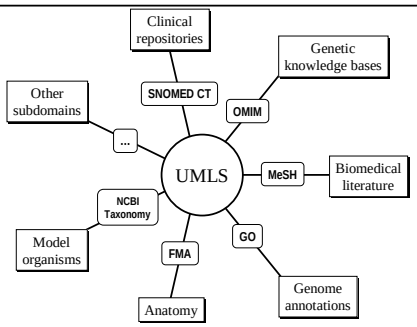
Biomedical terminologies (cont'd)

- 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)

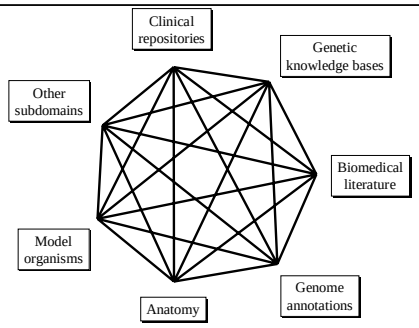
39

Integrating subdomains



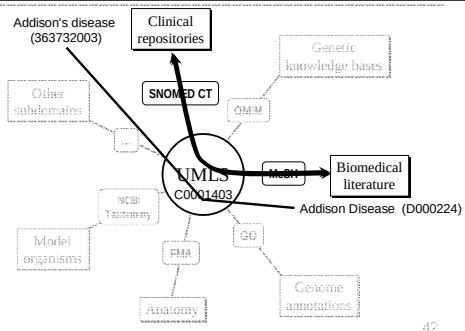
40

Integrating subdomains



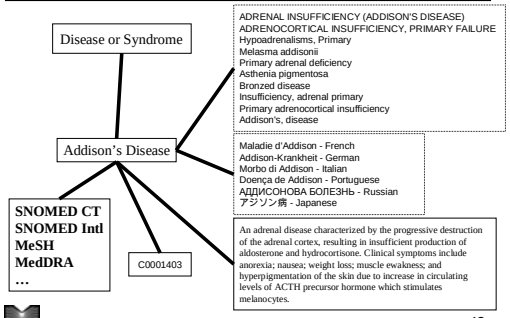
41

Trans-namespace integration



42

Addison's Disease: Concept



Metathesaurus Concepts (2007AB)

- Concept (~ 1.4M) CUI
  - Set of synonymous concept names
- Term (~ 5.3 M) LUI
  - Set of normalized names
- String (~ 5.9M) SUI
  - Distinct concept name
- Atom (~ 7.2M) 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

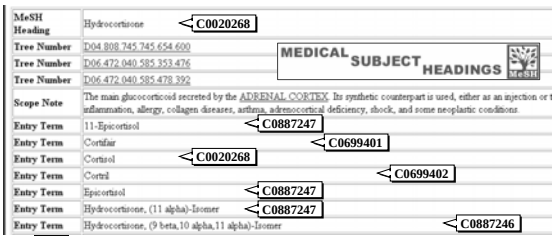
Cluster of synonymous terms

Concept C0001403

Term L0001403	S0354372 *Addison's disease* S0010794 Addison's Disease S0010792 Addison Disease S0010796 Addisons Disease S0033587 Disease, Addison S0469271 Addison's disease, NOS	[…]
Term L0494940	S5907336 *Primary Adrenocortical Insufficiency* S5901878 Insufficiencies, Primary Adrenocortical	
Term L0494851	S5907334 *Primary Adrenal Insufficiency* S5924573 Adrenal Insufficiency, Primary	[…]
Term L0585243	S5907343 *Primary Hypoadrenalism* S0718109 Primary hypoadrenalism	[…]
Term L3541031	S4115514 *primary; hypoadrenocorticism* S4090095 hypoadrenocorticism; primary	[…]
Term L1229627	S1471573 *Addison-Krankheit* GER	
Term L5345155	S6107160 *Maladie d'Addison* FRE	[…]

Beyond concepts Descriptor level

- In some vocabularies, the unit of information is an aggregate of concepts
  - e.g., descriptors in MeSH, for indexing purposes



Beyond concepts Descriptor level

A0070119 Hydrocortisone (MeSH)
A0043102 Cortisol (MeSH)
C0020268

A0066000 Epicortisol (MeSH)
A0016625 11-Epicortisol (MeSH)
A0055118 Hydrocortisone, (11 alpha)-Isomer (MeSH)
C0887247

A7801724 Hydrocortisone, (9 beta,10 alpha,11 alpha)-Isomer (MeSH)
C0887246

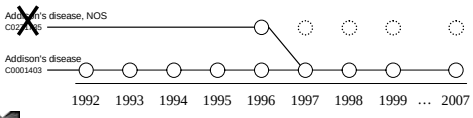
A7757592 Cortifair (MeSH)
C0699401

A7757595 Cortril (MeSH)
C0699402

D006854

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

49

Symbolic relations

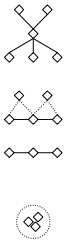
- Relation
  - Pair of "atom" identifiers
  - Type
  - Attribute (if any)
  - List of sources (for type and attribute)
- Semantics of the relationship: defined by its type [and attribute]

Source transparency: the information is recorded at the "atom" level

50

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`



51

Symbolic relationships Attribute

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

52

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)

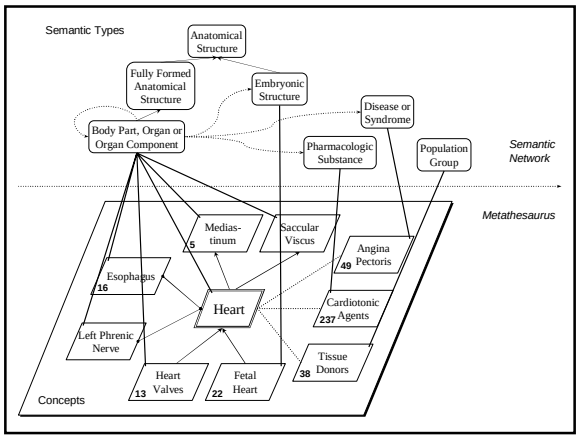
NB: partially redundant with relations in MRREL

53

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)
  - …

54



UMLS Semantic Network

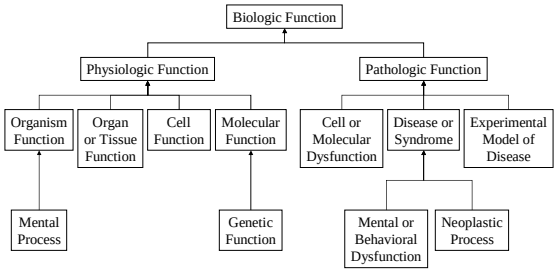
Semantic Network

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

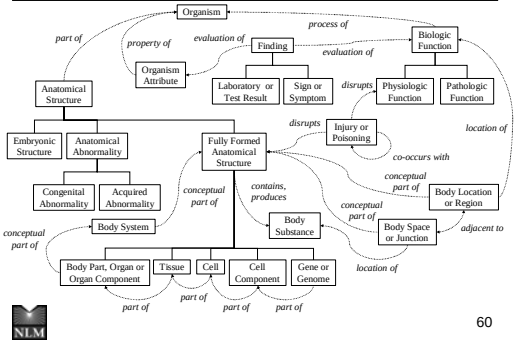
Semantic Network

- 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

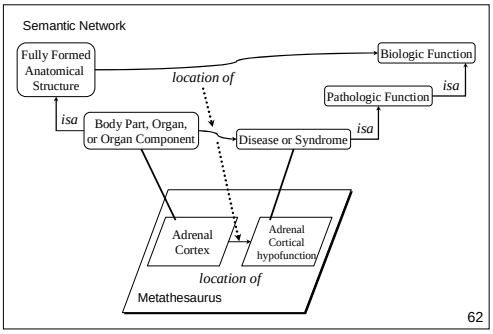


Why a semantic network?

- 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

61

Relationships can inherit semantics



62

SPECIALIST Lexicon and lexical tools

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

64

Morphology

- Inflection
  - noun — nucleus, nuclei
  - verb — cauterize, cauterizes, cauterized, cauterizing
  - adjective — red, redder, reddest
- Derivation
  - verb ⇔ noun — cauterize -- cauterization
  - adjective ⇔ noun — red -- redness

65

Orthography

- Spelling variants
  - oe/e — oesophagus - esophagus
  - ae/e — anaemia - anemia
  - ise/ize — cauterise - cauterize
  - genitive mark — Addison's disease
Addison disease
Addisons disease

66

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

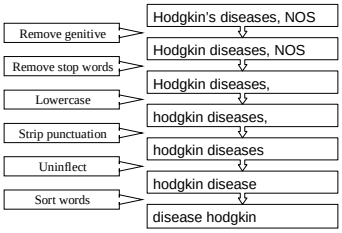
67

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

68

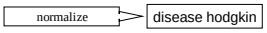
Normalization



69

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



70

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

71

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

72

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)

Part II

How to use the UMLS?

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

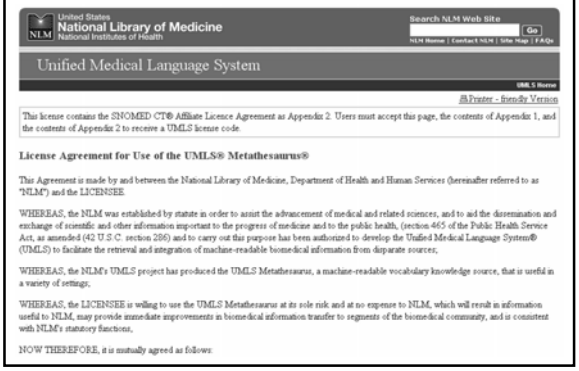
Part II
How to use the UMLS?

(1) Obtaining a license

First step License agreement

- Online Web-based license:
  http://www.nlm.nih.gov/research/umls/license.html
  - Read license — Accept and continue
  - Read appendix 1 and 2 — Accept and continue
  - Print a copy for your records — Printer-friendly Version
  - Complete the Web form — Submit
  - 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

http://www.nlm.nih.gov/research/umls/license.html



United States National Library of Medicine, National Institutes of Health

Search NLM Web Site

Unified Medical Language System

UMLS Home

Printer - friendly Version

This license contains the SNOMED CT® Affiliate Licence Agreement as Appendix 2. Users must accept this page, the contents of Appendix 1, and the contents of Appendix 2 to receive a UMLS license code.

License Agreement for Use of the UMLS® Metathesaurus®

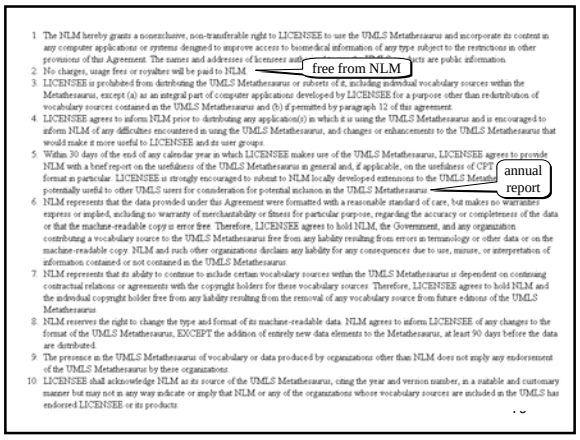
This Agreement is made by and between the National Library of Medicine, Department of Health and Human Services (hereinafter referred to as "NLM") and the LICENSEE.

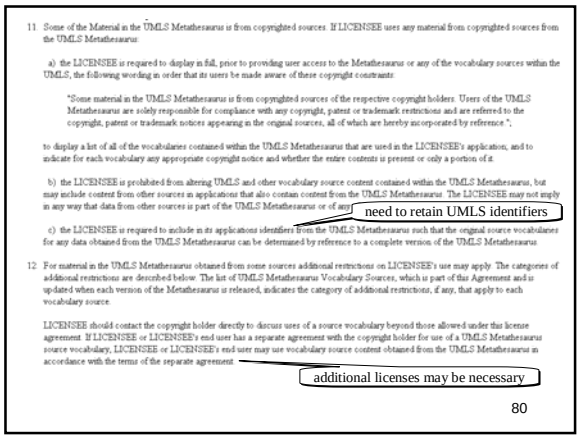
WHEREAS, the NLM was established by statute in order to assist the advancement of medical and related sciences, and to aid the dissemination and exchange of scientific and other information important to the progress of medicine and to the public health, (section 465 of the Public Health Service Act, as amended (42 U.S.C. section 286) and to carry out this purpose has been authorized to develop the Unified Medical Language System® (UMLS) to facilitate the retrieval and integration of machine-readable biomedical information from disparate sources;

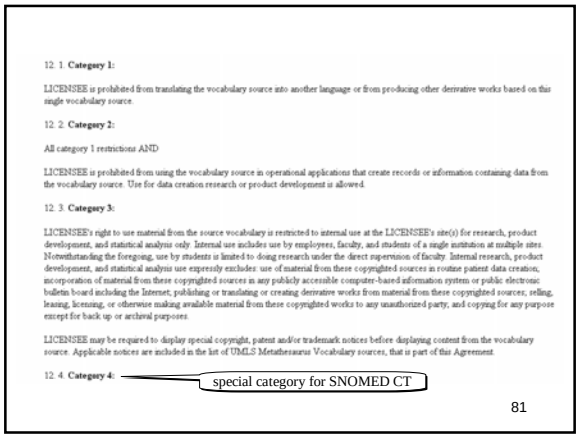
WHEREAS, the NLM's UMLS project has produced the UMLS Metathesaurus, a machine-readable vocabulary knowledge source, that is useful in a variety of settings;

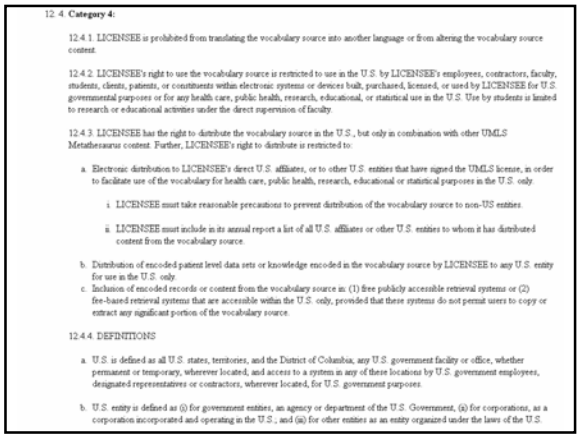
WHEREAS, the LICENSEE is willing to use the UMLS Metathesaurus at its sole risk and at no expense to NLM, which will result in information useful to NLM, may provide immediate improvements in biomedical information transfer to segments of the biomedical community, and is consistent with NLM's statutory functions;

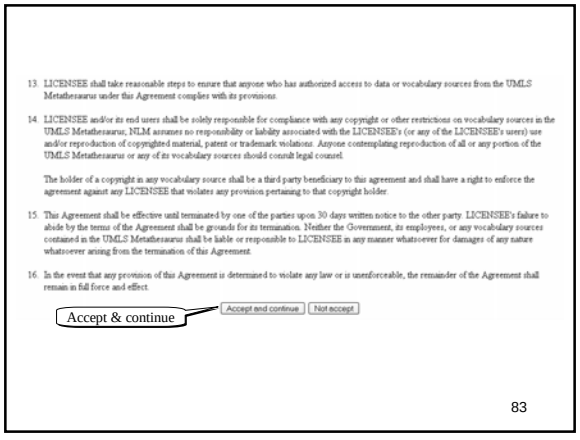
NOW THEREFORE, it is mutually agreed as follows:

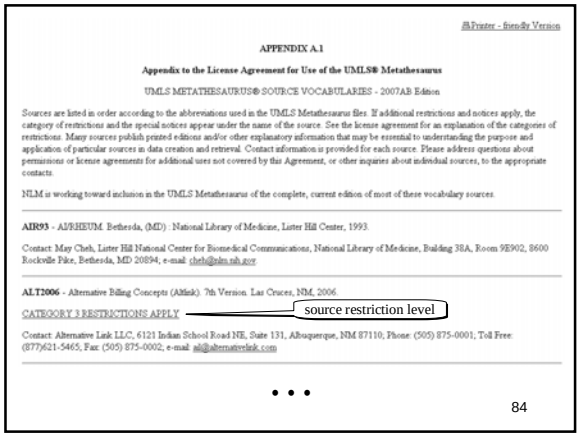

free from NLM
annual report


need to retain UMLS identifiers
additional licenses may be necessary
80


special category for SNOMED CT
81




Accept & continue
Accept and continue
Not accept
83


source restriction level
84

86

License Restriction Levels 0-4 (2007AB)

- 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!

87

Part II
How to use the UMLS?

(2) Remote access

Remote Access

- UMLS Knowledge Source Server:
  http://umlsks.nlm.nih.gov
- Web search interface
- Application Programming Interface (API)
- Coming soon: web services

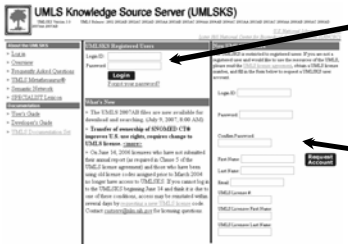
89

Knowledge Source Server
Web search interface

UMLSKS Web search interface

- Logging in
- Basic searching
- Advanced searching

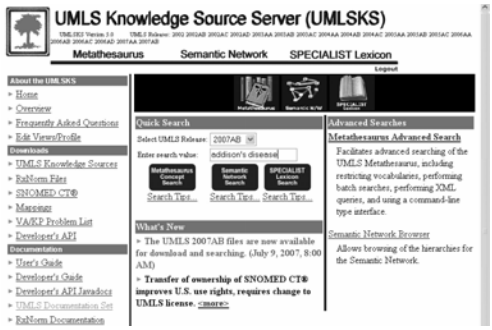
UMLSKS Web search interface log in



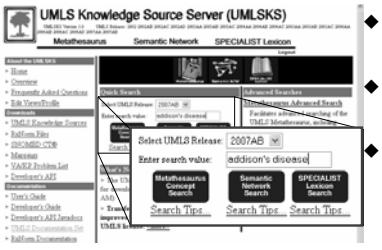
UMLS Knowledge Source Server Home Page

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

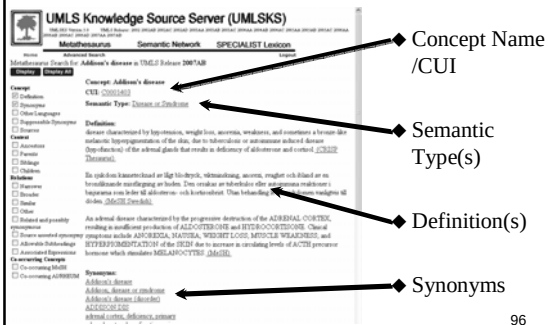
UMLS Knowledge Source Server Home Page



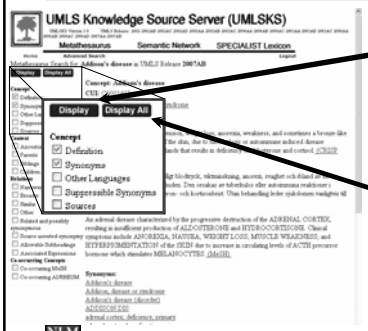
Metathesaurus Basic Search
Addison's disease



Concept Report *Addison's disease*

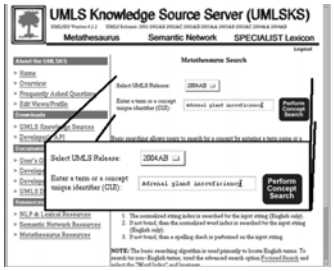


Display All



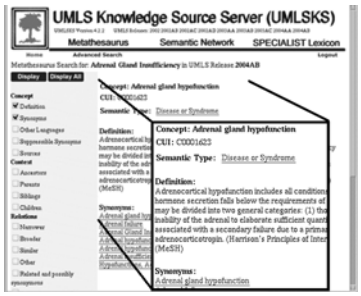
- "Display" shows results for selected options
- "Display All" shows results for all available options

Metathesaurus Basic Search
Adrenal gland insufficiency



- Specify:
  - UMLS Release
  - Search term
- Algorithm:
  - Search Normalized String
  - Search Normalized Word
  - Suggest Spelling

Basic Concept Report
Adrenal gland insufficiency



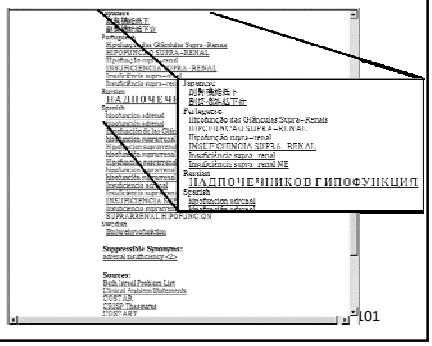
Concept Report Display All
Adrenal Gland Insufficiency

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

Concept Report Display All (continued)

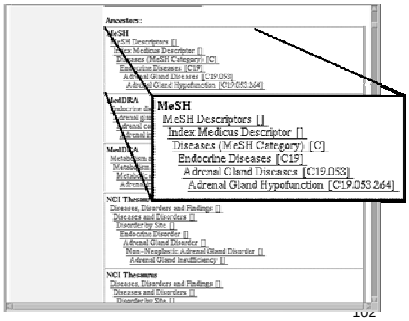
Synonyms

Sources



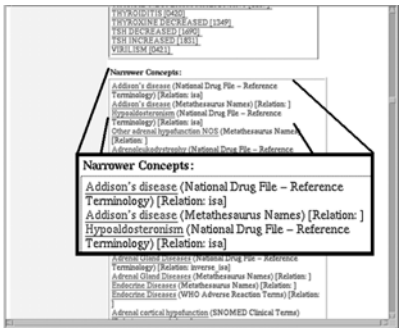
Concept Report Display All (continued)

Hierarchies



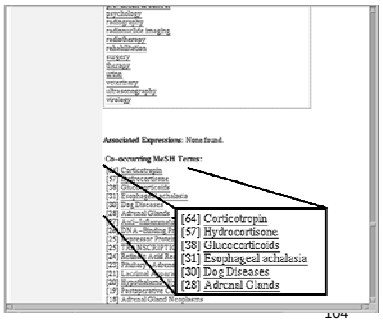
Concept Report Display All (continued)

Relations



Concept Report Display All (continued)

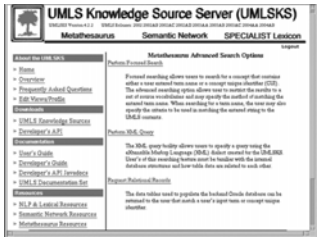
Co-occurrence data



104

Metathesaurus Advanced Search Options

- Focused Search
- Raw Relational Records



105

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

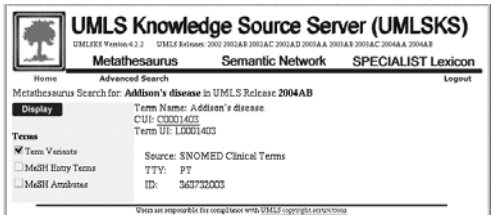
106

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

107

Addison's disease in SNOMED CT Preferred Term and Code

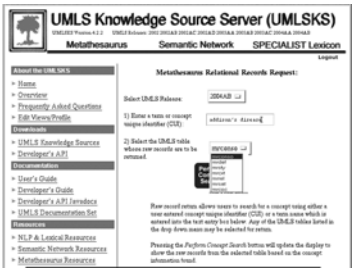


- TTY: Term Type
- ID: Source Code Descriptor

108

Metathesaurus Advanced Search Feature Relational Record Request

- UMLS Release
- Search Term
- UMLS Relational Table



109

Relational Records MRCONSO.RRF

110

Semantic Network Searching

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

111

Semantic Network Search

- Enter search string

-or-

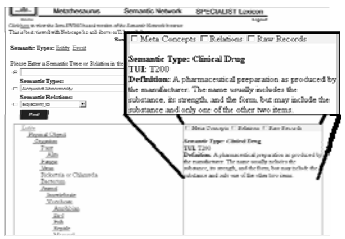
- Select semantic type

-or-

- Select semantic relation

112

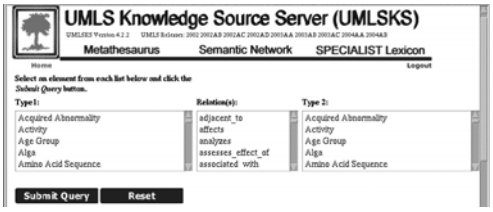
Semantic Type Clinical Drug



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

113

Show Relations Between Types



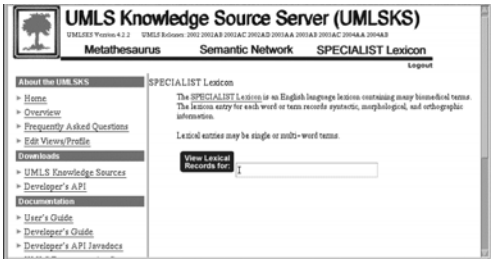
- Validates whether a selected Semantic Relationship (SR) holds between two selected Semantic Types (ST)

114

SPECIALIST Lexicon Searching

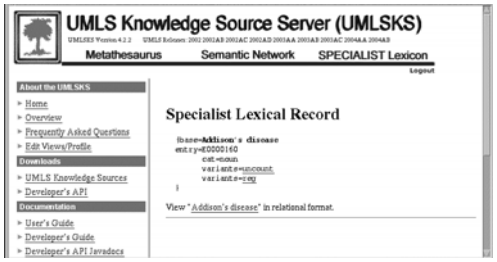
- Select Tab along top
- Quick search

SPECIALIST Lexicon Search



116

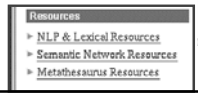
SPECIALIST Lexical Record



117

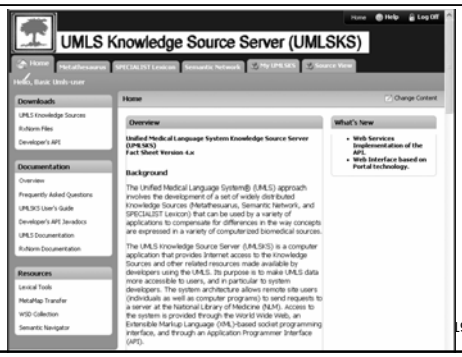
UMLS Resources

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

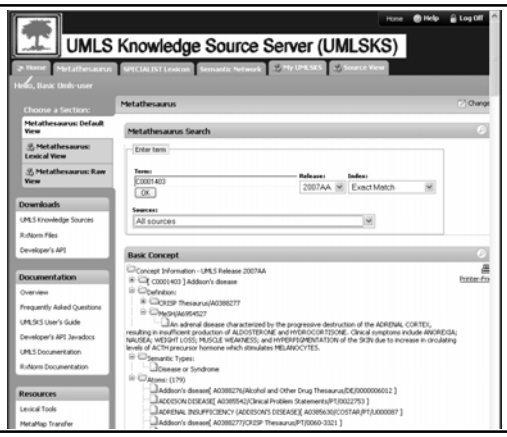


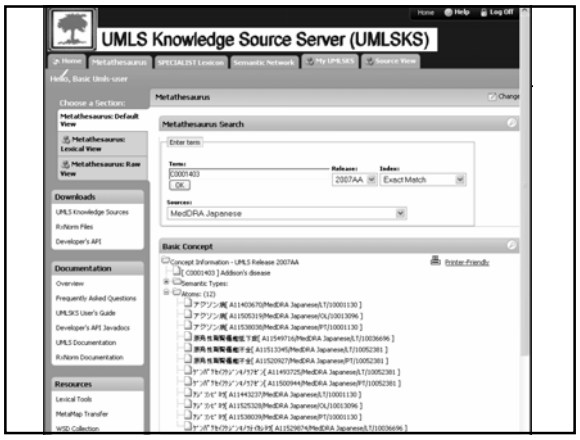
118

The "new" UMLSKS (coming soon)



19

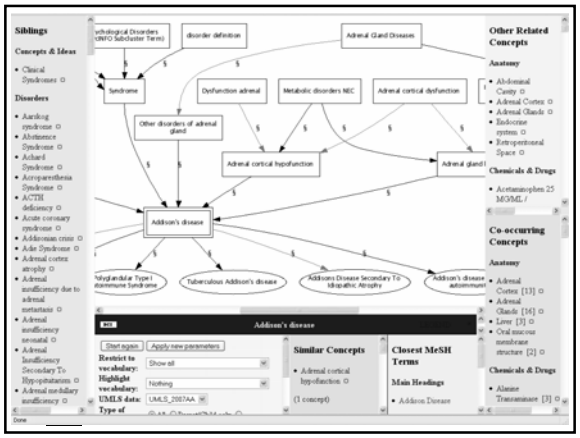




UMLS Semantic Navigator

- Web-based
  - http://mor.nlm.nih.gov/perl/semnav.pl
- Concept- and relation-centric
- Displays contexts graphically
- Displays all relations simultaneously
- Excludes hierarchical cycles in the UMLS graph
- Search
  - By CUI
  - By word

122

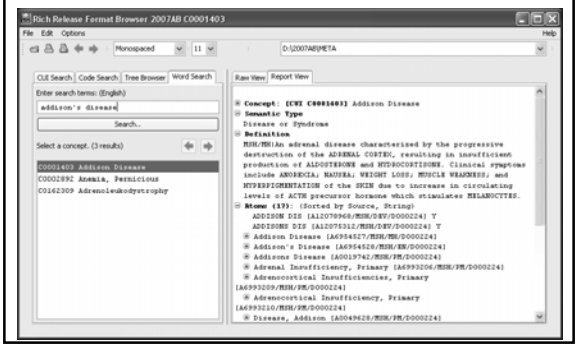


RRF Browser

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

124

RRF Browser



Knowledge Source Server
Application Programming Interface

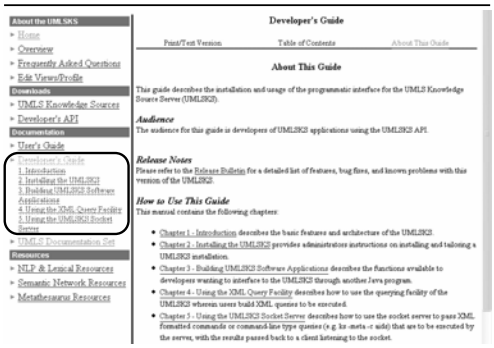
UMLSKS API basics

- Remote server at NLM
- Local application connected through

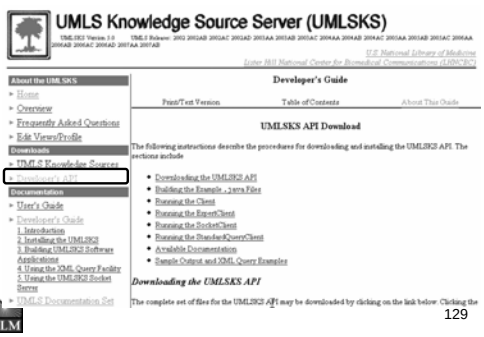
Java RMI	TCP/IP socket
◆ Java-based applications ◆ Developer's Guide: Chapter 3 ◆ Set of Java classes (part of the UMLSKS API download) ◆ Detailed *Javadoc* documentation online and with API download	◆ XML-based queries ◆ Developer's Guide: Chapter 5 ◆ XML schema ◆ Socket server • Host: umlsks.nlm.nih.gov • Port: 8042

127

Developer's Guide



Documentation Java API



129

Documentation Javadocs

130

Sample XML query (1) Current version

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

⇨

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

131

Sample XML query (2) Concepts by string

```
<?xml version="1.0"?>
<findCUI version="1.0">
<conceptName>appendicectomy</conceptName>
<language>ENG</language>
<exact/>
<noSuppressibles/>
</findCUI>
```

⇨

```
<?xml version="1.0"?>
<ConceptIdCollection version="1.0">
  <release>2004AB</release>
   <conceptId>
    <cui>C0003611</cui>
    <cn>Appendectomy</cn>
   </conceptId>
</ConceptIdCollection>
```

132

Sample XML query (3) Concepts properties

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

```
<?xml version="1.0"?>
<SemanticTypeCollection version="1.0">
<release>2004AB</release>
<cui>C0033572</cui>
<cn>Prostate</cn>
    <semanticType>
       <tui>T023</tui>
       <sty>Body Part, Organ,
          or Organ Component</sty>
    </semanticType>
</SemanticTypeCollection>
```

133

Sample XML query (4) Relationships

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

```
<?xml version="1.0"?>
<RelationCollection version="1.0">
[…]
   <relation>
     <aui>A3188910</aui>
     <sab>SNOMEDCT</sab>
     <relSource>
        <cui>C0007112</cui>
        <cn>Adenocarcinoma of prostate</cn>
        <aui>A3318222</aui>
        <rel>RO</rel>                    […]
        <rui>R54806623</rui>
        <rela>has_finding_site</rela>
     </relSource>
   </relation>
[…]
```

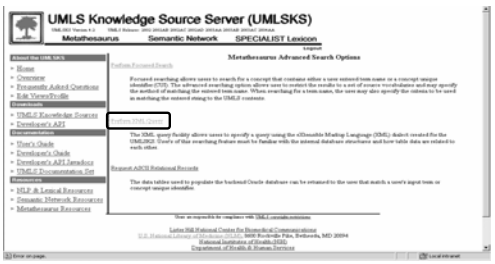
Sample XML query (5) All semantic type IDs

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

```
<?xml version="1.0"?>
<SemNetIdCollection version="1.0">
    <release>2004AB</release>
    <semnetId>
       <name>Acquired Abnormality</name>
       <ui>T020</ui>
       <semtype/>
    </semnetId>
    <semnetId>
       <name>Activity</name>
       <ui>T052</ui>
       <semtype/>
    </semnetId>
[…]
```

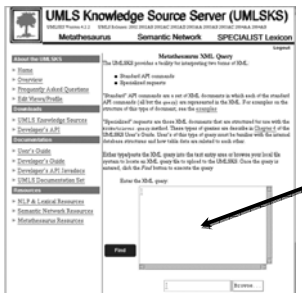
135

Performing XML queries from UMLSKS



136

Performing XML queries from UMLSKS




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

137

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

139

Using MetamorphoSys

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

140

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

141

Why use MetamorphoSys?

Changing Default Settings

- To alter the preferred name
- To alter suppressibility of specific source term types

142

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

143

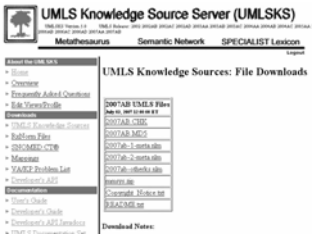
Machine Requirements

- 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

144

Download from UMLSKS …

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



145

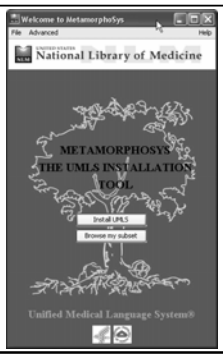
...or DVD?

- 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

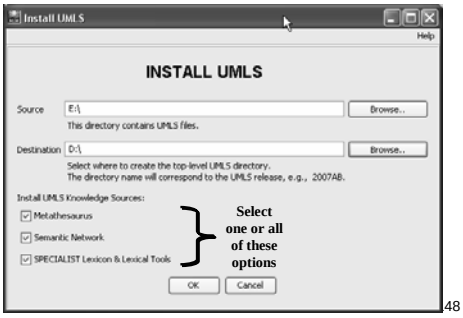
146

Welcome Screen



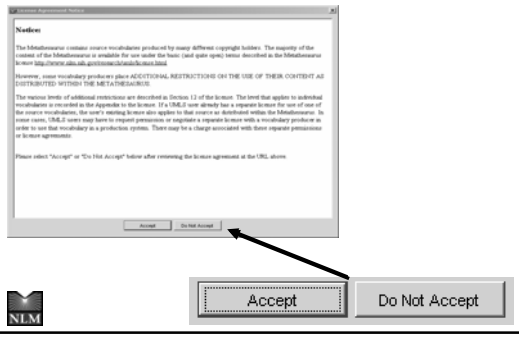
147

Install UMLS

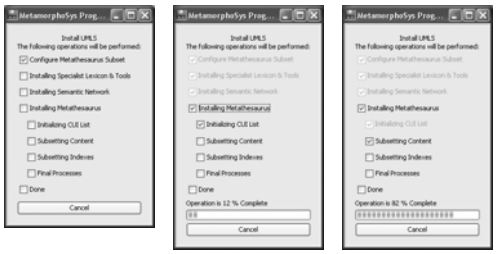


148

UMLS License Notice

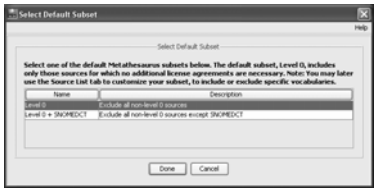


Installation progress monitor



150

Select a default subset



Level 0 → no separate additional license agreements

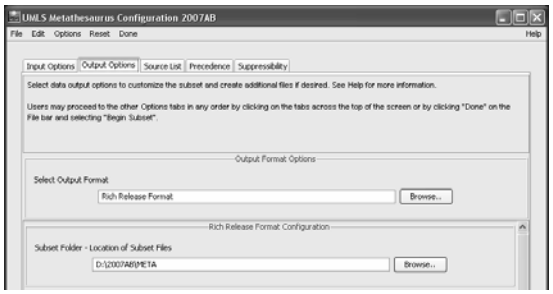
Level 0 + SNOMEDCT → Users from non-IHTSDO member countries must have separate license agreements

151

Input Options Tab

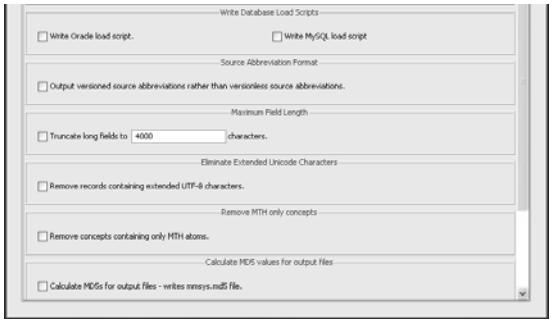
152

Output Options Tab

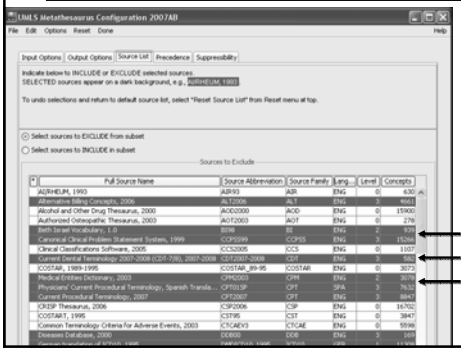


153

Output Options Tab



Source List Tab

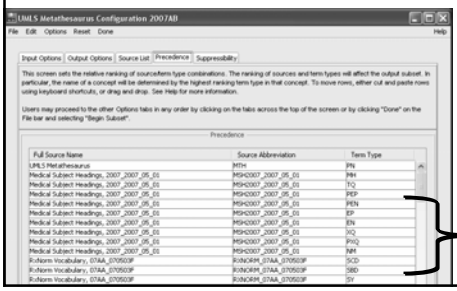


Highlighted rows are excluded from the subset.

155

Precedence Tab

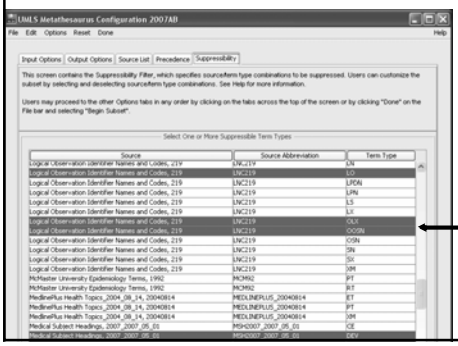
- Ranks names by types of terms within sources
- Highest ranking name determines the Preferred Name



Cut and paste rows to alter the preferred name

156

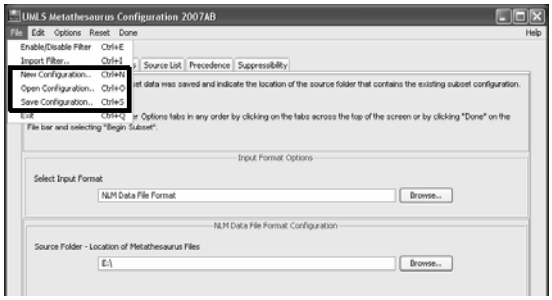
Suppressibility Tab



Highlighted source term types will be marked as suppressible

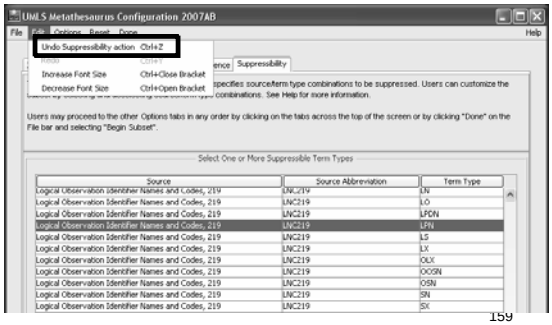
157

File menu



158

Edit menu

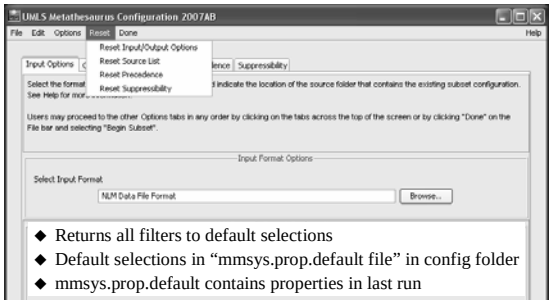


159

Options menu

160

Reset menu



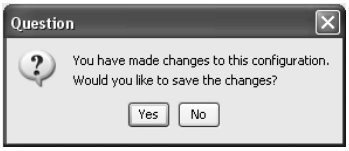
- Returns all filters to default selections
- Default selections in "mmsys.prop.default file" in config folder
- mmsys.prop.default contains properties in last run

161

Done – Begin Subset

162

Save configuration for next installation



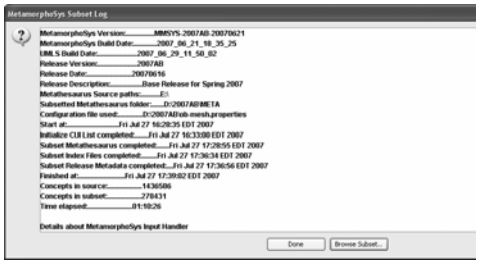
163

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

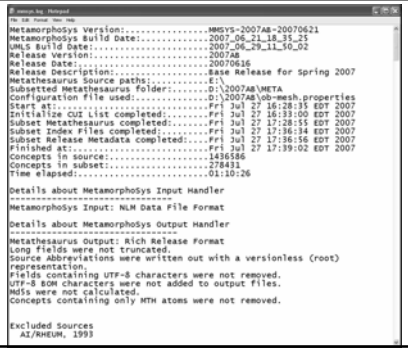
164

MetamorphoSys log



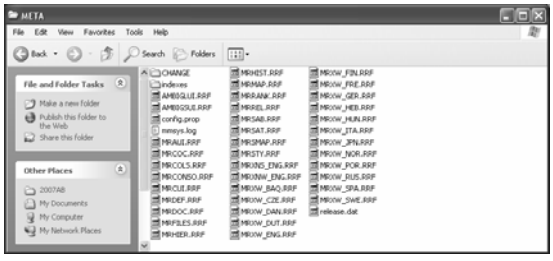
165

MetamorphoSys log



166

Output directory contents



167

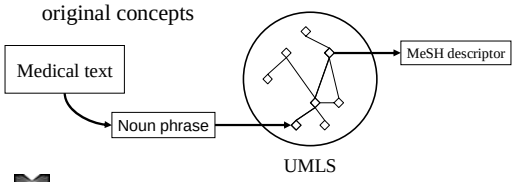
Part II
How to use the UMLS?

(4) A UMLS-based algorithm

Indexing Initiative

[Aronson & al., *AMIA*, 2000]

- 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



169

Restrict to MeSH

[Bodenreider & al., *AMIA*, 1998]

- 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

170

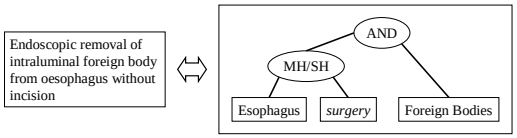
Restrict to MeSH Synonymy

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

171

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)



172

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

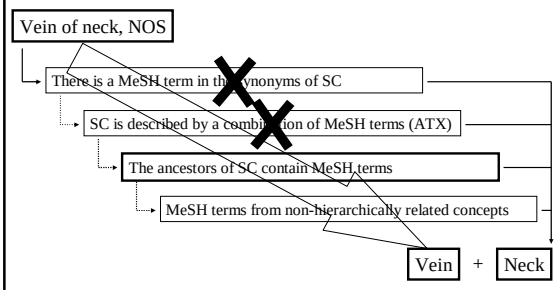
173

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)

174

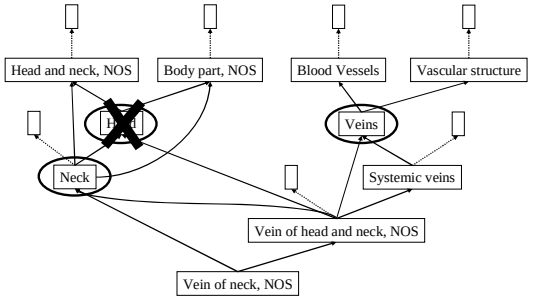
Restrict to MeSH Example



NLM

175

Restrict to MeSH Example

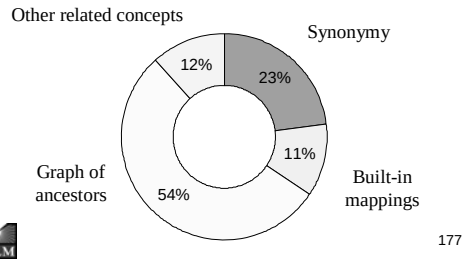


NLM

176

Restrict to MeSH Quantitative results

- 86% of UMLS concepts mapped to MeSH (2007)



NLM

177

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

NLM

178

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

181

Direct benefits

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

182

UMLS as an enabling resource

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

183

Limitations

Limitations

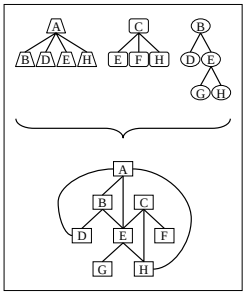
[Cimino, *JAMIA*, 1998]

- 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)

185

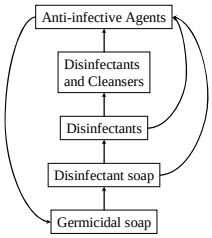
Structural inconsistency From trees to graph

- Multiple tree structures combined into a graph structure
- Expected: Directed acyclic graph (DAG)



186

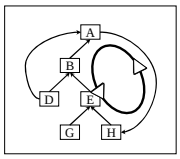
Structural inconsistency Cycles in the UMLS graph



187

Structural inconsistency Issues

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

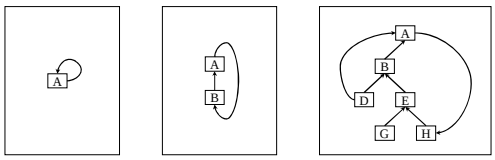


[Bodenreider, AMIA 2001]

188

Acyclicity

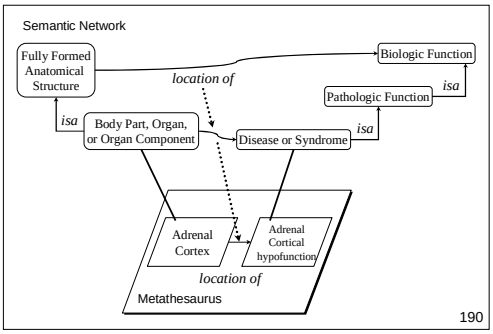
"back edge" from a child concept to a parent concept



Reflexive 13,000 Direct 1800 Indirect 120

189

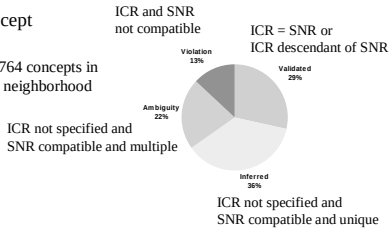
Semantic inconsistency A two-level structure



190

Semantic inconsistency A limited study

- 6894 interconcept relationships
  - among the 3764 concepts in the semantic neighborhood of "Heart"



[McCray A.T, Bodenreider O. A conceptual framework for the biomedical domain. In: Green R, Bean CA, Myaeng SH, editors. *The semantics of relationships: an interdisciplinary perspective*. Boston: Kluwer Academic Publishers; 2002. p. 181-198.]

191

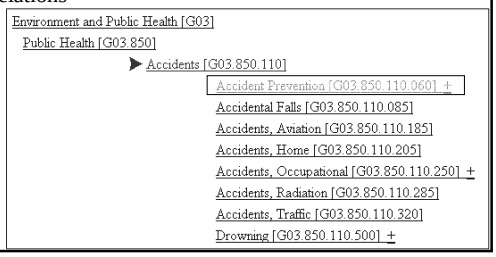
Semantic inconsistency Issues

- The UMLS integrates what terminologies represent
- Hierarchies in source vocabularies
  - Often task-driven rather than based on principles
  - Usually suitable for information retrieval
  - Not necessarily suitable for reasoning
- No automatic correction possible
  - Wrong categorization
  - Wrong inter-concept relationship
  - [Wrong semantic network relationship]

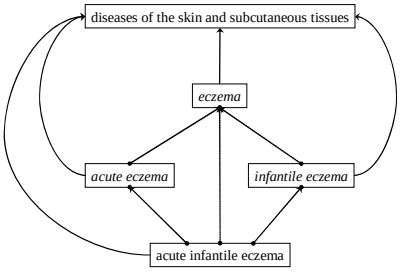
192

Underspecified relationships

- Relationship "attribute" not always present
- Relations used to create hierarchies vs. hierachical relations

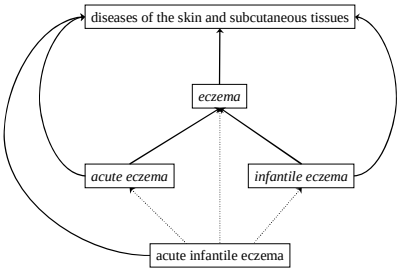


Missing relations Example



194

Missing relations Example



195

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]

196

Compensation mechanisms

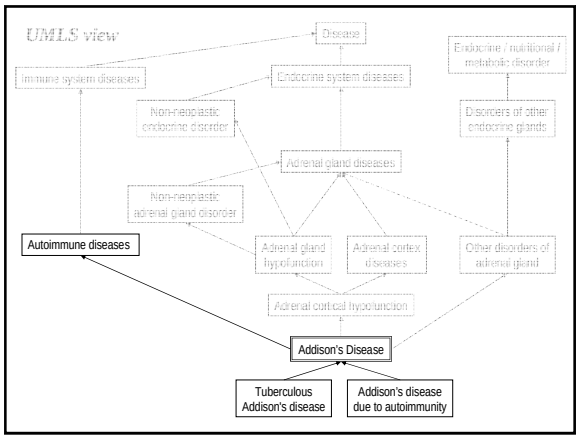
- Examples
  - Removing cycles from hierarchical relations
    - Using redundancy (number of sources asserting the relation)
    - Using terminological knowledge (e.g., NEC)
  - Lexically-suggested hyponymic relations
    - Properties of adjectival modification

197

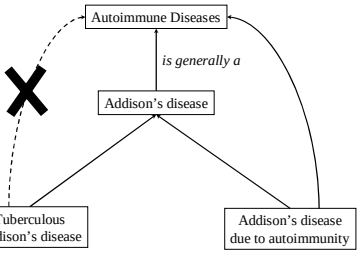
More limitations

- Semantics of hierarchical relations
- Some missing / wrong relations are hard to detect
- Some relations are present but hard to find

198

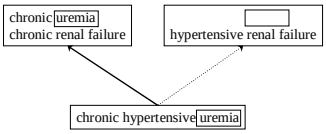


Semantics of hierarchical links



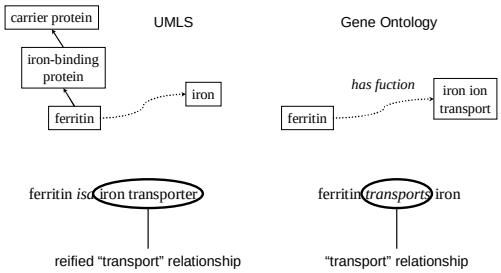
NLM

Relations Missing and difficult to detect



NLM

Relations Existing but difficult to find



NLM

How to address these limitations?

- Description logics
- Natural Language Processing (semantic interpretation of the terms)
- Comparing knowledge sources (alignment, inference)

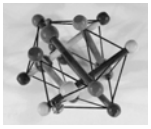
NLM

Summary

UMLS Summary

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

205



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
  - http:// www.nlm.nih.gov/research/umls/
- UMLS documentation
  - Formerly know as the "Green Book"
  - Now online documentation
  - http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML

208

References

- Short presentation
  - Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue), D267-70.
- UMLS as a research project
  - Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods Inf Med, 32*(4), 281-91.
  - Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc, 5*(1), 1-11.

209

References

- Technical papers
  - McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods Inf Med, 34*(1-2), 193-201.
- Comprehensive bibliography 1986-96

  http://www.nlm.nih.gov/pubs/cbm/umlscbm.html

210

Documentation and Support

UMLS documentation and support

- UMLS homepage
  - links to various UMLS resources
  - http://www.nlm.nih.gov/research/umls/
- UMLSKS homepage
  - links to the User's and Developer's guides
  - http://umlsks.nlm.nih.gov/
- UMLS mailing list
  - UMLSUSERS-L@LIST.NIH.GOV
- Email address for support
  - custserv@nlm.nih.gov

Appendix

UMLS files in
Rich Release Format

MRCONSO (sample rows 1..5) (2004AB)

	CUI	LAT	TS	LUI	STT	SUI	ISPREF	AUI	SAUI	SCUI	SDUI
1	C0001403	ENG	P	L0001403	PF	S0354372	Y	A4367951			
2	C0001403	ENG	P	L0001403	PF	S0354372	N	A2922421	485624014	363732003	
3	C0001403	ENG	P	L0001403	VC	S0010794	Y	A0019740		M0000346	D000224
4	C0001403	ENG	S	L0494851	PF	S2164152	N	A2018589			
5	C0001403	FRE	P	L3246333	PF	S3773545	Y	A3996251			D000224

	SAB	TTY	CODE	STR	SRL	SUPPRESS	CVF
1	MTH	PN	NOCODE	Addison's disease	0	N	
2	SNOMEDCT	PT	363732003	Addison's disease	4	N	
3	MSH	MH	D000224	Addison's Disease	0	N	
4	MDR	LT	10052381	Primary adrenal insufficiency	3	N	
5	MSHFRE	MH	D000224	Addison, maladie	3	N	

MRCONSO (sample rows 6..10) (2004AB)

	CUI	LAT	TS	LUI	STT	SUI	ISPREF	AUI	SAUI	SCUI	SDUI
6	C0001403	FRE	S	L1272481	PF	S1514427	Y	A1464383			
7	C0001403	GER	P	L1229627	PF	S1471573	Y	A4030156			D000224
8	C0001403	GER	S	L1239271	PF	S1481217	Y	A4034094			D000224
9	C0001403	JPN	P	L3437833	PF	S3965327	Y	A4264008			D000224
10	C0001403	JPN	S	L3465347	PF	S3992841	Y	A4291522			D000224

	SAB	TTY	CODE	STR	SRL	SUPPRESS	CVF
6	WHOFRE	IT	0410	MALADIE D'ADDISON	2	N	
7	MSHGER	MH	D000224	Addison-Krankheit	3	N	
8	MSHGER	SY	D000224	Bronzehautkrankheit	3	N	
9	MSHJPN	MH	D000224	Addison病	3	N	
10	MSHPJN	SY	D000224	副腎性黒皮症	3	N	

MRCONSO (sample rows 11-13) (2004AB)

	CUI	LAT	TS	LUI	STT	SUI	ISPREF	AUI	SAUI	SCUI	SDUI
11	C0001403	POR	P	L3302998	PF	S3831123	N	A6382080			
12	C0001403	RUS	P	L3336992	PF	S3864473	Y	A4157629			
13	C0001403	SPA	P	L1226877	PF	S1468823	Y	A1419475			

	SAB	TTY	CODE	STR	SRL	SUPPRESS	CVF
11	MDRPOR	LT	1001130	Doença de Addison	3	N	
12	MSHRUS	MH	D000224	АДДИСОНОВА БОЛЕЗНЬ	3	N	
13	WHOSPA	IT	0410	ADDISON, ENFERMEDAD	3	N	

MRHIER (sample rows)

(2004AB)

	CUI	AUI	CXN	PAUI	SAB	RELA
1	C0001403	A0019740	1	A0020270	MSH	
2	C0001403	A0019740	2	A0028022	MSH	
3	C0001403	A0019743	3	A1988358	PSY	member_of_cluster
4	C0001403	A2922421	1	A3307650	SNOMEDCT	isa
5	C0001403	A2922421	2	A3307650	SNOMEDCT	isa

	PTR	HCD	CVF
1	A0434168.A2367943.A2366890.A0135391.A0054194.A0020267.A0020270	C19.053.264.263	
2	A0434168.A2367943.A2366890.A0135391.A0072566.A0028022	C20.111.163	
3	A0449751.A1988279.A1988358		
4	A3684559.A3886745.A2880798.A3398606.A3399335.A3398961.A2872359.A2872360.A3307650		
5	A3684559.A3886745.A2880798.A3398606.A3399335.A3398961.A2872359.A2933400.A2989549.A3307650		

MRREL (sample rows)

(2004AB)

	CUI1	AUI1	STYPE1	REL	CUI2	AUI2	STYPE2
1	C0001403		CUI	RB	C0001621		CUI
2	C0001403	A0019738	AUI	SY	C0001403	A0049628	AUI
3	C0001403	A2922421	SCUI	CHD	C0085859	A2977940	SCUI
4	C0001403	A6326321	SCUI	RO	C0688490	A6339383	SCUI
5	C0001403	A0019743	AUI	PAR	C0935495	A1988358	AUI

	RELA	RUI	SRUI	SAB	SL	RG	DIR	SUPPRESS	CVF
1		R02837989		MTH			N	N	
2		R18849683		MSH	MSH			N	
3	isa	R19859511	1658795027	SNOMEDCT	SNOMEDCT	0	Y	N	
4	may_treat	R27600039		NDFRT	NDFRT			N	
5	has_member	R08110401		PSY	PSY			N	

MRDEF

(2004AB)

CUI	AUI	ATUI	SATUI	SAB	DEF	SUPPRESS	CVF
C0001403	A0019740	AT15061584		MSH	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.	N	

MRSAT (sample rows)

(2004AB)

	CUI	LUI	SUI	METAUI	STYPE	CODE
1	C0001403	L0001403	S0010792	A0019738	AUI	D000224
2	C0001403	L0001403	S0010794	A6326321	SCUI	C712
3	C0001403	L0001403	S0354372	A2922421	SAUI	363732003
4	C0001403			R15742591	SRUI	
5	C0001403				CUI	

	ATUI	SATUI	ATN	SAB	ATV	SUPPRESS	CVF
1	AT15321482		DID	MSH	D000224	N	
2	AT33411754		MESH_UI	NDFRT	D000224	N	
3	AT24166602		DESCRIPTION STATUS	SNOMEDCT	0	N	
4	AT27438950		REFINABILITY	SNOMEDCT	0	N	
5	AT02925340		ST	MTH	R	N	

MRSTY

(2004AB)

CUI	TUI	STN	STY	ATUI	CVF
C0001403	T047	B2.2.1.2.1	Disease or Syndrome	AT17683850	

MRHIST (sample rows)

(2004AB)

	CUI	SOURCEUI	SAB	SVER	CHANGETYPE
1	C0001403	1198962018	SNOMEDCT	20020731	0
2	C0001403	1212124016	SNOMEDCT	20020731	0
3	C0001403	1490869013	SNOMEDCT	20030131	0
4	C0001403	363732003	SNOMEDCT	20020129	0
5	C0001403	373662000	SNOMEDCT	20020731	0

	CHANGEKEY	CHANGEVAL	REASON	CVF
1	DESCRIPTIONSTATUS	0		
2	DESCRIPTIONSTATUS	0		
3	DESCRIPTIONSTATUS	0		
4	CONCEPTSTATUS	0		
5	CONCEPTSTATUS	0		