INTRODUCTION TO CONTROLLED VOCABULARIES AND ONTOLOGIES

Illhoi Yoo, Ph.D.
Dept. of Health Management and Informatics
School of Medicine
University of Missouri-Columbia

Agenda

• Introduction to Controlled vocabularies
  – Principles of vocabulary control
  – 4 different types of CVs
  – Terms and Relationships
• Ontologies
  • Medical Subject Headings (MeSH)
    – Descriptors and MeSH Tree
  • Unified Medical Language System (UMLS)
    – MetaThesaurus
    – Semantic Network

Controlled Vocabularies

• Guidelines for the construction, Format, and Management of Monolingual Controlled Vocabularies (ANSI/NISO Z39.19-2005)
  – Provides guidelines for the selection, formulation, organization, and display of terms making up a CV
  – Abstract:
    • presents guidelines and conventions for the contents, display, construction,...
    • CVs are used for the representation of content objects in knowledge organization systems including lists, synonym rings, taxonomies, and thesauri.
    • The primary purpose of vocabulary control is to achieve consistency in the description of content objects and to facilitate retrieval.

CV: Controlled Vocabularies

• Four principles of vocabulary control:
  – Eliminating ambiguity
    • Each term has only one meaning and only one term (i.e., heading) can be used to represent a given concept (or entity).
    • E.g., for cold, common cold and cold temperature
  – Controlling synonyms
    • Each concept is represented by a single preferred term (heading)
    • A set of synonyms should be provided for each concept
    • E.g., Lung neoplasms; pulmonary neoplasm; lung cancer; pulmonary cancer; cancer of lung, etc
  – Establishing relationships among terms
    • Various types of semantic relationships may be identified among the terms.
    – {Equality, hierarchical, associative} relationships
  – Testing and validation of terms (terms are changing!)
CV: Controlled Vocabularies

- 4 different types of CVs
  - **List**
    - a limited set of terms arranged as a simple list
    - E.g., a list of the US states
  - **Synonym ring**
    - A set of terms that are considered equivalent for the purposes of retrieval
    - Synonym rings are mainly used for document retrieval
    - E.g., Query expansion

- **Taxonomy**
  - Consists of preferred terms (headings)
  - No synonym for each concept
  - All terms are connected in a (poly)hierarchy
- **Thesaurus**
  - The most typical and complex form of CVs
  - Provides synonyms for each concept
  - For use in indexing and searching applications
    - Requiring **preferred terms** and **synonyms**
  - Provides the richest structure

### CV: Terms

- **Choice of terms**
  - The most fundamental factor in
  - Many issues to be considered
    - The information space or domain
      - You should ascertain whether an existing CV covers the same or an overlapping domain of knowledge.
      - Existing CVs can serve as a useful starting point
    - Literary, user, and organizational warrants *Why?*
      - **Literary** warrant: What terms are used in dictionaries, textbooks, journals, etc
      - **User** warrant: What terms users actually use in related IR systems (e.g., PubMed log data)
        - *Diabetes* is not found in the MeSH vocabulary
      - **Organizational** warrant: specific forms of terms that are preferred by organizations

### Property List Synonym Ring Taxonomy Thesaurus

<table>
<thead>
<tr>
<th>Property</th>
<th>List</th>
<th>Synonym Ring</th>
<th>Taxonomy</th>
<th>Thesaurus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preferred terms</td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Entry terms</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(synonyms)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Equivalence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>relationship</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hierarchy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>relationship</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scope note</td>
<td></td>
<td></td>
<td></td>
<td><strong>Optional</strong></td>
</tr>
<tr>
<td>History note</td>
<td></td>
<td></td>
<td></td>
<td><strong>Optional</strong></td>
</tr>
</tbody>
</table>
CV: Terms

• Choice of terms (con’t)
  – Many issues to be considered (con’t)
    • Specificity or granularity of the terms
      – A general CV and a specialized CV have different levels of term specificity.
        » E.g., MeSH vs. SNOMED Clinical Terms (CT) vs. NCI Thesaurus (an example next slide)

• Relationships with related CVs (advanced)
  – Various relationships among terms across multiple CVs should be identified
  – These relationships are retained and maintained for future use. What use?
    » Interoperability initiatives
    » Sometime, one CV is not enough!

# How to check it out
2. Select Metathesaurus Browser under the UTS Applications top-down menu
3. Type thyroid neoplasm
4. Select the first concept in the search results box
5. Find the Contexts branch under Report View
6. Find MSH, NCI, and SNOMEDCT branches

CV: Terms

• Grammatical forms of terms
  – The grammatical form of a term should be a noun or noun phrase (there are several different types)
    • Verbal nouns:
      – verbs should not be used alone as terms
        » E.g., Bleeding (not bleed), distillation (not distill)
    • Premodified (adjective) noun phrases:
      – are the preferred form
        » E.g., medical informatics
    • Postmodified noun phrases:
      – are also allowed but should be restricted to concepts that cannot be expressed in any other way
        » E.g., hospital for children (X), children’s hospital (O)
        » E.g., vaginal birth after Cesarean (O)

• Forms of count nouns
  – Count nouns should normally be expressed as plurals.
    • E.g., Neoplasms; lung neoplasms; viruses; bacteria; cells; genes
    – But the names of body parts are generally formulated in the singular
      • E.g., heart; brain; ear; eye; lung; breast; hand; leg
        » Blood vessels; arteries; microvessels; veins; heart valves; heart ventricles;
    • How about finger? “finger” or “fingers”
      – Fingers!
    • How about hair?
CV: Terms

- Selecting the preferred form
  - Neutral terms should be selected
    - E.g., Developing nations rather than underdeveloped countries
  - When two or more variants have literary warrant, the most frequently used term should be selected
    - "Myocardial Infarction" vs. "Myocardial infarct"
    - Which one is the more widely used in PubMed?
  - If a choice between spellings is made for dialectal reasons (e.g., American and British English), the choice should be adhered to consistently throughout the CV.
    - E.g., labor pain rather than labour pain

- Full names should be selected rather than abbreviations and acronyms.
  - If abbreviations and acronyms have become so well established, they should be selected as terms
    - E.g., HIV rather than Human immunodeficiency virus
    - E.g., DNA rather than Deoxyribonucleic Acid
    - E.g., laser (light amplification by stimulated emission of radiation)
    - How about AIDS (in MeSH)?
      - Not AIDS. Why?
      - But AIDS is widely used in MeSH e.g., AIDS-related complex, AIDS vaccines, etc.

CV: Terms

- Non-alphabetic characters
  - Hyphens
    - E.g., (Standard) nonfiction;
    - E.g., (MeSH) Carcinoma, Non-Small-Cell Lung, non-
  - Apostrophes
    - For medical eponyms, the use of the possessive form is becoming progressively less common.
      - E.g., Down Syndrome rather than Down’s Syndrome
      - E.g., Raynaud disease rather than Raynaud’s disease
      - E.g., Machado-Joseph disease rather than Machado-Joseph’s disease

CV: Relationships

- The relationships among terms in a CV are indicated by semantic linking
  - Three types of semantic relationships
    - Equivalency
      - Synonyms (e.g., HIV & AIDS virus)
      - Lexical variants (e.g., cancer & cancers)
      - Near synonyms (e.g., Benign Neoplasm & cancer)
    - Hierarchy
      - Generic, instance, and whole/part
        - E.g., arteries BT blood vessels
    - Associative
      - Many associative types (e.g., cause/effect)
      - UMLS provides associative R between concept categories (semantic types) not terms/concepts
CV: Relationships

- Hierarchical relationships
  - Polyhierarchical relationships
    - Some concepts belong, on logical grounds, to more than one category.
    - E.g., diabetes mellitus belongs to glucose metabolism disorders and endocrine system diseases
  - Node labels in hierarchies
    - Form a logical level in a hierarchy and serve to group a set of narrower terms
    - Node labels are not official CV terms, and must not be used as indexing terms.
      - E.g., Neoplasms by Site (MeSH)

What is (an) ontology?

- In the philosophy
  - Onto + logy
    - Onto: being or existence
    - logy: theory or science (e.g., biology and pathology)
    - The science of being or existence
      - “A branch of metaphysics concerned with the nature and relations of being”
      - Aristotle (Greek philosopher 384-322 B.C.)
        - Concerns universal categories for classifying everything that exists
        - Ontology is a uncountable noun

What is (an) ontology?

- In the IT field (AI, CS, IS, BMI, HI, etc)
  - “formal, explicit specification of a shared conceptualization” by Gruber
    - conceptualization: an abstract model of phenomena in the world that is created by identifying relevant concepts
    - shared: consensus on the conceptualization among experts
    - specification: consists of concepts, relationships among them, synonyms, concept hierarchy, constraints (rules), etc
    - explicit: the specification is explicitly defined.
    - formal: the ontology is unambiguous and there is a consensus on the specification so it can be machine readable/processable
  - knowledge (of a domain) itself! (countable noun)

- Ontology is a science and its result

What is (an) ontology?

- Benefits of use of ontologies
  - Communication between people
    - “Biologists would rather share their toothbrush than share a gene name”
  - Interoperability between intelligent systems
  - Make domain knowledge explicit
  - Reuse of domain knowledge
    - if it is represented in ontology
    - Introduction to biomedical ontologies (4/7mins)
  - National Center for Biomedical Ontology
Controlled vocabulary vs. Ontology

• How are they different?
  – No absolute answer!
  – Failed to reach a consensus!

• Ontologists vs. biomedical informaticians

Ontology = CV + description logic (DL)
• DL is for computational inferences (strong smell of AI!)
• UMLS, GO, etc are NOT ontologies

“An ontology is a kind of CV of well defined terms with specified relationships between those terms, capable of interpretation by both humans and computers.” (Broad view of ontology)

Controlled vocabulary vs. Ontology

• Methods in Medical Informatics
  – By Dr. Jules Berman
    • President of the Association for Pathology Informatics
  – “Because each MeSH term may be assigned multiple MeSH numbers, each with its own hierarchy, the MeSH data structure is more accurately thought of as a complex ontology…”

Medical Subject Heading (MeSH)

• What is MeSH?
  – developed by the National Library of Medicine (NLM, NIH) in 1960
  – a biomedical controlled vocabulary/ (informal) ontology
  – used for indexing MEDLINE articles as well as NLM documents.
    - [Q] Can we use it for retrieving MEDLINE?
      – updated annually
Medical Subject Heading (MeSH)

- **MeSH = {Vocabulary, MeSH Tree}**
  - **Vocabulary**
    - **Descriptors**
      - Topical Descriptors (DescriptorClass = "1")
        - Pharmacologic Actions[PA]
      - Publication types (DescriptorClass = "2")
      - Check Tag (DescriptorClass = "3")
      - Geographic Descriptors (DescriptorClass = "4")
        - Entry terms
    - Qualifiers (subheadings)
    - Supplementary Concept Records (SCRs)

- **Entry terms**
  - Qualifiers (subheadings)
  - Supplementary Concept Records (SCRs)

Medical Subject Heading (MeSH)

- **MeSH = {Vocabulary, MeSH Tree}**
  - **Vocabulary**
    - **Descriptors**
      - Topical Descriptors (DescriptorClass = "1")
        - Pharmacologic Actions[PA]
      - Publication types (DescriptorClass = "2")
      - Check Tag (DescriptorClass = "3")
      - Geographic Descriptors (DescriptorClass = "4")
        - Entry terms
    - Qualifiers (subheadings)
    - Supplementary Concept Records (SCRs)

- **Entry terms**
  - Qualifiers (subheadings)
  - Supplementary Concept Records (SCRs)

Medical Subject Heading (MeSH)

- **Descriptors**
  - main headings (or representatives of (near) synonyms)
    - E.g., neoplasms, tumors, cancers, benign neoplasms
    - E.g., hypotension, low blood pressure, vascular hypotension
  - Indicate the subject of an article
    - also include other types of terms for precise indexing.
      - E.g., publication types, geographic information, & check tag.
  - Revised every year (added, deleted, and renamed)
    - 2011 MeSH: 26,142 Descriptors
      - >177,000 entry terms
    - 2010 MeSH: 25,588 Descriptors
    - 2009 MeSH: 25,186 Descriptors
    - 2008 MeSH: 24,767 Descriptors
    - 2007 MeSH: 22,997 Descriptors

Medical Subject Heading (MeSH)

- **Descriptors (con’t)**
  - Have a 3-level structure
    (in fact, 4-level structure)
  - **Descriptor**
    - **Concepts**
    - **Terms**
      - (Strings or permuted terms)
    - => Entry terms
  - A group of a few similar people
    - People (individuals)
    - Names
    - All possible names
  - Each official object has a unique name and ID
**Lung Neoplasms (D008175) [Descriptor]**

- Lung Neoplasms (M0012749) [Concept, preferred]
- Lung Neoplasms (T024371) [Term, preferred, NP]
- Neoplasms, Lung (T024370) [Term, NP]
- Lung Neoplasm (T024370) [Term]
- Neoplasm, Lung (T024370) [Term]
- Neoplasms, Pulmonary (T024372) [Term, NP]
- Pulmonary Neoplasms (T024372) [Term]
- Lung Cancer (M0012750) [Concept]
- Lung Cancer (T024374) [Term, preferred, NP]
- Cancer, Lung (T024374) [Term]
- Cancers, Lung (T024374) [Term]
- Lung Cancers (T024374) [Term]
- Pulmonary Cancer (T364541) [Term, NP]
- Cancer, Pulmonary (T364541) [Term]
- Cancers, Pulmonary (T364541) [Term]
- Pulmonary Cancers (T364541) [Term]
- Cancer of the Lung (T364542) [Term, NP]
- Cancer of Lung (T364540) [Term, NP]
- Pulmonary Neoplasms (T024373) [Term, NP]

**MeSH**

- How to obtain MeSH?
  - Go to [www.nlm.nih.gov/mesh](http://www.nlm.nih.gov/mesh)

  - Read MeSH Memorandum of Understanding
  - Fill out the MeSH Registration Form
  - Save the “Files Available to Download or View” page
    - "Save As..." (IE) or "Save Page As..." (FireFox)
  - Email the web page to yourself for the future use.

**Medical Subject Heading (MeSH)**

- **MeSH = {Vocabulary, MeSH Tree}**
- **Vocabulary**
  - Descriptors
    - Topical Descriptors (DescriptorClass = "1")
      - Pharmacologic Actions[PA]
    - Publication types (DescriptorClass = "2")
    - Check Tag (DescriptorClass = "3")
    - Geographic Descriptor (DescriptorClass = "4")
      - Entry terms
  - Qualifiers (subheadings)
  - Supplementary Concept Records (SCRs)

- **Topical Descriptors**
  - Indicate the subject of an article
    - E.g., Lung Neoplasms, Myocardial Infarction, etc
  - Most Descriptors are “topical”

- **Publication Types [PT]**
  - Indicate genres of articles
    - E.g., meta-analysis, Randomized Controlled Trial, etc
  - Unlike other descriptors you must use the [pt] search tag on PubMed searches
    - [Q] What happened if you don't use the search tag (just type "meta-analysis")?
Medical Subject Heading (MeSH)

- **MeSH = {Vocabulary, MeSH Tree}**
  - **Vocabulary**
    - **Descriptors**
      - Topical Descriptors (DescriptorClass = "1")
        - Pharmacologic Actions[PA]
      - Publication types (DescriptorClass = "2")
      - **Check Tag** (DescriptorClass = "3")
      - Geographic Descriptor (DescriptorClass = "4")
        - **Entry terms**
      - Qualifiers (subheadings)
      - Supplementary Concept Records (SCRs)

Medical Subject Heading (MeSH)

- **Check Tag**
  - "Female" and "Male" (only two Descriptors!)
    - Gender of subjects (human & animal) of an experiment
  - Not placed in the MeSH Tree (no tree number)
  - [Q] How to check it out in the XML MeSH file?
    - by searching **DescriptorClass = "3"** in the XML MeSH file
  - Used to contain age-related terms in the past.
    - Now, they are topical Descriptors.
      - Listed in Category M of the MeSH Tree

Medical Subject Heading (MeSH)

- **Geographic Descriptors**
  - characterize physical locations of the study
    - E.g., Americas, North America, United States, Illinois, Chicago, etc
  - Listed in Category Z of MeSH Tree
  - An example (PMID: 20042561)
    - [Q] In what city was the study performed?
      - Chicago (located under the MeSH Terms)
Medical Subject Heading (MeSH)

- **MeSH Tree**
  - the hierarchy of MeSH Descriptors.
    - MeSH Descriptors are organized in 16 categories.
    - Each category is further divided into subcategories.
      - E.g., the Anatomy category has Digestive System, Respiratory System, Cardiovascular System, Nervous System, etc.
    - Descriptors are hierarchically arranged from most general to most specific in up to 11 hierarchical levels.
  - MeSH Descriptors are normally located in more than one place in MeSH Tree.
    - MeSH terms are usually classified into more than one category and have at least one address in the MeSH Tree.

Medical Subject Heading (MeSH)

- PubMed’s exploding (inclusive) search
  - When a MeSH descriptor is searched, PubMed automatically searches both the descriptor and its more specific (child) headings underneath in the MeSH Tree.
  - [Q] If you search the term lung neoplasms [major] in PubMed, what terms are you also searching?
  - [Q] the term “Neoplasms by site” is a node label and is NOT used for indexing. Explain why “Neoplasms by site” [mh] retrieves a lot of citations.
  - You can turn off the automatic explosion feature by using [mh:noexp] or [major:noexp].

Medical Subject Heading (MeSH)

- **MeSH = {Vocabulary, MeSH Tree}**
- **Vocabulary**
  - Descriptors
    - Topical Descriptors (DescriptorClass = "1")
      - Pharmacologic Actions[PA]
    - Publication types (DescriptorClass = "2")
    - Check Tags (DescriptorClass = "3")
    - Geographic Descriptors (DescriptorClass = "4")
      - Entry terms
  - Qualifiers (subheadings)
  - Supplementary Concept Records (SCRs)

Medical Subject Heading (MeSH)

- **Qualifiers (Subheadings)**
  - further describe a particular aspect of MeSH descriptors.
  - E.g., if you are interested in adverse effects of aspirin, type “aspirin/adverse effects”[MeSH]
    - “aspirin” is a descriptor
    - “adverse effects” is a subheading (qualifier)
  - What does “liver/drug effects” [mh] indicate?
    - Articles about the drug effects on the liver.
  - Only one subheading may be attached to a descriptor at a time.
Medical Subject Heading (MeSH)

- 83 qualifiers or subheadings (basic or adv. view)
  - aspirin/adverse effects[mh] = aspirin/ae[mh]
- Not every qualifier is suitable for use with every a descriptor
  - Quiz: Is either aspirin/drug effects or liver/adverse effects correct?
- Hierarchically structured
  - Like MeSH Descriptors, subheadings usually belong to more than one category.
  - You can “free float” qualifiers using the tag [sh]
    - E.g., aspirin[mh] AND ae[sh]
    - Quiz: Is that the same as aspirin/ae[mh]?

Use of MeSH in Indexing

- MeSH is designed to index the MEDLINE DB.
- 3 principles of MEDLINE Indexing using MeSH
  - Multiplicity
    - Each article generally discusses multiple subjects so an indexer supplies multiple Descriptors.
      - Around 5-15 Descriptors are usually assigned to each article.
    - [Q] Which article is assigned more MeSH Descriptors?
      - A research article vs. a survey (literature review) article
  - Co-ordination
    - If possible, multiple MeSH descriptors and/or qualifiers are combined to index a complex subject
      - Not creating a new Descriptor for every new subject/concept.
      - When a particular complex subject occurs frequently in MEDLINE for long time, a new descriptor may be created.
    - E.g., Aspirin-induced Asthma concept or subject (2010 MeSH term)
      - Aspirin and Asthma Chemically induced had been used for +35 years
      - This is why you should know qualifiers (subheadings).
Use of MeSH in Indexing

- Co-ordination (con’t)
  - If there is a right complex descriptor, that is used for indexing (not combining two or more simple concepts)
    - E.g., for a subject of arm injuries,
      (X) Combining the Descriptor Arm with the qualifier Injuries
      (O) Simply using the single Descriptor Arm Injuries
  - [Q] How is the multiplicity different from the co-ordination?
    - Multiplicity indicates multiple terms are used to represent an article discussing several topics/subjects
    - Co-ordination indicates multiple terms are used to represent a complex concept/subject

Use of MeSH in Indexing

- Specificity
  - Indexers are required to use the most specific MeSH Descriptor available (not the broader subjects).
    - E.g., an article about pulmonary pathology is indexed under the Descriptor Lung rather than the more general Descriptor Respiratory System.
  - [Q]: Why does NLM use most specific MeSH Descriptors?
    - A specific Descriptor provides more information!
    - It avoids multiple, redundant, indexing because of PubMed’s exploding (inclusive) search.
  - [Q] Explains why you should be familiar with use of the MeSH Tree (or the MeSH term hierarchy)

UMLS

- History
  - Started at NLM in 1986
- UMLS is updated 2-4 times per year
- Purpose
  - To facilitate the development of intelligent systems that behave as if they “understand” the language of biomedicine and healthcare
    - An example in next slide

UMLS

- What domain knowledge do you need to know that the two sentences are semantically identical?
  - Melatonin is a sleeping hormone.
  - A supplement is nonprescription medicine.
  - Insomnia is a disorder characterized by difficulty falling asleep.
- UMLS supplies this information!

Melatonin is a safe, effective medicine, not requiring a doctor’s prescription, for insomnia
The sleeping hormone supplement is recommended for people with difficulty falling asleep
**UMLS**

- UMLS has three knowledge sources
  - **Meta**thesaurus (MTH)
    - Meta- means “more comprehensive”
  - Semantic Network
    - Semantic Types (concept categories)
    - Semantic Relations between semantic types
  - SPECIALIST Lexicon and lexical tools
    - Lexicons and natural language processing (NLP) tools for biomedical text

❖ UMLS is more than a collection of biomedical vocabularies/ontologies (why?)
  - Semantic network and the NLP tools and its own terms

**How to use the UMLS**

- Prerequisite: you must obtain a license from NLM
  - UMLS Terminology Services (**UTS**) (Formerly, UMLS Knowledge Source (UMLSKS) Server)
    - You must create an account
  - UTS Web Services (ongoing project)
    - Formerly, UMLSKS Java API (no longer available)
      - Programmatically access UMLSKS via the Internet
      - an old-fashion API (a language-dependent approach)
    - Local installation
      - You can download UMLS files from NLM or use UMLS DVD to install using **MetamorphoSys**.

**Not** an end-user application but user-unfriendly resources (data) you must customize to use!

- Why customize UMLS?
  - You don’t need all of them for a specific problem or application
    - Do you need Spanish terms or vocabulary “housekeeping” attributes?
  - The default “preferred name” for a concept might not be best for your applications.
    - For clinical applications SNOMED CT terms are preferred.
    - For PubMed applications MeSH terms are preferred.
  - You don’t have the license required for operational use of all source vocabularies
    - UMLS is free to download. But that doesn’t necessarily mean you can use it for any purposes.

**UMLS - Meta**Thesaurus

- means a more comprehensive thesaurus
- integrates all major biomedical vocabularies
  - 2011AB version uses 161 source vocabularies
- contains
  - biomedical concepts, their various names, and
  - the relationships among them.
- organized by concept (or meaning)
  - 2011AB contains more than 2.6 M concepts and 8.6 M unique names
  - 2010AB: > 2.3 M concepts and 8.5 M unique names
**UMLS - MetaThesaurus**

- **Principle of integration**
  - The Metathesaurus preserves everything in ontologies rather than “corrects” them even though there are a lot of inconsistence (hierarchies)
  - E.g., when two different sources use the same name for different concepts, UMLS represents both of the meanings and indicates which meaning is present in which source (two AUIs).
  - NLM thinks that inconsistent data are not wrong but each ontology has just a different view.
  - NLM adds extra relationships and terms.
  - Kind of some “glue” among ontologies
  - UMLS is more than a collection of biomedical CVs

**UMLS - MetaThesaurus**

- **has 4-tier structure: CUI – LUI – SUI – AUI**
  - **AUI**: atom UI (source)
    - The basic building blocks or “atoms” are the concept names from each of the source vocabularies.
    - Every occurrence of a string in each source vocabulary even for the same concept is assigned an AUI
      - A0027665: Atrial Fibrillation (from MSH)
      - A0027667: Atrial Fibrillation (from PSY)
      - A0027668: Atrial Fibrillations (from MSH)
      - A0027930: Auricular Fibrillation (from PSY)
      - A0027932: Auricular Fibrillations (from MSH)
    - A single AUI is always linked to a single concept (CUI)
      - Meaning that AUI has only one meaning!

**UMLS - MetaThesaurus**

- **has 4-tier structure: CUI – LUI – SUI – AUI**
  - **SUI**: string UI (spelling & language)
    - Any variation in character set, upper-lower case, or punctuation is a separate string with a separate SUI
      - A0027665: Atrial Fibrillation (from MSH) >>> S0016668
      - A0027667: Atrial Fibrillation (from PSY) >>> S0016668
      - A0027668: Atrial Fibrillations (from MSH) >>> S0016669
    - If there are same strings (same meaning/concept) in different languages (e.g., English and Spanish), they will have a different SUI for each language.
      - A1756664: acute coryza (BAQ) >>> S1805178
      - A1505070: acute coryza (DAN) >>> S1555988
    - Each unique concept name has a SUI

**UMLS - MetaThesaurus**

- **has 4-tier structure: CUI – LUI – SUI – AUI**
  - **SUI**: string UI (con’t)
    - If the same string has more than one meaning (e.g., Cold - S0026353), the string identifier will be linked to more than one CUI
      - A15576916 : Cold (MSH) >>> S0026353 >>> C0009264 (cold temperature)
      - A0040708 : Cold (COSTAR) >>> S0026353 >>> C0009443 (common cold)
  - [Q] SUI and CUI have (one-to-one, one-to-many, or many-to-many) relationship.
    - Many to many
UMLS - MetaThesaurus

- has 4-tier structure: CUI – LUI – SUI – AUI
  - **LUI**: term UI (lexical variation)
    - Each concept name is linked to all of its **lexical variants** or minor variations by means of a LUI
  - **CUI**: concept UI (meaning)
    - Each concept or meaning has a unique concept identifier (CUI).
    - A concept or meaning can have many different names.

<table>
<thead>
<tr>
<th>Atoms (AUI)</th>
<th>Strings (SUI)</th>
<th>Terms (LUI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0027665: Atrial Fibrillation</td>
<td>S0016668: Atrial Fibrillation</td>
<td>L0004238: Atrial Fibrillation (preferred)</td>
</tr>
<tr>
<td>A0027668: Atrial Fibrillations</td>
<td>S0016669: Atrial Fibrillations</td>
<td></td>
</tr>
<tr>
<td>A0027930: Auricular Fibrillation</td>
<td>S0016899: Auricular Fibrillation</td>
<td></td>
</tr>
<tr>
<td>A0027932: Auricular Fibrillations</td>
<td>S0016900: Auricular Fibrillations</td>
<td></td>
</tr>
</tbody>
</table>

- To detect English lexical variants the Lexical Variant Generator (lvg) S/W (one of the UMLS Lexical Tools) is used.

MRCONSO.RRF shows the 4-tier structure.

UMLS - MetaThesaurus

- Includes **relationships** between concepts
  - 2 types of relationships: **Intra** and **inter**
    - **Intra**: intra-source vocabulary relationships
      - Most of them come from individual sources.
    - **Inter**: inter-source vocabulary relationships
      - Most of them are added by NLM.
    - Relationship labels are simple e.g., Broader, Narrower, translation, Qualifier of, etc
      - MRREL.RRF contains these relationships.
        - Abbreviations for relationship labels (use MRDOC.RRF)
    - Every relationship has a RUI.

Understanding of Metathesaurus data or files

- Two problems: **No column heads** and **a lot of abbreviations** in data (open **MRFILES.RRF**)
  - **MRFILES.RRF**
    - Summary of MTH files (Table of Contents)
    - contains basic information about files (column heads, short descriptions, # of records/rows, file size, etc)
  - **MRCOLS.RRF**
    - provides short descriptions of data attributes (column heads) used in MRFILES.RRF
      - All attribute names (column heads) are abbreviations
    - Chapter 3 of the **UMLS Reference Manual** contain a little detailed information for some MTH files.
UMLS - MetaThesaurus

- Understanding of Metathesaurus data or files
  - MRDOC.RRF
    - Challenging: attribute values are also abbreviations
    - Provides allowed values (abbrev.) of data attributes, full names, and short descriptions
    - Format: Attribute name | allowed values | type | full names/short descriptions
  - MRSAB.RRF contains source vocabulary information

- MRCONSO.RRF (the most fundamental resource)
  - There is exactly one row (record) for each AUI
  - [Q] If there are 4.5M records in the file, how many AUIs exists?
    - 4.5M AUIs
  - [Q] Did you install UMLS FULLY or PARTIALLY?
    - You did not fully install UMLS but partially.
      - You selected the default installation including frequently used vocabularies.
  - Every string or concept name appears in this file
    - Connected to its language, source, SUI, LUI, and CUI
UMLS - MetaThesaurus

- **MRCONSO.RRF** (the most fundamental resource)
  - Has the following attributes:
    - CUI, LAT, TS, LUI, STT, SUI, ISPREF, AUI, SAUI, SCUI, SDUI, SAB, TTY, CODE, STR, SRL, SUPPRESS, CVF
  - [Q] Among them what attribute can be used as a primary key?
    - AUI
  - The key is how to read data?

- **C0009443 | ENG | P | L0009443 | VO | S0026750 | Y | A0041267 | M0004864 | D003139 | MSH | PM | D003 139 | Common Colds | 0 | N | 1536 |

The UMLS Reference Manual contain a little detailed information for each file of 14 Metathesaurus files

<table>
<thead>
<tr>
<th>CUI</th>
<th>TS</th>
<th>LUI</th>
<th>STT</th>
<th>SUI</th>
<th>ISPREF</th>
<th>AUI</th>
<th>CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>VW</td>
<td>50026363</td>
<td>N</td>
<td>A15661752</td>
<td>Cold, Common</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>VW</td>
<td>50026363</td>
<td>N</td>
<td>A18018470</td>
<td>Cold, Common</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>VW</td>
<td>50026363</td>
<td>Y</td>
<td>A0040725</td>
<td>Cold, Common</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>VO</td>
<td>50026365</td>
<td>Y</td>
<td>A0040727</td>
<td>Colds, Common</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>PF</td>
<td>N</td>
<td>A0041261</td>
<td>Common Cold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>PF</td>
<td>N</td>
<td>A0041263</td>
<td>Common Cold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>PF</td>
<td>N</td>
<td>A15662792</td>
<td>Common Cold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>PF</td>
<td>N</td>
<td>A17994729</td>
<td>Common Cold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>PF</td>
<td>N</td>
<td>A7569706</td>
<td>Common Cold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>PF</td>
<td>Y</td>
<td>A0041262</td>
<td>Common Cold</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The physical file name is MRCONSO.RRF and comma separated list of column names (COL), in order:

- CUI, LAT, TS, LUI, STT, SUI, ISPREF, AUI, SAUI, SCUI, SDUI, SAB, TTY, CODE, STR, SRL, SUPPRESS, CVF

- **MRFILES.RRF** - MRCONSO.RRF
- **MRCOLS.RRF (simple)** - MRCONSO.RRF

<table>
<thead>
<tr>
<th>Column or data element name</th>
<th>descriptive name</th>
<th>physical file name in which this field occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUI</td>
<td>Unique identifier for concept</td>
<td>MRCONSO.RRF</td>
</tr>
<tr>
<td>LAT</td>
<td>Language of Term(s)</td>
<td>MRCONSO.RRF</td>
</tr>
<tr>
<td>TS</td>
<td>Term status</td>
<td>MRCONSO.RRF</td>
</tr>
<tr>
<td>LUI</td>
<td>Unique identifier for term</td>
<td>MRCONSO.RRF</td>
</tr>
<tr>
<td>STT</td>
<td>String type</td>
<td>MRCONSO.RRF</td>
</tr>
<tr>
<td>SUI</td>
<td>Unique identifier for string</td>
<td>MRCONSO.RRF</td>
</tr>
<tr>
<td>ISPREF</td>
<td>Indicates whether AUI is preferred</td>
<td>MRCONSO.RRF</td>
</tr>
<tr>
<td>AUI</td>
<td>Unique identifier for atom</td>
<td>MRCONSO.RRF</td>
</tr>
</tbody>
</table>

Source abbreviations (Lab): Open MRSAB.RRF to identify MSH
- Check if the MeSH term has SDUI and SCUI. [Q] How to check?
- TTY - term type in source vocabulary. What's PM? Refer to xxxxx.RRF?
- CODE - most useful source asserted identifier or Metathesaurus-generated source entry identifier
- String
- SRL - source restriction level (0-4)
- SUPPRESS - suppressible flag (not used)
- CVF - content view flag (internally used)

- SCUI - source asserted concept identifier (optional)
- SDUI - source asserted descriptor identifier (optional)
- SAB - Source abbreviation.
- Lab: Open MRSAB.RRF to identify MSH
- Lab: Check if the MeSH term has SDUI and SCUI. [Q] How to check?
- SCUI - source asserted concept identifier (optional)
UMLS - MetaThesaurus

- **MRREL.RRF**
  - There is one row for each relationship between concepts or atoms (RUI is the ID of the table)
  - Intra and Inter (or Internal & External) relationships

- **Lab**
  - Go to [http://uts.nlm.nih.gov](http://uts.nlm.nih.gov) and log in
  - Select Metathesaurus Browser under Application menu
  - Select CUI
  - Search for the concept names in the table (in the handout) by CUI

- **MRREL.RRF (con't)**
  - TID 1-2 (See the handout)
    - **Lab:** Find out the names of A0040727 and A0041267
      - How?
        » You can either search in MRCONSO.RRF or use UTS
        » Search the AUIs in the search result of C0009443 (under Report View)
      - A0040727: Colds, Common
      - A0041267: Common Colds
      - They are permuted terms
      - Officially, they are synonyms (**SY**)
  - Relationships within the concept
    - Relationships between AUIs

- **MRREL.RRF (con't)**
  - TID 3-5
    - As the SAB indicates, the three AUIs are translation of A0006735
      - A11023458: hoofdverkoudheid
      - A11094248: Rhume de cerveau
      - A11136019: Schnupfen
      - You cannot find those AUIs in your UMLS files but UTS. **Why?**
        » You selected the default setting that does not install many foreign vocabularies

- **MRREL.RRF (con't)**
  - TID 6-7 (Explain the relationships)
    - Allowed qualifiers for the concept (Common Cold)
    - **Lab:** Check if they are allowed qualifiers by searching with “common cold” in the MeSH DB
      - nutritional management is a UMLS term.
      - The MeSH term of the concept is **diet therapy**
The relationships are between SCIUs in MeSH.

- **Lab**: What is the descriptor-concept relationship (in MeSH) of the UMLS concept?
  - How to identify the descriptor-concept relationship in MeSH?
    - Search the XML MeSH file (i.e., desc2012.xml)
    - How to search? What is a proper search keyword?
      - "common cold" is NOT a good search keyword
    - You need the descriptor ID of common cold for the search
    - How to get the descriptor ID?
      - MeSH Browser not the MeSH database; Google it
  - Finally, how to compose the search query with the descriptor ID?
    - `<DescriptorUI>D003139</DescriptorUI>`

Acute coryza and catarrh are child concepts of C.C.

**MRREL.RRF (con't)**

- **TID 10**
  - **RB** indicates "has a broader relationship"
    - Virus Diseases is a broader/parent concept of the concept C.C.

- **TID 11-12**
  - **SIB** indicates "has sibling relationship in a Metathesaurus source vocabulary"
    - Hepatitis A and Herpesviridae Infections are sibling concepts of C.C.

**UMLS – Semantic Network**

- Provides a categorization of all the UMLS concepts and a set of useful relationships between categories (not concepts)
  - **133 Semantic Types (categories)** (ST)
    - Every UMLS concept belongs to at least one ST
    - Three STs were deleted and one ST was added
  - **54 Semantic Relations (relationship labels)** (SR) between semantic types
    - No semantic relations between concepts but STs
    - STs and SRs are hierarchically arranged.
    - STs are the nodes and SRs are the links in the Semantic Network.
UMLS – Semantic Network

Coherent UMLS Example

UMLS Metathesaurus (Concepts)
- Raynaud Disease
- Blood Viscosity
- Fish Oils

UMLS Semantic Types
- Disease or Syndrome
- Physiologic Function
- Biologically Active Substance

UMLS Semantic Relations
- Disease or Syndrome
- Physiologic Function
- Biologically Active Substance

UMLS – Semantic Network

- Synonyms > (Concepts) > [Semantic Types]
  - [Q] If a relationship exists between two semantic types, do their concepts have the relationship?
  - The relations do NOT necessarily apply to all instances of (concepts) that have been assigned to the [semantic types].
  - [Sign] <evaluation_of> [Organism Attribute]
    - (overweight) <evaluation_of> (body weight)
    - (fever) <evaluation_of> (body temperature)
    - How about the following relationship?
      - (fever) <evaluation_of> (body weight)

UMLS – Semantic Network

A Portion of the UMLS Semantic Network

UMLS – Semantic Network

SRFIL: Table of Contents of Semantic Network files

<table>
<thead>
<tr>
<th>File name</th>
<th>Description of the file</th>
<th>Format of the file (fields in a comma-separated list)</th>
<th># of rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRDEF</td>
<td>Basic information about the Semantic Types and Relations</td>
<td>RT,U1,STY/R1,STN/R1N,RTN,DFF,FX,NT,N,NH,ABR,RIN</td>
<td>187</td>
</tr>
<tr>
<td>SRFIL</td>
<td>File Description</td>
<td>FIL,DES,FMT,CLS,RWS,BTS</td>
<td>6</td>
</tr>
<tr>
<td>SRFLD</td>
<td>Field Description</td>
<td>COL,DES,REF,FIL</td>
<td>21</td>
</tr>
<tr>
<td>SRSTRE1</td>
<td>Fully inherited set of Relations (UIs)</td>
<td>U1,U1,U1</td>
<td>6704</td>
</tr>
<tr>
<td>SRSTRE2</td>
<td>Fully inherited set of Relations (Names)</td>
<td>STY,RL,STY</td>
<td>6704</td>
</tr>
<tr>
<td>SRSTR</td>
<td>Structure of the Network</td>
<td>STY/RL,RL,STY/RL,LS</td>
<td>603</td>
</tr>
</tbody>
</table>

Why 187?
STY|T058|Health Care Activity|B1.3.1|An activity of or relating to the practice of medicine or involving the care of patients. |Ambulatory Care; Clinic Activities; Geriatric Nursing; Preventive Health Services||hlca||

STY: semantic type, RL: semantic relation (relation label)
T058: ID of STY or RL
Health Care Activity: the name of the semantic type
B1.3.1 or R1.1: Tree number of the ST or SR
An activity of or ...: the definition of the ST (Health Care Activity)
Ambulatory Care; Clinic Activities: the concept examples in the ST

HLCA: abbreviation of the ST or SR
performed_by and prevented_by: inverse of the relation (SR/RL only not for ST)

The relations are generally inherited via the “isa” link by all the children of STs.

- [Biologic Function]<process of>[Organism]
  - Link Status: D (Defined for the argument and its children)
- [Organ or Tissue Function]<process of>[Animal]

Do you remember semantic types and relations are hierarchically arranged like concepts (see ST and SR)
UMLS – Semantic Network

- Link Status: B (blocked)
  - In some cases the inheritance of the relation link is said to be blocked.
  - [Biologic Function]<process_of>[Organism]
  - [Mental Process]<process_of>[Virus]
  - [Mental Process]<process_of>[Plant]

UMLS – Semantic Network

- Defined but Not Inherited (DNI) relations
  - A relation is defined for two semantic types but blocked for all the children of those semantic types.
  - [Body System] <conceptual part of> [Fully Formed Anatomical Structure]

UMLS – SPECIALIST Lexicon

- Consists of a set of lexical entries with one entry for each spelling or set of spelling variants.
  - E.g., {treat, treats, treated, treating}
- Several lexical (NLP) tools are provided (implemented in Java)
  - To address the high degree of variability in natural language words and terms.
- word order variants for multi-word terms
- This approach is an ideal (but extremely expensive) solution for detecting concepts from MEDLINE articles.

Quiz: What are the relationships among the three files?
Questions or Comments?