AMIA 2000 Fall Symposium - Tutorial T13

Customizing the UMLS Metathesaurus for Your Applications

William T. Hole, MD
Betsy L. Humphreys, MLS
Suresh Srinivasan, MS
Laura Roth, MLS

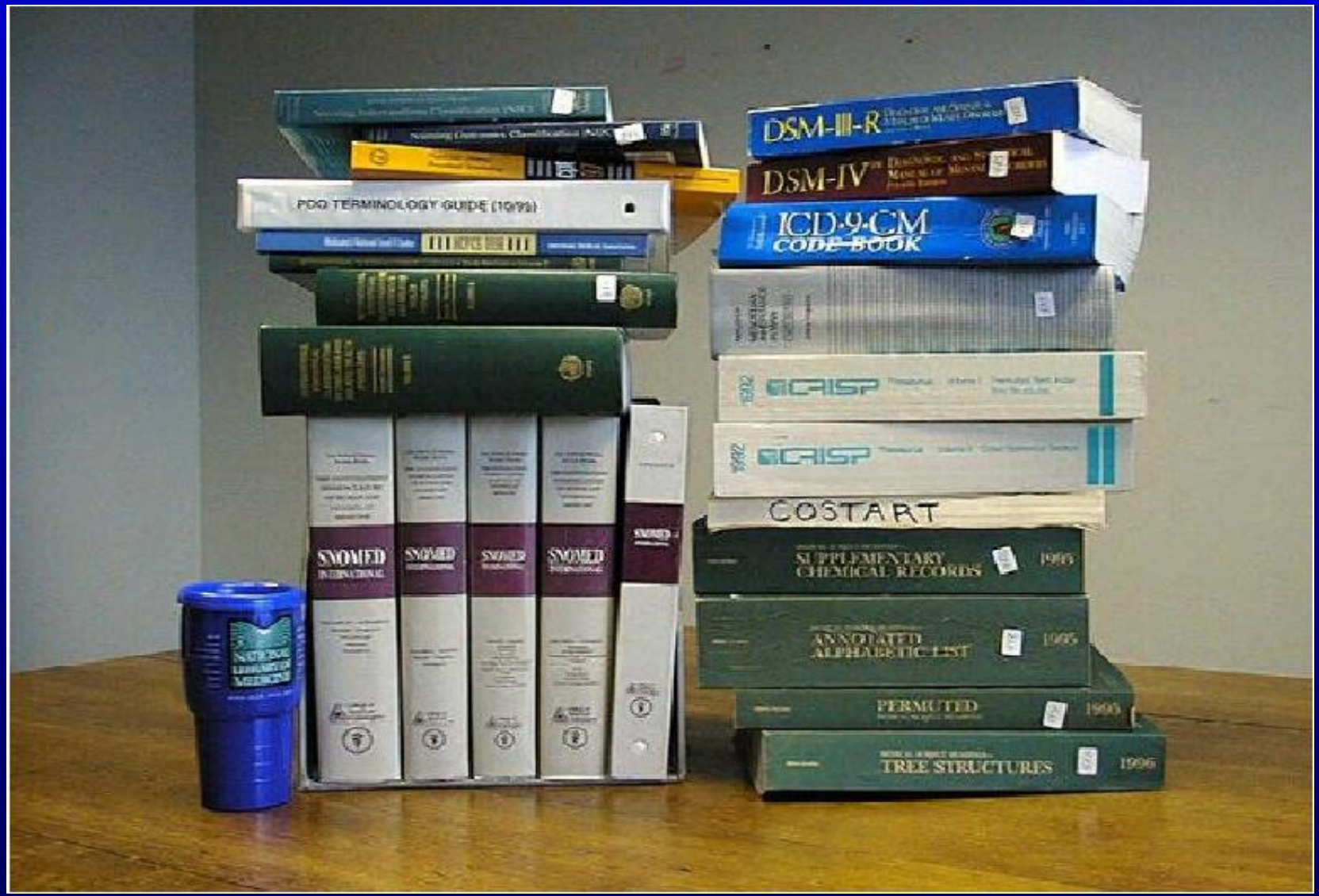
National Library of Medicine

Saturday, November 4, 2000 - 1:00-4:30 pm

Outline of Tutorial

- *Why customize - Betsy Humphreys*
- How to customize - *Bill Hole*
- A tool to help you customize - *Laura Roth and Suresh Srinivasan*
- Adding “local” terminology - *Bill Hole*

A few Metathesaurus Ingredients...



UMLS Source "Vocabularies"

- Widely varying purposes, structures, properties that do ***not*** add up to single ontology or view of the world:
  - Thesauri, e.g., MeSH
  - Statistical Classifications, e.g., ICD
  - Billing Codes, e.g., CPT
  - Clinical coding systems, e.g., SNOMED
  - Lists of controlled terms, e.g., HL7 valid values

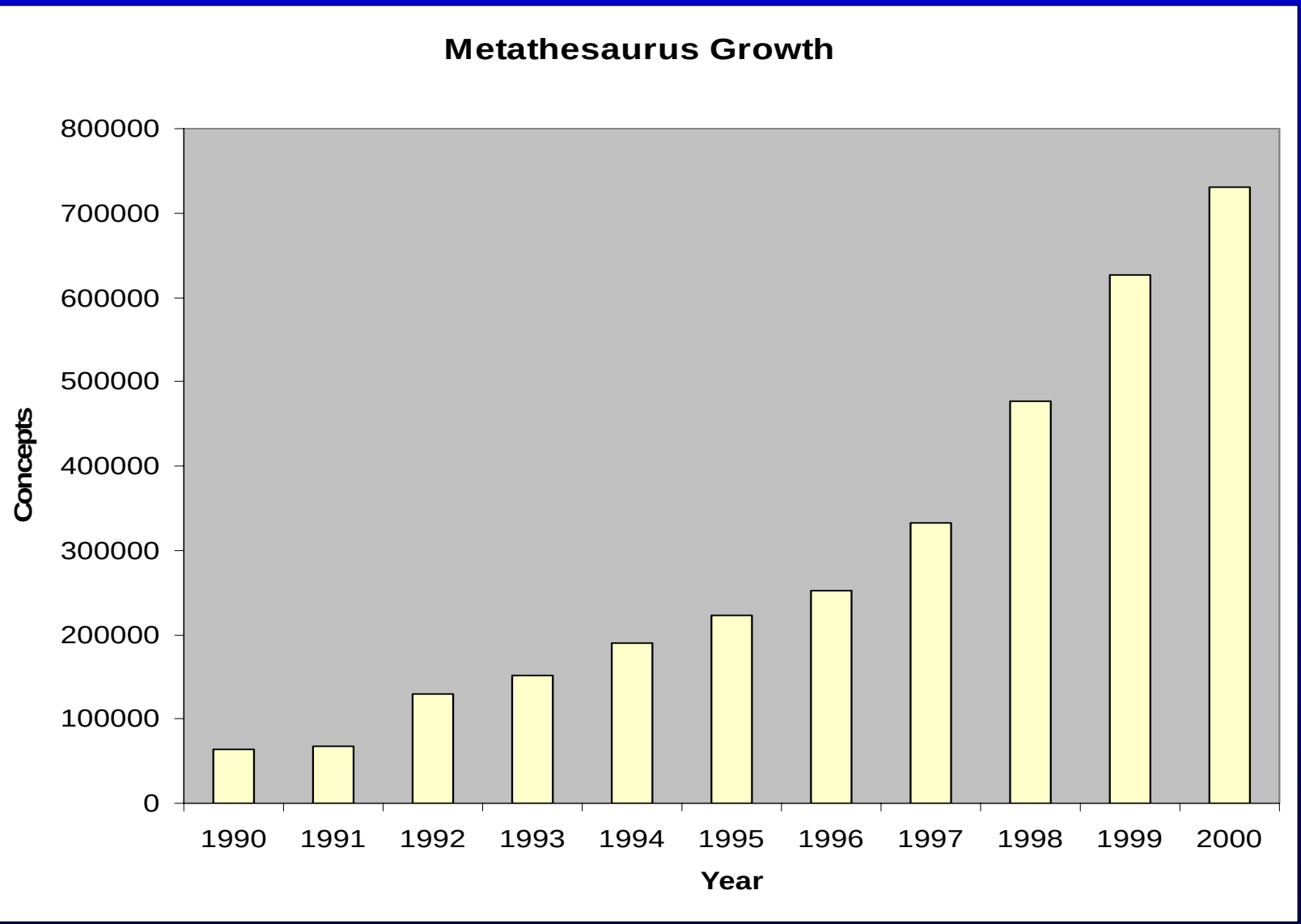
How to combine them?

Meta Processor,
Alpha 0.001

Not really ….

- “The Metathesaurus preserves the meanings, hierarchical connections, and other relationships between terms present in its source vocabularies, while adding certain basic information about each of its concepts and establishing new relationships between concepts and terms from different source vocabularies.”

It keeps getting bigger….



2000 UMLS Metathesaurus

- 730,000 concepts
- 1,338,650 “terms” (Eye, Eyes, eye = 1)
- 1,593,730 “strings”/concept names - (Eye, Eyes, eye = 3)
- >50 source vocabularies

UMLS Metathesaurus

Why Customize?
3 basic reasons

- Because *nobody* needs or wants all of it for any specific set of purposes
  - extraneous vs. pernicious content
- Because you don’t have the licenses required for operational use of all source vocabularies
- Because the default “preferred name” is not best for your applications

Possibly Extraneous, e.g.,

- Terms in languages other than English
- Redundant minor variations
- Procedure codes, when your application is focused on problems

Possibly Pernicious, e.g.,

- Terms that lack face validity
- Abbreviations and short forms
- Other less than beautiful “suppressible synonyms” already identified by NLM
- Unhelpful “views of the world”

- Home
- Metathesaurus
- Semantic Network
- SPECIALIST Lexicon
- Expert Search
- Download Results
- Comments
- Help

Your query term is "prostate"

This query term has multiple concepts associated with it in the Metathesaurus. Select a concept and click on submit button to obtain information about that concept.

(•) Prostate
Semantic Type:
Body Part, Organ, or Organ Component

( ) Prostatic Diseases
Semantic Type:
Disease or Syndrome

( ) Benign neoplasm of prostate
Semantic Type:
Disease or Syndrome
Semantic Type:
Neoplastic Process

( ) Carcinoma in situ of prostate
Semantic Type:
Neoplastic Process

( ) Neoplasm of uncertain or unknown behavior of prostate
Semantic Type:
Neoplastic Process

SUBMIT

- Home
- Metathesaurus
- Semantic Network
- SPECIALIST Lexicon
- Expert Search
- Download Results
- Comments
- Help

Your query term is "ER"

This query term has multiple concepts associated with it in the Metathesaurus. Select a concept and click on submit button to obtain information about that concept.

◉ Endoplasmic Reticulum

Semantic Type:

Cell Component

○ Estrogen Receptors

Semantic Type:

Amino Acid, Peptide, or Protein

Semantic Type:

Receptor

Definition:

Cytoplasmic proteins that bind estrogens and migrate to the nucleus where they regulate DNA transcription. Evaluation of the state of estrogen receptors in breast cancer patients has become clinically important.

Definition:

ER. Protein found on some breast cancer cells to which estrogen will attach. Breast cancer cells that are estrogen receptor positive (ER+) need the hormone estrogen to grow and usually respond to hormone treatment.

SUBMIT

Examples of customization

- Natural Language Processing version of the Metathesaurus, ***excluding*** “suppressible synonyms”, highly structured terms such as those in LOINC, etc.

UMLS Knowledge Source Server

ALMECILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AMDINOCILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AMIKACIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AMOXICILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AMOXICILLIN AND CLAVULANATE:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AMPHOTERICIN B:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AMPICILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AMPICILLIN AND SULBACTAM:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AZITHROMYCIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AZLOCILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
AZTREONAM:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
BACAMPICILLIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER
BUTIROSIN:SUSCEPTIBILITY:POINT IN TIME:ISOLATE AND SERUM:ORDINAL:SERUM BACTERICIDAL TITER

Examples of customization

- PubMed’s UMLS/MeSH mapping table

About Entrez

Entrez PubMed
Overview
Help | FAQ
New/Noteworthy

PubMed Services
Journal Browser
MeSH Browser
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Cubby NEW

Related Resources
Order Documents
Grateful Med
Consumer Health
Clinical Alerts
ClinicalTrials.gov

1: *Urology* 2000 Sep 1;56(3):544

Related Articles, Books, LinkOut

Incidence of acute myocardial infarction and cause-specific mortality after transurethral treatments of prostatic hypertrophy.

Weisman KM, Larijani GE, Goldberg ME

Publication Types:
- Letter

MeSH Terms:
- Cause of Death
- Human
- Male
- Myocardial Infarction/mortality
- Myocardial Infarction/complications*
- Prostatic Hyperplasia/mortality
- Prostatic Hyperplasia/complications*

Examples of customization

- National Cancer Institute's Metaphrase™ implementation, which changes default "preferred name" precedence

- Home
- Metathesaurus
- Semantic Network
- SPECIALIST Lexicon
- Expert Search
- Download Results
- Comments
- Help

BASIC CONCEPT INFORMATION

Concept Name: Persian Gulf Syndrome

UI: C0282550

Semantic Type: Disease or Syndrome

Definition (MSH2000):
Unexplained symptoms reported by veterans of the Persian Gulf War with Iraq in 1991. The symptoms reported include fatigue, skin rash, muscle and joint pain, headaches, loss of memory, shortness of breath, gastrointestinal and respiratory symptoms, and extreme sensitivity to commonly occurring chemicals. (Nature 1994 May 5;369(6475):8)

Synonyms :
Gulf War Syndrome

Sources: MSH2000

Other Languages :
Persianlahden oireyhtymae - Finnish
GOLFE PERSIQUE, SYNDROME - French
GUERRE GOLFE, SYNDROME - French
SYNDROME GOLFE PERSIQUE - French

ANCESTORS

MSH2000
Diseases (MeSH Category) [C]
Disorders of Environmental Origin [C21]
Occupational Diseases [C21.447]
Persian Gulf Syndrome [C21.447.653]

SNMI98
DISEASES/DIAGNOSES
INJURIES AND POISONINGS
POISONINGS
TOXIC EFFECTS OF NONMEDICINAL
Gulf war syndrome, NOS [DD-84D10]

Metaphrase **NCI-v3.0 Version 1.0**

persian gulf syndrome Search Concepts T

Max Returns: 10 Sources: All

For String For Code Short Circuit By Score

Matching Concepts

Persian Gulf Syndrome (Gulf war syndrome, NOS)	Disease or Syndrome
Persian Gulf	Geographic Area
Persian	Population Group
Gulf butterfish (Peprilus burti)	Fish
Gulf darter (Etheostoma swaini)	Fish
Gulf flounder (Paralichthys albigutta)	Fish
Gulf grouper (Mycteroperca jordani)	Fish

Please send questions and comments to: hartel@nih.gov.

Definitions | Synonyms | Sources | Broader_Concepts | Narrower_Concepts | Related_Concepts | Medications | Procedures | Laboratory | Diagnosis | *Suggest New Term*

C0282550: Gulf war syndrome, NOS

Disease or Syndrome

Definition(s)

MSH2000 Unexplained symptoms reported by veterans of the Persian Gulf War with Iraq in 1991. The symptoms reported include fatigue, skin rash, muscle and joint pain, headaches, loss of memory, shortness of breath, gastrointestinal and respiratory symptoms, and extreme sensitivity to commonly occurring chemicals. (Nature 1994 May 5;369(6475):8)

Synonym(s)

· Gulf war syndrome, NOS
· Persian Gulf Syndrome

Sources

MSH2000 SNMI98

View neighborhood in NCI OK

Broader Concepts

Occupational Diseases
TOXIC EFFECTS OF NONMEDICINAL SUBSTANCES

Examples of customization

- CliniWeb International, which uses MeSH and several of its translations

CliniWeb International

Browse Search Help Feedback About

OHSU Home Page

Welcome to the new CliniWeb International!

CliniWeb is an index and table of contents to clinical information on the World Wide Web. It now allows search terms to be entered in five different languages (English, German, French, Spanish, Portuguese) and has direct links to MEDLINE searches via the PubMed system at the National Library of Medicine. The CliniWeb database can be accessed by:

- Searching - using the SAPHIRE look-up system
- Browsing - through the MeSH hierarchy

Additional information about CliniWeb is also available.

We welcome your feedback by e-mail. Please notify us if you know of clinical sites we have overlooked or disagree with any indexing term selections.

DISCLAIMER

Customization is critical, but it ***requires*** a clear understanding of:

- Your functional requirements
- Characteristics of relevant UMLS source vocabularies
- Your license arrangements
- -- ***and*** Technical expertise
- Therefore, it is usually a team sport.

Outline of Tutorial

- Why customize - *Betsy Humphreys*
- *How to customize - Bill Hole*
- A tool to help you customize - *Laura Roth and Suresh Srinivasan*
- Adding “local” terminology - *Bill Hole*

How to Customize your Metathesaurus

- Limit vocabularies
- Limit languages
- Limit Relationships
- Aggregate by Relationships
- Add and Use Suppressibility
- Change Naming Precedence
- Select by attribute e.g. STYs
- Add local terminology

To work with Meta files:

- You will have to select the rows and columns you need, and combine them in ways which best meet your needs.
- You need your own programming and data management tools to work with the files, e.g.:
  - Unix: grep, sed, cut, awk, sort, join ...
  - RDBMS: Mysql, Access, Oracle, Ingres…
  - Other languages and extensions, e.g. Java, C++, , Visual Basic, Perl; interfaces to B-trees, RDBMS …
- We supply *MetamorphoSys*, a tool to exclude vocabularies and alter naming, described later.
- Some helpful files, such as RDBMS load scripts, are being developed - check **`http://umlsinfo.nlm.nih.gov`**

Using Unique Identifiers:

Simplified example: Concept Unique Identifier C0002871 links tables:

MRCON: ***Concept names***

C0002871|ENG|P|L0002871|PF|S0013742|Anemia|0|
C0002871|ENG|S|L0376533|PF|S0500659|Oligocythemia of red blood cells|3|
C0002871|FRE|P|L0162748|PF|*S0227229*|ANEMIE|2|

MRSO: ***Sources (vocabulary information)***

C0002871|L0002871|S0013742|MSH2000|MH|D000740|0|
C0002871|L0376533|S0500659|SNMI98|SY|DC-10010|3|
C0002871|L0162748|*S0227229*|INS2000|MH|D000740|3|
C0002871|L0162748|*S0227229*|WHOFRE|PT|0544|2|

MRSTY: ***Semantic Types***

C0002871|T047|Disease or Syndrome|

MRDEF: ***Definitions***

C0002871|CSP98|subnormal levels or function of erythrocytes, resulting in symptoms of tissue hypoxia.|
C0002871|MSH2000|A reduction in the number of circulating erythrocytes or in the quantity of hemoglobin.|
C0002871|PDQ99|A condition in which the number of red blood cells is below normal.|

First, Select the data you need:

The ‘Big Five’ Relational files:

MRCON: Concept Names

MRSO: Source vocabulary information

MRSTY: Semantic Types

MRREL: Most Relationships

MRSAT: Simple Attributes

MRCON: Concept Names

```
 CUI      |LAT|TS|LUI      |STT|SUI      |STR                                |LRL|
C0154009|ENG|P|L0180842|PF|S0245368|Benign neoplasm of prostate|0|
C0154009|ENG|S|L0524759|PF|S0598915|Benign prostatic tumour|3|
C0154009|ENG|s|L0033572|PF|S0999020|Prostate <3>|0|
C0154009|ENG|s|L0033572|VO|S0077252|Prostate|3|
C0154009|GER|P|L1258213|PF|S1500159|Gutartige Neubildung: Prostata|1|
 …
```

CUI Unique identifier for concept (“The name that never changes”;)
- Links ALL Concept information

LAT Language of Term

TS Term status

LUI Unique identifier for term (lexically similar strings, defined by lvg)

STT String type

SUI Unique identifier for string

STR String

LRL Least Restriction Level (lowest license restriction* for a source of this string)
- lowest license restriction level for this string; see MRSO Source Restriction level

** License restriction levels are defined in the UMLS License Agreement.*
A copy of the License Agreement is provided in the Appendix

Customizing with MRCON - Examples:

- Select or exclude sets of concepts by *CUI* identified elsewhere
- Select or exclude a language by *LAT*
- Select or exclude by Term Status TS, e.g. :

  TS="P" Preferred;

  TS="S" Synonym

  TS="s" Supressible Synonym
- Use Preferred Terms *(TS="P")* and preferred form *(STT="PF")*
- Select or exclude by Least Restriction Level (*LRL*) , e.g.

  No additional restrictions *LRL =0*

```
CUI      |LAT|TS|LUI    |STT|SUI      |STR   | LRL
C0002871|ENG|P|L0002871|PF|S0013742|Anemia|0|
```

MRSO: Source vocabulary information

```
CUI      | LUI     |SUI      |SAB     |TTY|SCD    |SRL|
C0002871|L0002871|S0013742|LCH90   |PT|U000235|0|
C0002871|L0002871|S0013742|MSH2000|MH|D000740|0|
C0002871|L0002871|S0013742|MTH|PT |U000161   |0|
C0002871|L0002871|S0013742|PSY94   |PT|024503 |0|
```

...

CUI Unique identifier for concept
LUI Unique identifier for term
SUI Unique identifier for string
SAB Source abbreviation
TTY Term type in source
SCD Unique Identifier or code for string in source
SRL Source Restriction Level

** on this slide and many which follow, trailing spaces have been added to align columns for readability. They do not exist in the Metathesaurus files.*

Source Restriction Level *vs.* Least Restriction Level

- Source (Vocabulary) Restriction Level (SRL), in MRSO:

  0 Means there are no additional restrictions

  1 - 3 Means there are additional restrictions as defined in the License Agreement.

- Least Restriction Level (LRL), in MRCON:

The lowest restriction level (SRL) for all vocabularies providing this string

* *All Restrictions are defined and listed for each vocabulary in the UMLS License Agreement*

Customizing with MRSO - Examples:

- **Better to use MetamorphoSys to select or exclude sources**

  it removes ALL source information - demo later

- **Select or exclude a TTY, e.g.**

  TTY="PT" : Preferred term in source vocabulary;

  TTY="AB" : Abbreviation in source vocabulary;

- **Select or exclude or order by SCD, e.g.**

  ICD-9-CM codes with 2 digits before decimal are Procedures

- **Select or exclude by *SRL* (Source Restriction Level)**

```
* SABs are listed in Manual Section B.2 and the SABDOC file
  TTYs are described Manual Section B.4 and the TTYDOC file
```

MRSTY: Semantic Types

***** Don't assume you know what Semantic Types mean!***

Use Knowledge Source Server (http://umlsks.nlm.nih.gov)

•Browser display for STY "Finding":

```
Semantic Types
  |
  Entity
    |
    Conceptual Entity
      |
      Finding  ...
        |
        Laboratory or Test Result
        Sign or Symptom
```

•Definitions: ...
•See a list of Metathesaurus concepts with Semantic type: "Finding".
•Relations and Related Types for "Finding".
•View records for Finding in relational format

* ***hint:*** *Tree numbers in SRDEF can be useful for walking hierarchies*

MRSTY: Semantic Types

```
 CUI    |TUI |STY|
C0002871|T047|Disease or Syndrome|
C0004057|T109|Organic Chemical|
C0004057|T121|Pharmacologic Substance|
```

CUI Unique identifier of concept
TUI Unique identifier of Semantic type
STY Semantic type, defined in the Semantic Network.

Customizing with MRSTY - Examples:

- Select or exclude by Semantic Type or sets of Types

- Use Knowledge Source Server to see definitions, hierarchy, examples: **http://umlsks.nlm.nih.gov**

e.g. select or exclude all Laboratory or Test Result

```
CUI     |TUI |STY|
C0005798|T034|Laboratory or Test Result|
```

MRREL: Most Relationships

CUI1	*REL*	*CUI2*	*RELA*	*SAB*	*SL*	*MG*
C0002871	CHD	C0002891	isa	MSH2000	MTH	
C0002871	RB	C0221016		MTH	MTH	
C0002871	RL	C0002886	mapped_to	SNMI98	SNMI98	
C0000665	RN	C0612003	mapped_to	MSH2000	MSH2000	G

CUI1 Unique identifier of first concept
REL Relationship of second to first concept
CUI2 Unique identifier of second concept
RELA Relationship attribute
SAB Abbreviation of the source of relationship
SL Source of relationship labels
MG Machine-generated and unverified indicator (optional)

MRREL: Most Relationships

```
CUI1     |REL|CUI2     |RELA|SAB     |SL  |MG|
C0002871|CHD|C0002891|isa  |MSH2000|MTH||
```

Anemia Anemia , Neonatal

"Neonatal Anemia" is child of "Anemia" (REL = CHD)
"Neonatal Anemia" is_a "Anemia" (RELA= isa)

SAB - Source of REL is MeSH;
SL - Source of labels (type of relationship) is MTH
MG - Not Machine generated.

MRREL: Most Relationships

Picking the Relationships you want:

RELDOC file, A bonus file available on http://umlsinfo.nlm.nih.gov

SAB	SL	REL	RELA	Count
ICD2000	ICD2000	CHD		19579
ICD2000	ICD2000	PAR		19579
MSH2000	MSH2000	CHD	connected_to	43
MTH	MTH	RB	ingredient_of	112
MTH	MTH	RB		157580
RAM99	RAM99	RO	clinically_similar	8200
UWDA99	UWDA99	CHD	isa	16591
UWDA99	UWDA99	CHD	part_of	7697

Customizing with MRREL - Examples:

- Select or exclude RELS, RELAs, and SAB or SL
  use *RELDOC* or determine suitability empirically

- Use selected RELS and RELAs to aggregate
  e.g., all ‘clinically_similar’

- Combine with other criteria
  e.g., selected RELs to concepts with selected Semantic Types

MRSAT: Simple Attributes

```
CUI     |LUI     |SUI     |SCD       |ATN|SAB    |ATV     |
C0002871|L0002871|S0013742|D000740   |MN |MSH2000|C15.378.71|
C0002871|L0002871|S0013742|D000740   |TH |MSH2000|POPLINE (1994)|
C0002871|L0002871|S0414880|208/04453|SOS|PDQ99  |secondary related
                                                  condition|
```

CUI Unique identifier for concept or, for LT element only, LUI for term
LUI Unique identifier for term (optional)
SUI Unique identifier for string (optional)
SCD Unique identifier or code for entry in the source of the attribute
ATN Attribute name *(see manual, ATNDOC file)*
SAB Abbreviation of the source of the attribute
ATV Attribute value *(see manual, ATNDOC file)*

**note: few attribute values exceed 2,200 characters.*

Customizing with MRSAT - Examples:

- **Wide variety of Attributes, many very detailed, described in manual or in *ATNDOC file* available on umlsinfo; e.g.:**

SAB	*ATN*	*Count*	
MSH2000	SOS	88355	-- Details of scope of concept in MeSH
MTH	AM	11056	-- Ambiguous string indicator
MTH	ST	730155	-- Status - ‘R’eviewed or ‘U’nreviewed

- **Attributes may be attached to CUIs, LUIs, or SUIs**
- **Choose the attributes you need!**

Customizing with MRSAT, continued ...

You may also identify *sets* of concepts by attribute

- **For example, to Select or exclude by Attribute for *Reviewed Status***
  - ATN “Status” = ST,
  - Attribute Value ATV = ‘R’ (Values are ‘R’eviewed or ‘U’nreviewed)

- **All Unreviewed Concepts (which are MeSH Supplementary concepts) have *ATN ‘ST’ and ATV ‘U’***

- **To use or exclude these concepts, sselect their CUIs from the MRSAT rows with *ATN ‘ST’ and ATV ‘U’***

Select Other data you need:

- MRDEF: Definitions
- MRCXT: Contexts
- MRLO: Locator
- MRATX: Mapping one vocabulary to Another
- MRCOC: Co-occurrences
- Indexes:
  - **MRXW.ENG, MRXW.FRE, Other languages ...**
  - **MRXNW.ENG**
  - **MRXNS.ENG**

MRDEF: Definitions

```
 CUI      |SAB    | DEF
C0002871|MSH2000|A reduction in the number of circulating
                  erythrocytes or in the quantity of
                  hemoglobin.|
```

CUI	Unique identifier for concept
SAB	Abbreviation of the source of the definition
DEF	Definition

*** Note - there are nearly 32,000 definitions.*
Some definitions are very long, up to 6 kb

MRCXT: Contexts - Examples:

```
MSH2000
  Diseases (MeSH Category) [C]
     Virus Diseases [C2]
      RNA Virus Infections [C2.782]
       Arenaviridae Infections [C2.782.82]
        Lassa Fever [C2.782.82.545]
 MSH2000
    Diseases (MeSH Category) [C]
     Virus Diseases [C2]
      RNA Virus Infections [C2.782]
       Hemorrhagic Fevers, Viral [C2.782.417]
        Lassa Fever [C2.782.417.505]
 ICD10
    Certain infectious and parasitic diseases
     Arthropod-borne viral fevers and viral haemorrhagic fevers
      Arenaviral haemorrhagic fever
       Lassa fever [A96.2]
```

Using MRCXT: Contexts

-- See the documentation!

```
CUI     |SUI     |SAB    |SCD   |CXN|CXL|RNK|CXS|CUI2    |HCD|REL|XC|
C0002871|S0013742|MSH2000|D000740|1|ANC|1|MeSH|C0220876|||||
C0002871|S0013742|MSH2000|D000740|1|ANC|2|Diseases (MeSH Category)|C0012674|C|||
C0002871|S0013742|MSH2000|D000740|1|ANC|3|Hemic and Lymphatic Diseases|C0018981|C15|||
C0002871|S0013742|MSH2000|D000740|1|ANC|4|Hematologic Diseases|C0018939|C15.378|isa||
C0002871|S0013742|MSH2000|D000740|1|CCP||Anemia|C0002871|C15.378.71|isa|+|
C0002871|S0013742|MSH2000|D000740|1|CHD||Anemia, Aplastic|C0002874|C15.378.71.85|isa|+|
C0002871|S0013742|MSH2000|D000740|1|SIB||Blood Protein Disorders|C0005830|C15.378.147|isa|+|
C0002871|S0013742|MSH2000|D000740|1|SIB||Bone Marrow Diseases|C0005956|C15.378.190|isa|+|
```

CUI	Unique identifier of concept
SUI	Unique identifier for string used in this context
SAB	Source abbreviation
SCD	Unique Identifier or code for string in that source
CXN	The context number (to distinguish multiple contexts , same source, same SUI).
CXL	Context member label, e.g.., ANC for ancestor of this concept\
RNK	For rows with a CXL value of ANC, the rank of the ancestors
CXS	String for context member.
CUI2	Unique concept identifier of context member (may be empty)
HCD	Hierarchical number or code of context member in this source (optional).
REL	Relationship of concept to parent, if applicable and known.
XC	Plus(+) sign indicating that the concept has children (see manual)

MRCXT: Contexts - Caveats:

MRCXT hierarchies are viewable and incomplete.

- **MRCXT was created to provide *reasonably viewable* hierarchies.**
- **Many vocabularies do not have full, principled hierarchies**
- **Many hierarchies have too many children or siblings to be viewable**

MRREL relationships are authoritative and complete.

MRLO: Locator

CUI	ISN	FR	UN	SUI	SNA	SOUI
C0002871	DXPLAIN			S0352787		
C0002871	MEDLINE(1990-1995)	1666	*CITATIONS	S0013742		
C0002871	MEDLINE(1996-Fall 1998)	707	*CITATIONS	S0013742		
C0002871	PDQ			S0414880		
C0002871	QMR			S0013787		

CUI Unique identifier of concept

ISN Name of information source or database in which concept appears

FR Frequency count of number of occurrences of concept in the information source (optional)

UN Meaning of frequency (optional)

SUI Unique identifier of string if name used in information source appears in MRCON (optional)

SNA Actual name that occurs in the information source if not otherwise present in the Metathesaurus (optional)

SOUI Unique identifier of record in which the concept appears in source (optional)

MRATX: Mapping one vocabulary to another

```
CUI      | SAB   |REL|  ATX   |
C0494958|MSH2000|S|<Kidney Failure, Acute> AND <Kidney Papillary
                    Necrosis>|
```

"Acute renal failure with medullary necrosis"
in ICD-9-CM and ICD-10

CUI Unique identifier of concept to which the expression is related
SAB Abbreviation of source of terms *in expression*
REL Relationship of meaning of expression to main concept
ATX Associated expression (with complex internal syntax)

Most Metathesaurus concept do not have ATXs
Most are *FROM* ICD-9-CM *TO* MeSH
Need feedback on usefulness

MRCOC: Co-occurrences

Co-occurrence information:

Most are MEDLINE co-occurrences

Also includes:

Patient record co-occurrences from CCPSS, the *Canonical Clinical Problem Statement System* from Vanderbilt University

Knowledge base co-occurrences from AI Rheum, a Rheumatology expert system

MRCOC: Co-occurrences

(see manual)

CUI1	*CUI2*	*SOC*	*COT*	*COF*	*COA*
C0002871	C0000727	MBD	L	1	BL=1,ET=1
C0002871	C0000737	MBD	L	1	ET=1
C0002871	C0000772	MBD	L	2	CN=2
C0000731	C0000737	CCPSS99	PP	10	

CUI1 Unique identifier of first concept

CUI2 Unique identifier of second concept

SOC Abbreviation of the Source of co-occurrence information if applicable

COT Type of co-occurrence

COF Frequency of co-occurrence, if applicable

COA Attributes of co-occurrence, if applicable

Indexes:

MRXW.ENG, MRXW.FRE, Other languages … lowercased words
MRXNW.ENG … lowercased *normalized* words
MRXNS.ENG … lowercased *normalized* strings

Example:

MRCON:

C0026927|ENG|P|L0026927|PF|S1396311|Mycobacteria, Atypical|0|

MRXW.ENG:

ENG|atypical|C0026927|L0026927|S1396311|

ENG|mycobacteria|C0026927|L0026927|S1396311|

MRXNW.ENG :

ENG|atypical|C0026927|L0026927|S1396311|

ENG|mycobacterium|C0026927|L0026927|S1396311|

MRXNS.ENG :

ENG|atypical mycobacterium|C0026927|L0026927|S1396311|

Using Indexes: The Specialist Lexical Programs

- Very powerful, standardized tools are available as part of the UMLS Specialist Lexicon and Lexical Programs
- Source code and binary programs for Unix and NT are included
- “norm” will process your terms and queries to match the Meta indexes
- “lvg” does much more …
- All are included in the UMLS distribution

Recap: How to Customize the Metathesaurus

- Limit vocabularies, languages
  - MetamorphoSys does this best (demo is next!)
- Use Semantic Types
- Limit and aggregate by Relationships
  - MRREL
- Add and Use Suppressibility
  - MRCON “TS”
- Change Naming Precedence -
  - MetamorphoSys - GUI or custom configuration
- Select by attribute e.g. ST
  - MRSAT
- Use other MR Files as needed …
- Adding “local” terminology (after demo)

Outline of Tutorial

- Why customize - *Betsy Humphreys*
- How to customize - *Bill Hole*
- *A tool to help you customize - Laura Roth and Suresh Srinivasan*
- Adding “local” terminology - *Bill Hole*

MetamorphoSys

- A tool distributed for use with the UMLS Knowledge Sources
  - Already present in UMLS distribution
- Multi-platform Java software
- Creates a customized version of the Metathesaurus

Why Use MetamorphoSys?

- Exclude vocabularies as required by the UMLS License Agreement
  - Default is to select only vocabularies that have no special restrictions (category zero)
  - Example: Remove SNOMED if user does not have an agreement with CAP for use
- Exclude vocabularies not needed for users' specific purposes
  - Example: Nursing vocabularies

Use (cont'd)

- Alter "preferred name" precedence
  - Precedence is determined by MRRANK table supplied by NLM
  - Highest precedence source provides the "preferred name" of the concept name
  - User may want the preferred names of concepts to be other than the default provided
  - Example: User wants Read vocabulary to be highest precedence source

Use (cont'd)

- Add suppressibility to any Source-TTY combination
  - Metathesaurus already marks some combinations as suppressible by setting TS=s in MRCON
  - Suppressibility allows users to program their applications to remove certain groups of terms that may not be suitable
  - Example: Remove LOINC for Natural Language Processing purposes

How does MetamorphoSys work?

- What it does: removes all information from MR* files that is supplied by the excluded vocabularies
  - Allows added source-termgroup suppressibility
  - Allows altered “preferred name” precedence
- What results: A full Metathesaurus of the remaining vocabularies

How to Use MetamorphoSys

- Machine requirements
- Configuration interface
- Demo on subset of Metathesaurus

Machine Requirements

- A minimum of 256 MB of physical memory, as well as 8 GB recommended free disk space
  - Full UMLS distribution needs to be present
- Can be run on: Sun Solaris, Windows NT, Linux and Windows 98

Configuration Interface

- User sees a Java graphical user interface
- Interactively prompts for information
- Started by the MetamorphoSys program once UMLS distribution has been unpacked
  - Found in UMLS2000/META/METAMSYS/ directory
  - MetamorphoSys.sh in the UNIX environment
  - MetamorphoSys.bat in Windows

Configuration

- Simple to use

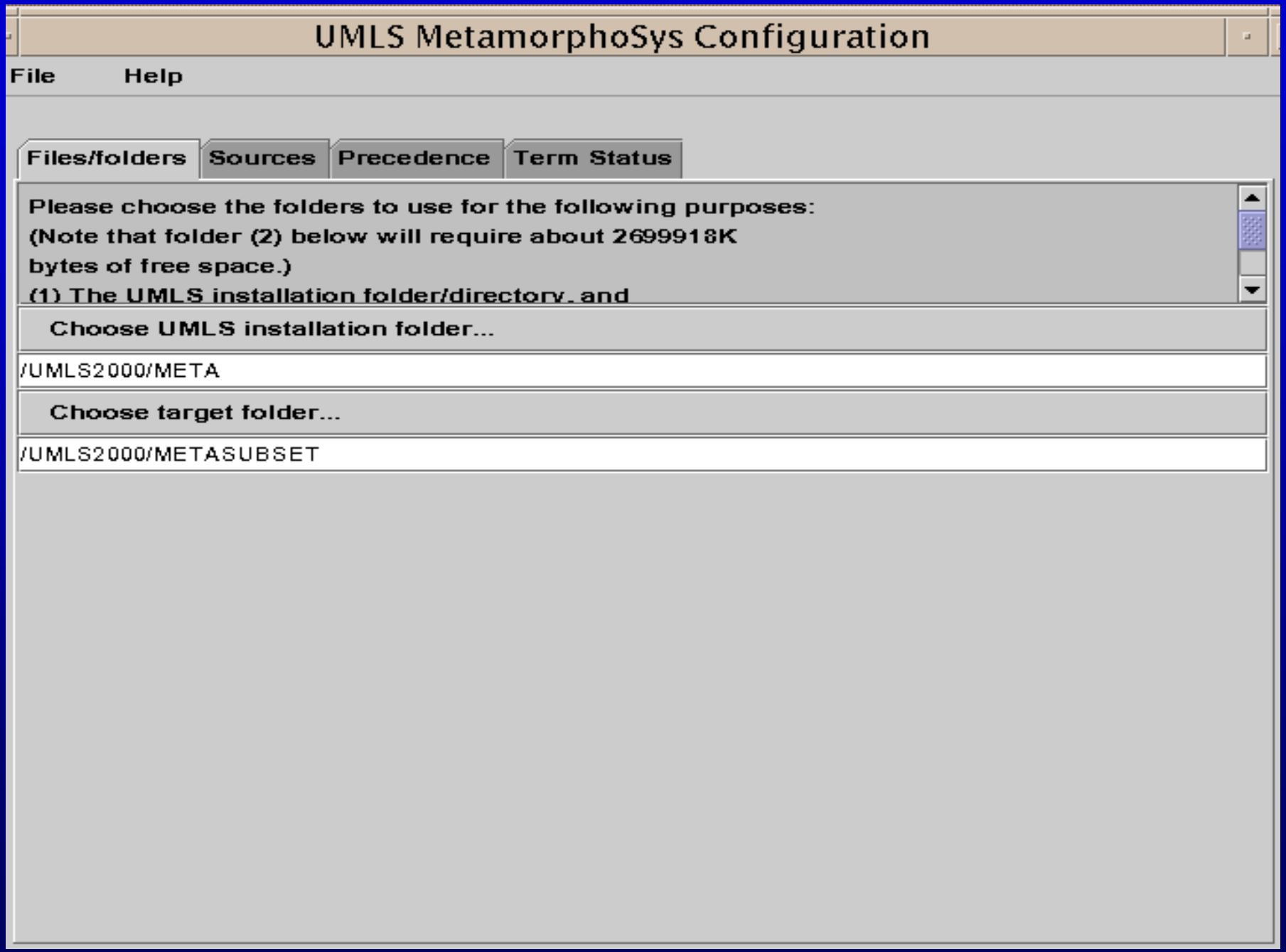
- No configuration needed to exclude all non-zero category vocabularies
  - Default is a Metathesaurus of just category zero vocabularies, i.e., to exclude those with category 1, 2 or 3 restrictions (see License Agreement in Appendix for details)

Configuration

- Four tabs present in the MetamorphoSys interface
  - Files/folders
  - Sources
  - Precedence
  - Term Status

Files/Folders

- Indicate where UMLS distribution is present on user's machine
- Indicate where the customized Metathesaurus should go once MetamorphoSys is finished
- Default directories are provided


UMLS MetamorphoSys Configuration
File
Help
Files/folders
Sources
Precedence
Term Status
Please choose the folders to use for the following purposes:
(Note that folder (2) below will require about 2699918K
bytes of free space.)
(1) The UMLS installation folder/directory, and
Choose UMLS installation folder...
/UMLS2000/META
Choose target folder...
/UMLS2000/METASUBSET

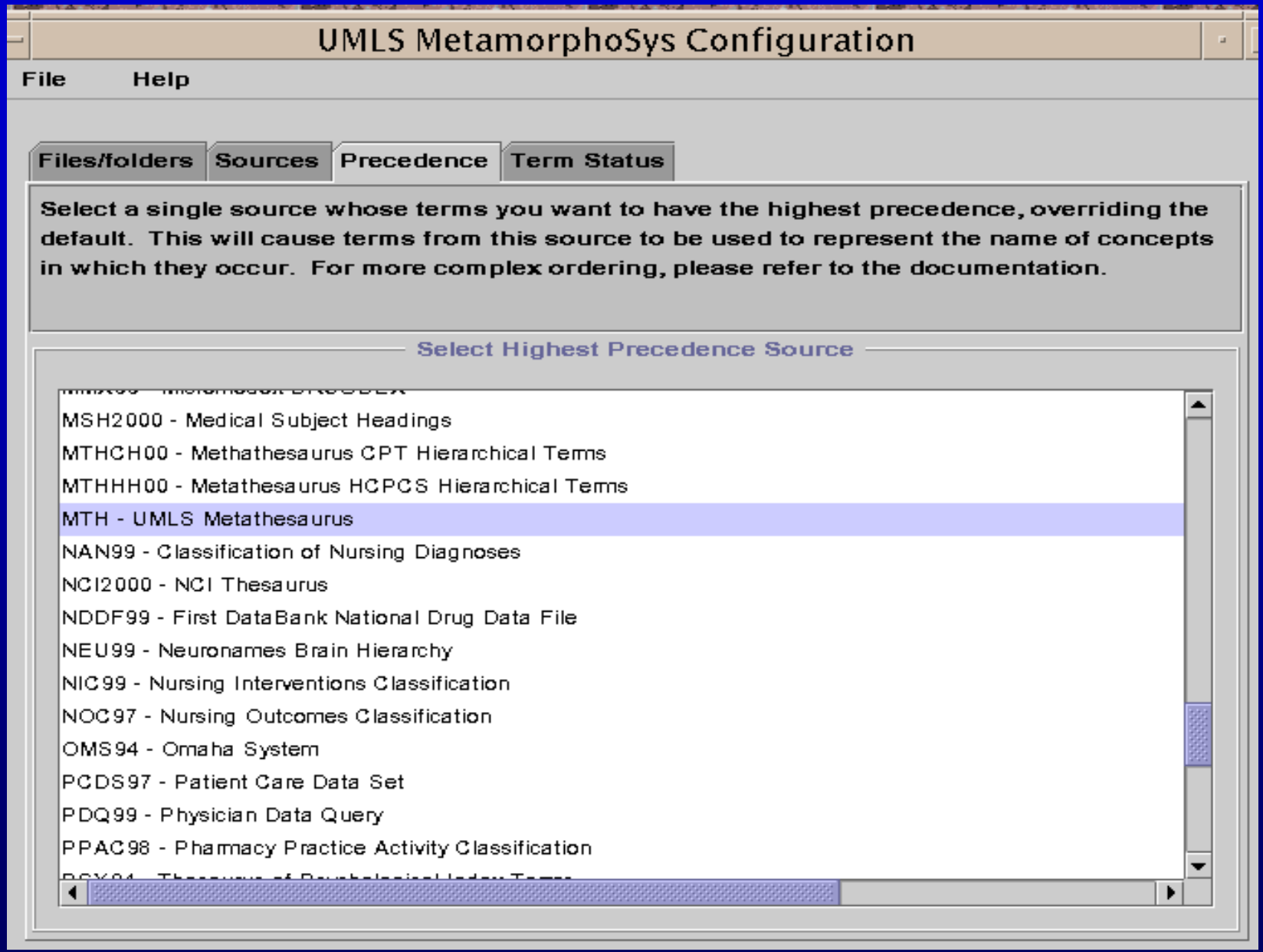
Sources

- All sources are alphabetically listed
- Ones highlighted are the ones to be excluded
- Can change to include or exclude any vocabulary

Sources Configuration

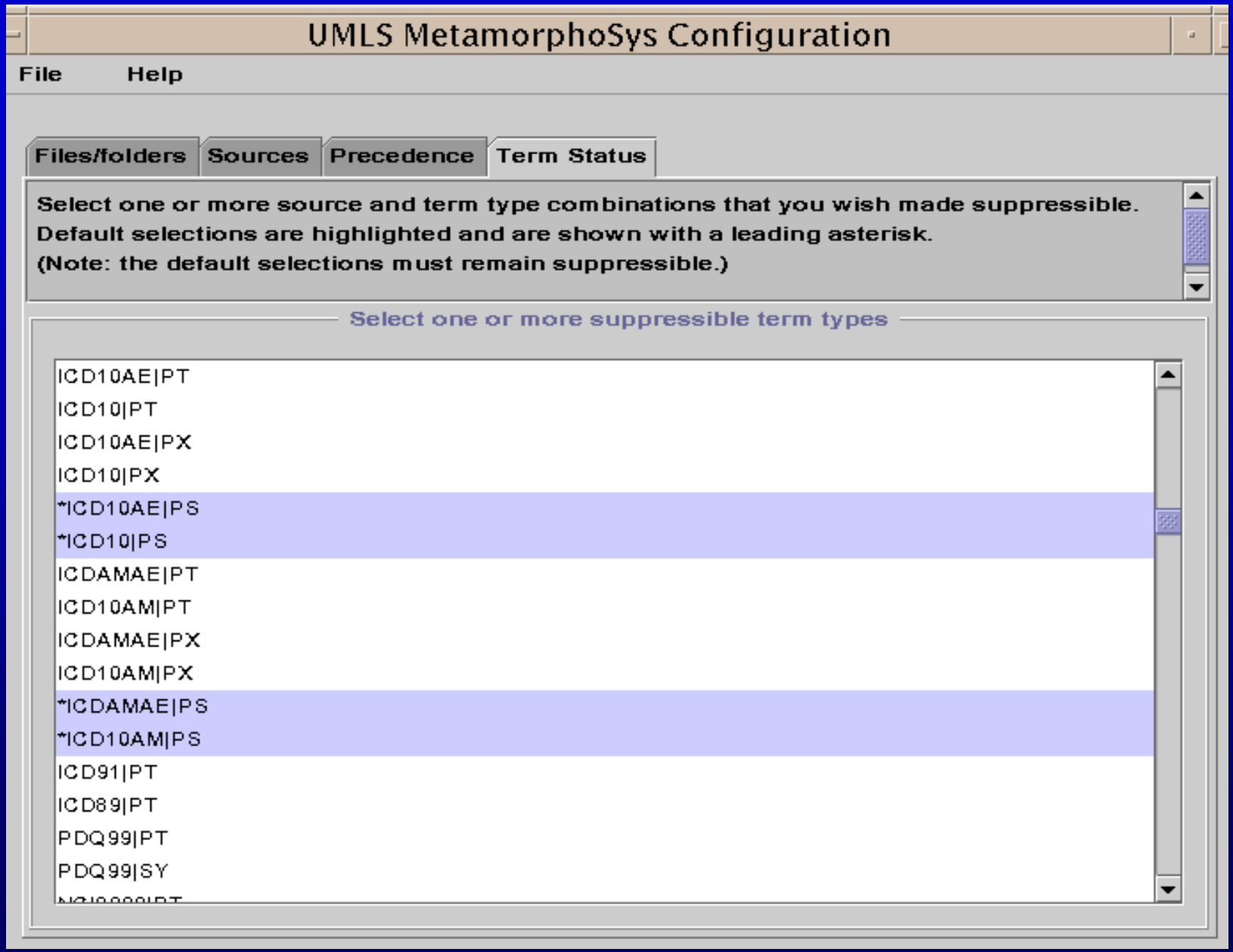
Precedence

- MTH source is the default highest precedence source
- Sources are arranged in list alphabetically and not by precedence
- Highlighting a source will select it as the highest precedence source
  - Only one source can be chosen


UMLS MetamorphoSys Configuration
File
Help
Files/folders
Sources
Precedence
Term Status
Select a single source whose terms you want to have the highest precedence, overriding the default. This will cause terms from this source to be used to represent the name of concepts in which they occur. For more complex ordering, please refer to the documentation.
Select Highest Precedence Source
MSH2000 - Medical Subject Headings
MTHCH00 - Methathesaurus CPT Hierarchical Terms
MTHHH00 - Metathesaurus HCPCS Hierarchical Terms
MTH - UMLS Metathesaurus
NAN99 - Classification of Nursing Diagnoses
NCI2000 - NCI Thesaurus
NDDF99 - First DataBank National Drug Data File
NEU99 - Neuronames Brain Hierarchy
NIC99 - Nursing Interventions Classification
NOC97 - Nursing Outcomes Classification
OMS94 - Omaha System
PCDS97 - Patient Care Data Set
PDQ99 - Physician Data Query
PPAC98 - Pharmacy Practice Activity Classification

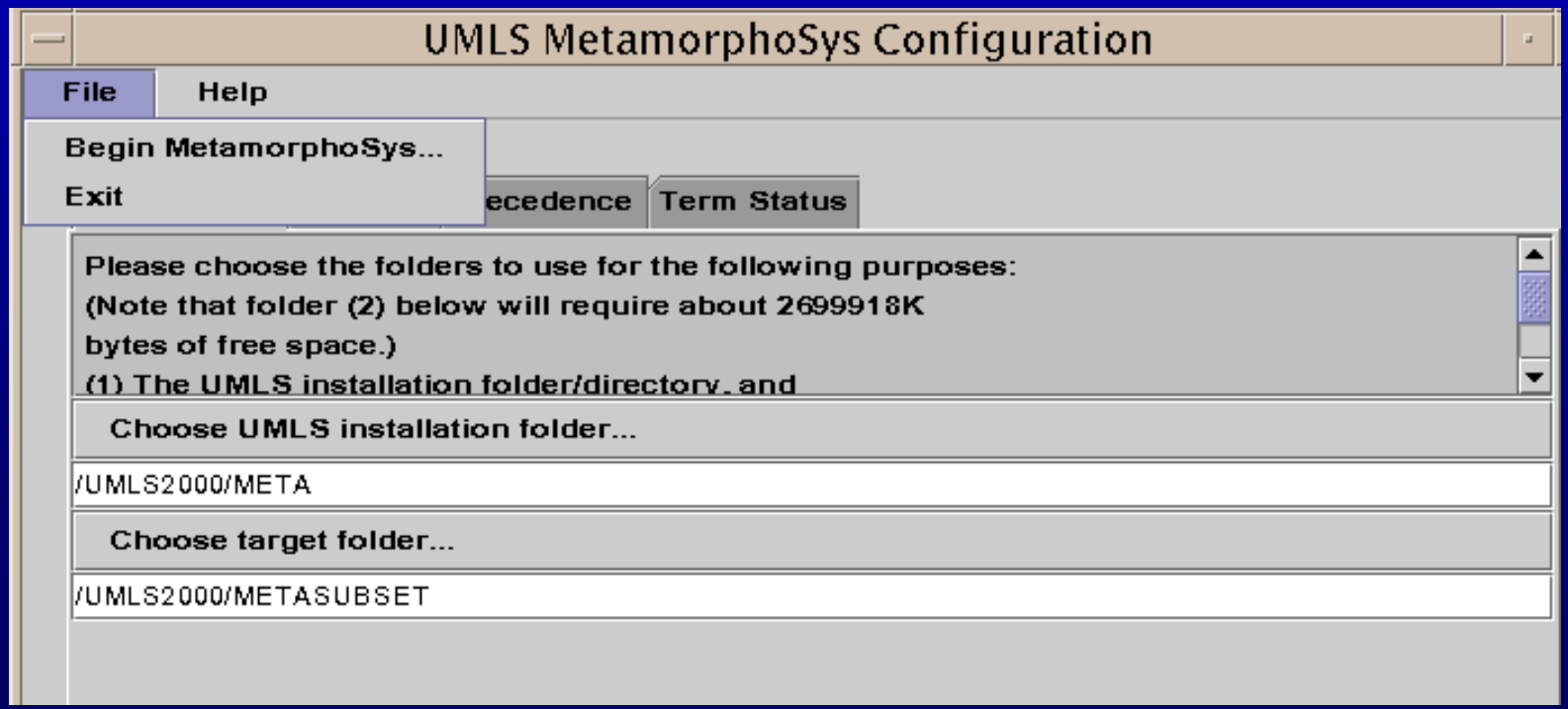
Term Status

- Used to add suppressibility
- All Source-TTY combinations that are suppressible are highlighted
- Cannot change these to be non-suppressible in interface
- New combinations can be highlighted to make them suppressible


UMLS MetamorphoSys Configuration
File
Help
Files/folders
Sources
Precedence
Term Status
Select one or more source and term type combinations that you wish made suppressible.
Default selections are highlighted and are shown with a leading asterisk.
(Note: the default selections must remain suppressible.)
Select one or more suppressible term types
ICD10AE|PT
ICD10|PT
ICD10AE|PX
ICD10|PX
*ICD10AE|PS
*ICD10|PS
ICDAMAE|PT
ICD10AM|PT
ICDAMAE|PX
ICD10AM|PX
*ICDAMAE|PS
*ICD10AM|PS
ICD91|PT
ICD89|PT
PDQ99|PT
PDQ99|SY

Running MetamorphoSys

- Once configuration is defined, a simple file selection starts subsetting of Metathesaurus



Output

- Once started, a message tells user that the program has started and the interface disappears
- Program exits automatically upon completion of data subsetting
- Your customized Metathesaurus will be present in defined directory
  - UMLS2000/METASUBSET by default

Outline of Tutorial

- Why customize - *Betsy Humphreys*
- How to customize - *Bill Hole*
- *A tool to help you customize*

  - Laura Roth and Suresh Srinivasan
- Adding “local” terminology - *Bill Hole*

MetamorphoSys Output

Distributed 2000 Metathesaurus Data for C0221233

Distributed Data for C0221233...

MRCON

```
C0221233|ENG|P|L0294378|PF|S0625803|Cyclic edema|3|
C0221233|ENG|P|L0294378|VW|S0368795|EDEMA, CYCLIC|0|
C0221233|ENG|P|L0294378|VW|S0369867|Edema, cyclic|3|
C0221233|ENG|S|L0298620|PF|S0368798|EDEMA, PERIODIC|0|
C0221233|ENG|S|L0298620|VC|S0369871|Edema, periodic|0|
C0221233|ENG|S|L0298620|VW|S0395649|Periodic, edema|0|
C0221233|ENG|S|L0803934|PF|S0852190|Cyclic oedema|3|
```

MRSO

```
C0221233|L0294378|S0368795|DXP94|SY|NOCODE|0|
C0221233|L0294378|S0369867|SNM2|PT|M-36680|3|
C0221233|L0294378|S0625803|SNMI98|PT|M-36480|3|
C0221233|L0298620|S0368798|DXP94|SY|NOCODE|0|
C0221233|L0298620|S0369871|ICD91|IT|995.1|0|
C0221233|L0298620|S0395649|ICD91|IT|995.1|0|
C0221233|L0803934|S0852190|SNMI98|SY|M-36480|3|
```

Distributed Data for C0221233...

MRREL

```
C0013604|CHD|C0221233||SNM2|SNM2||
C0013604|CHD|C0221233||SNMI98|SNMI98||
C0221233|PAR|C0013604||SNM2|SNM2||
C0221233|PAR|C0013604||SNMI98|SNMI98||
C0221233|RB|C0238094||MTH|MTH||
C0238094|RN|C0221233||MTH|MTH||
```

MRSAT

```
C0221233|||DA|MTH|19950417|
C0221233|||MR|MTH|20000101|
C0221233|||ST|MTH|R|
```

1. Default Subset for C0221233

Removing all restricted sources

Default Subset for C0221233 (Before)

MRCON

```
C0221233|ENG|P|L0294378|PF|S0625803|Cyclic edema|3|
C0221233|ENG|P|L0294378|VW|S0368795|EDEMA, CYCLIC|0|
C0221233|ENG|P|L0294378|VW|S0369867|Edema, cyclic|3|
C0221233|ENG|S|L0298620|PF|S0368798|EDEMA, PERIODIC|0|
C0221233|ENG|S|L0298620|VC|S0369871|Edema, periodic|0|
C0221233|ENG|S|L0298620|VW|S0395649|Periodic, edema|0|
C0221233|ENG|S|L0803934|PF|S0852190|Cyclic oedema|3|
```

MRSO

```
C0221233|L0294378|S0368795|DXP94|SY|NOCODE|0|
C0221233|L0294378|S0369867|SNM2|PT|M-36680|3|   ←
C0221233|L0294378|S0625803|SNMI98|PT|M-36480|3| ←
C0221233|L0298620|S0368798|DXP94|SY|NOCODE|0|
C0221233|L0298620|S0369871|ICD91|IT|995.1|0|
C0221233|L0298620|S0395649|ICD91|IT|995.1|0|
C0221233|L0803934|S0852190|SNMI98|SY|M-36480|3| ←
```

Default Subset for C0221233 (After)

MRCON

```
C0221233|ENG|P|L0294378|PF|S0368795|EDEMA, CYCLIC|0|
C0221233|ENG|S|L0298620|PF|S0368798|EDEMA, PERIODIC|0|
C0221233|ENG|S|L0298620|VC|S0369871|Edema, periodic|0|
C0221233|ENG|S|L0298620|VO|S0395649|Periodic, edema|0|
```

MRSO

```
C0221233|L0294378|S0368795|DXP94|SY|NOCODE|0|
C0221233|L0298620|S0368798|DXP94|SY|NOCODE|0|
C0221233|L0298620|S0369871|ICD91|IT|995.1|0|
C0221233|L0298620|S0395649|ICD91|IT|995.1|0|
```

MRREL

```
C0221233|RB|C0238094||MTH|MTH||
C0238094|RN|C0221233||MTH|MTH||
```

2. Altering Precedence for C0221233

Making ICD91 the highest precedence source

Changing Precedence for C0221233 (Before)

MRCON

```
C0221233|ENG|P|L0294378|PF|S0625803|Cyclic edema|3|
C0221233|ENG|P|L0294378|VW|S0368795|EDEMA, CYCLIC|0|
C0221233|ENG|P|L0294378|VW|S0369867|Edema, cyclic|3|
C0221233|ENG|S|L0298620|PF|S0368798|EDEMA, PERIODIC|0|
C0221233|ENG|S|L0298620|VC|S0369871|Edema, periodic|0|
C0221233|ENG|S|L0298620|VW|S0395649|Periodic, edema|0|
C0221233|ENG|S|L0803934|PF|S0852190|Cyclic oedema|3|
```

MRSO

```
C0221233|L0294378|S0368795|DXP94|SY|NOCODE|0|
C0221233|L0294378|S0369867|SNM2|PT|M-36680|3|
C0221233|L0294378|S0625803|SNMI98|PT|M-36480|3|
C0221233|L0298620|S0368798|DXP94|SY|NOCODE|0|
C0221233|L0298620|S0369871|ICD91|IT|995.1|0|
C0221233|L0298620|S0395649|ICD91|IT|995.1|0|
C0221233|L0803934|S0852190|SNMI98|SY|M-36480|3|
```

Changing Precedence for C0221233 (After)

MRCON

```
C0221233|ENG|P|L0298620|PF|S0369871|Edema, periodic|0|
C0221233|ENG|P|L0298620|VC|S0368798|EDEMA, PERIODIC|0|
C0221233|ENG|P|L0298620|VO|S0395649|Periodic, edema|0|
C0221233|ENG|S|L0294378|PF|S0625803|Cyclic edema|3|
C0221233|ENG|S|L0294378|VO|S0368795|EDEMA, CYCLIC|0|
C0221233|ENG|S|L0294378|VO|S0369867|Edema, cyclic|3|
C0221233|ENG|S|L0803934|PF|S0852190|Cyclic oedema|3|
```

MRSO

```
C0221233|L0294378|S0368795|DXP94|SY|NOCODE|0|
C0221233|L0294378|S0369867|SNM2|PT|M-36680|3|
C0221233|L0294378|S0625803|SNMI98|PT|M-36480|3|
C0221233|L0298620|S0368798|DXP94|SY|NOCODE|0|
C0221233|L0298620|S0369871|ICD91|IT|995.1|0|
C0221233|L0298620|S0395649|ICD91|IT|995.1|0|
C0221233|L0803934|S0852190|SNMI98|SY|M-36480|3|
```

3. Adding Suppressibility for C0221233

Suppressing SNOMED synonyms (SNMI98-SY) strings

Adding Suppressibility for C0221233 (Before)

MRCON

```
C0221233|ENG|P|L0294378|PF|S0625803|Cyclic edema|3|
C0221233|ENG|P|L0294378|VW|S0368795|EDEMA, CYCLIC|0|
C0221233|ENG|P|L0294378|VW|S0369867|Edema, cyclic|3|
C0221233|ENG|S|L0298620|PF|S0368798|EDEMA, PERIODIC|0|
C0221233|ENG|S|L0298620|VC|S0369871|Edema, periodic|0|
C0221233|ENG|S|L0298620|VW|S0395649|Periodic, edema|0|
C0221233|ENG|S|L0803934|PF|S0852190|Cyclic oedema|3|
```

MRSO

```
C0221233|L0294378|S0368795|DXP94|SY|NOCODE|0|
C0221233|L0294378|S0369867|SNM2|PT|M-36680|3|
C0221233|L0294378|S0625803|SNMI98|PT|M-36480|3|
C0221233|L0298620|S0368798|DXP94|SY|NOCODE|0|
C0221233|L0298620|S0369871|ICD91|IT|995.1|0|
C0221233|L0298620|S0395649|ICD91|IT|995.1|0|
C0221233|L0803934|S0852190|SNMI98|SY|M-36480|3|
```

Adding Suppressibility for C0221233 (After)

MRCON

```
C0221233|ENG|P|L0294378|PF|S0625803|Cyclic edema|3|
C0221233|ENG|P|L0294378|VO|S0368795|EDEMA, CYCLIC|0|
C0221233|ENG|P|L0294378|VO|S0369867|Edema, cyclic|3|
C0221233|ENG|S|L0298620|PF|S0368798|EDEMA, PERIODIC|0|
C0221233|ENG|S|L0298620|VC|S0369871|Edema, periodic|0|
C0221233|ENG|S|L0298620|VO|S0395649|Periodic, edema|0|
C0221233|ENG|s|L0803934|PF|S0852190|Cyclic oedema|3|
```

MRSO

```
C0221233|L0294378|S0368795|DXP94|SY|NOCODE|0|
C0221233|L0294378|S0369867|SNM2|PT|M-36680|3|
C0221233|L0294378|S0625803|SNMI98|PT|M-36480|3|
C0221233|L0298620|S0368798|DXP94|SY|NOCODE|0|
C0221233|L0298620|S0369871|ICD91|IT|995.1|0|
C0221233|L0298620|S0395649|ICD91|IT|995.1|0|
C0221233|L0803934|S0852190|SNMI98|SY|M-36480|3|
```

MetamorphoSys Configuration

- GUI and “Batch” Components
- Data in the “config” Subdirectory,
  e.g., sources.to.remove, prec.order
- GUI creates “Batch” config files
- Or Hand-Edit and Re-run

Why Hand-Edit?

- Better precedence control
- Saving configuration between runs
- Not for novices!
- Details are on http://umlsinfo.nlm.nih.gov

MetamorphoSys Configuration Files

File	Use	Description
Sources.to.remove	Batch	Abbreviations of all sources that will be removed – one per line;
Prec.order	Batch	Source abbreviation and term type in order of precedence; one per line. Leading asterisk implies suppressible combination
Mmsys.properties	both	Properties file; attribute=value format, e.g., begin=true

MetamorphoSys Help Files

File	Use	Description
Source.txt	GUI	Abbreviation and description of all sources
Sources.category<n>	n/a	Abbreviations of sources for each restriction level: 0, 1, 2, 3
Prec.txt	GUI	Source and term types in NLM's default order of precedence

sources.to.remove *file*

- Default is to remove all sources with restriction level>0
- Edit with any text editor - add additional source abbreviations (or remove existing ones)
- Ensure that added ones are in source.txt
- Save file prior to re-run

prec.order *file*

- Order of Source-TTY lines decides precedence (lines in decreasing order of precedence)
- Edit with any text editor to change order
- Make additional ones suppressible - by adding a leading ‘*’
- Save file prior to re-run

Expert(!) Mode Summary

- Edit sources.to.remove, prec.order and mmsys.properties as needed
- Change to installation directory
- Run the “batch” component directly
- Full details on umlsinfo

Post Subset Examples

- Removing Suppressible Synonyms
  - awk -F'|' '$3!="s"{print $1}' MRCON
  - SELECT CUI FROM MRCON WHERE TS!='s'
- Select concepts by STY
  - join -F'|' MRSTY MRCON | awk -F'|' '$2=="Disease or Syndrome"{print $1}'
  - SELECT MRCON.CUI FROM MRCON, MRSTY WHERE MRCON.CUI=MRSTY.CUI AND MRSTY.STY='Disease or Syndrome'

Notes

- Resulting STT not always accurate without LVG (see umlsinfo.nlm.nih.gov)
- Sort order - By CUI and ASCII within CUI
- Save your configuration files if needed
- Next version will be better!

Outline of Tutorial

- Why customize - *Betsy Humphreys*
- How to customize - *Bill Hole*
- A tool to help you customize - *Laura Roth and Suresh Srinivasan*
- *Adding "local" terminology - Bill Hole*

Adding Terminology

Create Distinct Unique Identifiers for *your* Terminology...

- e.g., for your concepts, use:
  ‘CA000001 …’ as CUIs instead
  of Meta’s ‘C0000001 ….’ for CUIs
- Similarly, use ‘LA000001 ...’ for LUIs
  and ‘SA000001 …’ for SUIs, as needed
- Create a table recording your UIs, and for mapping them to UMLS UIs

Adding Terminology, Continued

Which of your terms are Meta Synonyms?

- Normalize them and look for matches in the Normalized String Index (MRXNS).
- Use other sensible approaches to searching:
  - normalized word searches;
  - explore alternate naming styles and conventions
- Use Meta CUIs for Synonyms

** See "Discovering Missed Synonymy in a Large Concept-Oriented Metathesaurus" - AMIA 2000 paper (Session S7)*

Adding Terminology, Continued

Bonus - add relationships ...

- As you look for Meta Synonyms, add *relationships to Meta*

- Assign a REL and RELA to label the particular kinds of relationships you need and will use, e.g. to map or aggregate

Updating to a New Meta Release

- Repeat MetamorphoSys processing scripts from previous release
- Re-use previously found UIs for your terms to map synonyms, etc.
- Check for new synonyms of your terms which were not Meta synonyms before
- Check for any deleted CUIs and map them

Incremental updates are coming!

Same procedures will apply to the changes

- Update consists of deletes and adds only;
- Any changed concept is deleted, then added with the changes
- Any deleted concept is deleted
- Any new Concept is added

Online Resources:

WWW:

http://www.nlm.nih.gov/research/umls/
http:// umlsks.nlm.nih.gov
http:// umlsinfo.nlm.nih.gov

E-mail:

umlsmeta@nlm.nih.gov
umlslex@nlm.nih.gov
umlsnet@nlm.nih.gov
umlsks@nlm.nih.gov

umls-users listserv:

To subscribe to the listserv, send a message to

listserv@nlm.nih.gov

which includes the following line:

subscribe umls-users

To post a message to the umls-users listserv AFTER subscribing:
send email to: umls-users@nlm.nih.gov