Medical Data, Standard Vocabularies, Communication Standards

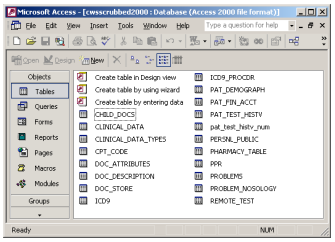
6.872/HST950

Peter Szolovits
(with some material from Chris Cimino)

Recall Children's Clinicians' Workstation Database

- Demographics
- Problems
- Allergies
- Medications
  - Immunizations
- Lab Data
- Clinical Measurements
  - Growth Charts
- Visit History
- Reports and Letters

The Database



- General Information
  - CPT_CODE
  - ICD9
  - ICD9_PROCDR
  - PROBLEM_NOSOLOGY
- Documents
  - DOC_STORE
  - DOC_ATTRIBUTES
  - DOC_DESCRIPTION
  - CHILD_DOCS
- Doctors
  - PERSNL_PUBLIC
  - PPR
- Patients
  - PAT_DEMOGRAPH
  - PAT_FIN_ACCT
  - PAT_TEST_HISTV
  - REMOTE_TEST
  - PHARMACY_TABLE
  - PROBLEMS

Vocabularies and Terminology

- Why?
  - Surrogate for "messy reality"
  - Uses
- How?
  - Flat list
  - Taxonomy (Hierarchy, Nosology, …)
  - Heterarchy
  - Combinatorial Language
    - Derivation rules
    - Inference
    - … knowledge representation

"Ontology" for Computer Folks

- An organization of concepts (hierarchy or heterarchy)
- (Some) concepts are defined in terms of others
  - A *triangle* is a *polygon* with exactly 3 *sides*
  - A *dachshund* is a *dog* (with ???)
- Automatic classification
  - If P is a 3-sided polygon with …, it is recognized automatically as a triangle

Definitions

- Word – a set of characters including punctuation delimited by white space.
- Term – one or more words used as a unit.
- Concept – an idea, action, or thing.
- Synonym – two terms for the same concept.

Vocabulary Uses

- Indexing – Finding what you want
- Cataloging – Putting away what you have
  - E.g., WHO, DRGs
- Knowledge Representation
  - Representing the facts
  - Blurring the facts
  - Creating new shades of meaning

Describe a term for a Laboratory Test.

- Where was it done?
- How was it done?
- Under what conditions was it done?
- How many minutes after eating carbohydrate was it measured?

Describe a Vocabulary for a Gene

- Whose gene?
- Gene fragment?
- Open Reading Frame?
- Promoter + all exons and introns
- Promoter + all exons + all introns + other binding sites affecting function?
- Final/draft/species/SNP/Alternative splicing?

Knowledge vs. Language

- Get two or more people to enumerate terms to describe the same set.
  - Do any terms match exactly?
  - Do terms differ by word order?
  - Do terms differ by word suffix or prefix?
  - Are there terms that some people think are synonyms that other people think are not?

History of 3 Vocabularies

- MeSH — Index
- ICD — Precoodinated
- SNOMED — Post-coordinated

History

- The modern history of medical controlled vocabularies begins with the U.S. Army General Surgeon who petitioned Congress to fund a medical library. (~Civil War)
- The position eventually became “The US Surgeon General” and the library the National Library of Medicine
  - http://www.nlm.nih.gov/

History

- Library collection was indexed with Index Medicus (created by NLM) which is published in book form.
- Index Medicus was extended to index medical literature articles.
- Index Medicus was extended further to provide on-line indexing (1960). This became the Medical Subject Headings (MeSH).

MeSH

- Purpose is to index the medical literature.
- Content of MeSH is driven by publications.
- Who "owns" MeSH?
- What impact do vocabulary changes have?

MeSH – Structure

http://www.nlm.nih.gov/mesh/

- MeSH is organized into a series of "trees". (e.g. physical findings, diseases, chemicals)
- A MeSH main heading is a "concept". (e.g. "Neurologic Disease", "Epilepsy")
- Main Heading (MH) is often called a term. (Try to avoid doing this.)

MeSH – Structure

- Each MH has a unique identifier.
- Each MH may have multiple synonyms.
- Each MH may have multiple locations in multiple trees. Each of these "contexts" has a unique tree address. The concept of "context" is synonymous with "multiple inheritance".

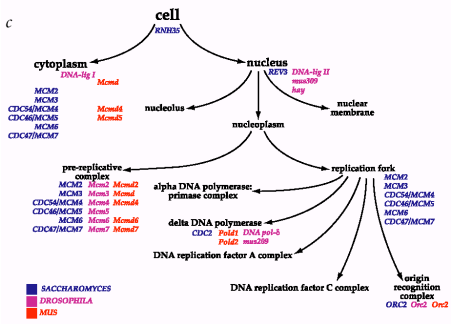
MeSH – Structure

- There is a small set of subheadings (50) that "modify" MH based on tree address. (e.g. "diagnosis" applies to MH in the "Disease" tree but not to the "Chemical" tree).
- There is a small set of tag terms (15) which exist unrelated to the rest of MeSH. (e.g., "Review Article", "Human", "Animal")

MeSH – Structure

- Every article is indexed with tag terms.
- Every article is indexed with MH terms for focus (main index term) and mention (minor index term).
- Every index term is checked for subheadings.
- This is all done by trained reviewers.
- The MeSH Vocabulary is revised annually.

MESH Redux—The Genome "Ontology"



International Classification of Disease (ICD)

- Any agency that dispenses funds for health care needs a way to assess needs and effectiveness.
- The United Nations World Health Organization (WHO) funds health care prevention projects world wide and gathers statistics for member nations.
- Who "owns" ICD?
- What impact will changes have?

ICD – Structure

- ICD is divided into categories based on a 5-digit numeric code. (e.g., "133.21")
- Usually round numbers are more general concepts (e.g., "100" subsumes "130" which subsumes "133")
- The fourth and fifth digit is called a modifier but it isn't really.

ICD – Structure

- The code is both the concept and the unique identifier. Multiple terms are linked to the same code.
- Every patient is coded with as many terms as possible.
- Terms should be the most specific one to describe a particular problem.

ICD – Structure

- Coding scheme limits the size of the vocabulary.
- Obsolete codes must be reused.
- Base ten results in limited flexibility and the need for "other", "NOS", and "NOC" terms.

ICD – Structure

- Lack of multiple contexts or multiple inheritance results in duplicate terms.
- Lack of overall organization results in ambiguous terms.

ICD – Structure

- ICD has been adopted by most insurance companies as a method for controlling billing and payment.
- Economic forces drive how the vocabulary is used which drives how ICD is modified which drives changes in reimbursement which drives how the vocabulary is used…
- Who "owns" ICD?
- The Vocabulary is revised sporadically.

SNOMED – Structure

- Developed by the American College of Pathologists to overcome the faults of ICD.
- Really describes 6 [now 12] different vocabularies, one for each "axis" of a concept (e.g., anatomy, environment, history).
- Every concept is built up from a term from each "axis" (e.g., "surgery of" "blue" "nevus" "of left" "forearm").

SNOMED – Structure

- There is some overlap of the axes so it is possible to form two different versions for the same concept (e.g., "blue nevus" "nevus colored blue").
- There are few rules for how to combine axes terms so it is possible to form valid nonsense terms (e.g., "nevus" "of left" "esophagus").
- Who "owns" SNOMED?

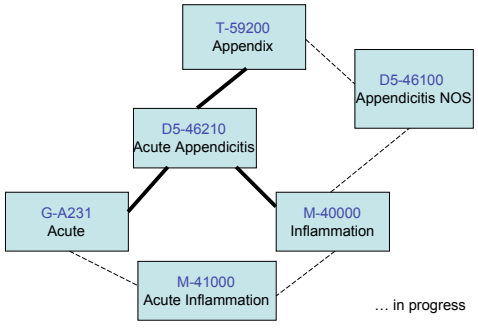
SNOMED Axes

- D – Diseases
- C – Drugs
- F – Function
- L – Living Organisms
- X – Manufacturers
- G – Modifiers
- M – Morphology
- J – Occupations
- A – Physical Agents
- P – Procedures
- S – Social Context
- T -- Topography

"Postcoordination"

- D5-46210 Acute Appendicitis
- G-A231 Acute, D5-46100 Appendicitis NOS
- G-A231 Acute, M-40000 Inflammation, T-59200 Appendix
- M-41000 Acute Inflammation, T-59200 Appendix
- T-59200 Appendix, M-41000 Acute Inflammation

Semantics of Postcoordination



SNOMED-CT (Clinical Terminology)

- Combined SNOMED + Reed Codes
  - SNOMED for diseases
  - Reed for symptoms
- Licensed by NLM for anyone to use royalty-free for 5 years starting 2004.
  - Attempt to encourage standardization

History

- For everyone who wants to "own" a medical vocabulary, there is a set of terms which are likely to overlap but be inconsistent with every other vocabulary.
- Read, CPT, COSTART, ChemAbstracts, …
- In theory they are all describing agreed upon concepts. A single standard vocabulary would improve the automated flow of medical information.

Ideal Vocabulary

?

Ideal Vocabulary

- Boundary
- Organization
- Completeness
- Absence of ambiguity
- Growth
- Aging

Ideal Vocabulary – Definitions

- String – a unique sequence of characters. The same set of characters may represent different concepts.
- Lexical variants – synonyms with "minor" differences. Word order, capitalization, and punctuation are usually included. Suffixes (plural) and prefixes may be included.
- One man's lexical variant is another's synonym.

Ideal Vocabulary – Definitions

- Related terms – distinct terms whose concepts overlap in some way. The most used relations are "broader" and "narrower" (e.g., "Neurologic Disease" includes but is broader than "Epilepsy".)
- One man's related term is another's synonym.

Ideal Vocabulary – Definitions

- Controlled versus "free" text
  - Freedom of expression
  - Automatic indexing accuracy
- Atomic versus enumerated (Pre vs Post)
  - Handle the unexpected
  - Predict what to expect
- Definitions
  - "Free" text versus semantic

Unified Medical Language

- The Unified Medical Language System (UMLS) started as an NLM collaborative program with 7 centers around the country.
- Proceeded in 3 – 3 year phases.
  - Explore ideas (1986)
  - One "winner" selected and developed (1988)
  - Usage Testing (1991)

UMLS – Structure

- Three components
  - Metathesaurus (META)
  - Semantic Network
  - Information Sources Map (ISM)
    - Recently dropped
  - Specialist Lexicon
    - Recently added

META – Structure

- NOT a controlled vocabulary.
- Database of information about other controlled vocabularies.
- Contains sufficient info to recreate most of the component vocabularies. (Who "owns" META? Who "owns" the components?)
- Basic unit is the concept. A concept is linked to multiple strings from multiple vocabularies.

META - Structure

- Each concept-string pair is either a preferred term, synonym, or lexical variant.
- The same string may be linked to multiple concepts but a term, synonym, or lexical variant will only link to one concept each.
- Other links exist based solely on the existence of those links in a source vocabulary.

META – Structure

- Each concept has only one preferred term chosen from all linked terms based on order of precedence of source vocabularies. With a few exceptions, MeSH is number one.
- Each concept is linked to semantic types in the semantic network.
- NOT a controlled vocabulary – or is it?

Semantic Network – Structure

- Small vocabulary that attempts to implement an ideal vocabulary
- Terms defined with free text definitions and by linkage.

UMLS Semantic Network

"T001" "Organism" "A1.1"
"Generally, a living individual, including all plants and animals."
"Homozygote; Radiation Chimera; Sporocyst"

"T002" "Plant" "A1.1.1"
"An organism having cellulose cell walls, growing by synthesis of inorganic substances, generally distinguished by the presence of chlorophyll, and lacking the power of locomotion. Plant parts are included here as well."
"Pollen; Potatoes; Vegetables"

"T003" "Alga" "A1.1.1.1"
"A chiefly aquatic plant that contains chlorophyll, but does not form embryos during development and lacks vascular tissue."
"Chlorella; Laminaria; Seaweed"

… 188 terms

UMLS Sem Net: Relations

H: isa
R: associated_with
R1: physically_related_to
R1.1: part_of
R1.2: consists_of
R1.3: contains
R1.4: connected_to
R1.5: interconnects
R1.6: branch_of
R1.7: tributary_of
R1.8: ingredient_of
R2: spatially_related_to
R2.1: location_of
R2.2: adjacent_to
R2.3: surrounds
R2.4: traverses

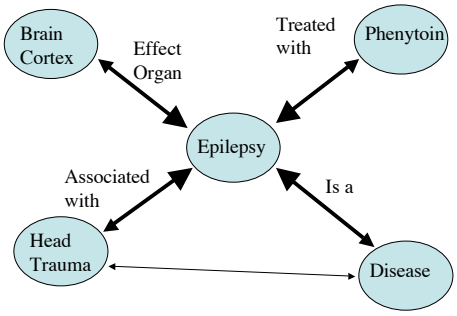
R3: functionally_related_to
R3.1: affects
R3.1.1: manages
R3.1.2: treats
R3.1.3: disrupts
R3.1.4: complicates
R3.1.5: interacts_with
R3.1.6: prevents
R3.2: brings_about
R3.2.1: produces
R3.2.2: causes
R3.3: performs
R3.3.1: carries_out
R3.3.2: exhibits
R3.3.3: practices
R3.4: occurs_in
R3.4.1: process_of

R3.5: uses
R3.6: manifestation_of
R3.7: indicates
R3.8: result_of
R4: temporally_related_to
R4.1: co-occurs_with
R4.2: precedes
R5: conceptually_related_to
R5.1: evaluation_of
R5.10: method_of
R5.11: conceptual_part_of
R5.12: issue_in
R5.2: degree_of
R5.3: analyzes
R5.3.1: assesses_effect_of
R5.4: measurement_of
R5.5: measures
R5.6: diagnoses
R5.7: property_of
R5.8: derivative_of
R5.9: developmental_form_of

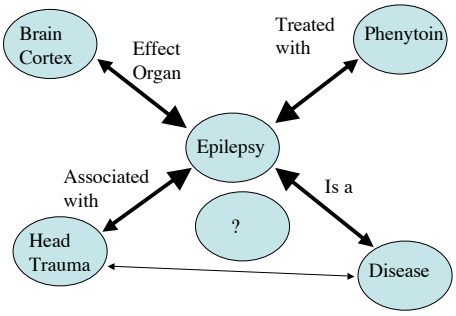
Definition of relations

("RL" "T132" "physically_related_to" "R1"
"Related by virtue of some physical attribute or characteristic." "" "" ""
"PR" "physically_related_to")
("RL" "T133" "part_of" "R1.1"
"Composes, with one or more other physical units, some larger whole. This includes component of, division of, portion of, fragment of, section of, and layer of."
"" "" "" "PT" "has_part")
("RL" "T134" "contains" "R1.3"
"Holds or is the receptacle for fluids or other substances. This includes is filled with, holds, and is occupied by."
"" "" "" "CT" "contained_in")
("RL" "T135" "location_of" "R2.1"
"The position, site, or region of an entity or the site of a process." "" ""
"" "LO" "has_location")
("RL" "T136" "temporally_related_to" "R4"
"Related in time by preceding, co-occuring with, or following." "" "" "" "TR"
"temporally_related_to")
("RL" "T137" "co-occurs_with" "R4.1"
"Occurs at the same time as, together with, or jointly. This includes is co-incident with, is concurrent with, is contemporaneous with, accompanies, coexists with, and is concomitant with."
"" "" "" "CW" "co-occurs_with")

Semantic Network



Semantic Network (Reification)



State of the Art

- UMLS is not sufficient.
  - META is not complete. Still weak for clinical terms (sign and symptom terms).
  - META has superficial organization. Links between vocabularies is based primarily on lexical matches. Inter-vocabulary links growing slower than total size.
  - Ambiguous sources mean META is ambiguous.

State of the Art

- Semantic typing does scale up so META and the semantic network can form a starting point.
- Semantic rules are being added to SNOMED which may remove its ambiguity problem. This would greatly strengthen SNOMED and META.
- Who "owns" the rules?

State of the Art

- Semantic tools are being developed to provide end user management of vocabularies. The same tools would allow users to add (or nominate) new terms and help the user understand the semantic definition of existing terms.
- Links back to META allow institutions to "own" a vocabulary while complying with other organizations' requirements.

Representation Languages

- Capabilities
  - Naming: intensional as well as extensional
    - E.g., "people who provide me healthcare"
  - Definitions
  - Assertions
- Examples
  - K-Rep
  - GALEN and GRAIL
  - (?) E31 of ANSI
  - DAML+OIL, OWL (Semantic Web, RDF, …)

ASTM E31 Subcommittees

- E31.01 Controlled Health Vocabularies for Healthcare Informatics
- E31.10 Pharmaco-informatics Standards
- E31.13 Clinical Laboratory Information Management
- E31.16 Interchange of Electrophysiological Waveforms & Signals
- E31.17 Privacy, Confidentiality, and Access
- E31.19 Electronic Health Record Content and Structure
- E31.20 Data and System Security for Health Information
- E31.22 Health Information Transcription and Documentation
- E31.23 Modeling for Health Informatics
- E31.24 Electronic Health Record (EHR) System Functionality
- E31.25 XML Document Type Definitions (DTDs) for Health Care
- E31.26 Personal (Consumer) Health Records
- E31.27 Data Capture and Reporting
- E31.90 Executive
- E31.95 Education and Publicity

General Rules for Gene Nomenclature

- **1.1. Requirements for designation by gene symbol**

In order for a gene symbol to be allocated at least one of the following criteria must apply.

1.1.1. A gene symbol may be used to designate a clearly defined phenotype shown to be inherited as a monogenic Mendelian trait. (Example:TSC1).

1.1.2. Gene symbols may be allocated to as yet unidentified genes contributing to a complex trait shown by linkage or association with a known marker (for example IDDM6).

1.1.3. A gene symbol may be used to designate a cloned segment of DNA with sufficient structural, functional, and expression data to identify it as a transcribed entity. However, alternate transcripts from the same gene should not in general be given different gene symbols.

1.1.4. Gene symbols are also allocated to non-functional copies of genes (pseudogenes).

1.1.5. Genes encoded by the opposite (anti-sense) strand of a known gene will be given their own symbols.

1.1.6. A gene symbol may be given to a transcribed but untranslated DNA segment eg XIST.

1.1.7. A cellular phenotype from which the existence of a gene or genes can be inferred may have its own designation. Example: LOH#CR#.

1.1.8. If insufficient data are available to allocate a unique and meaningful gene symbol, a putative gene may be designated by the symbol C#orf#. This symbol will also be used for EST clusters. Other fragments of expressed sequence will be designated by a D-number.

Communication Standards

- HL7, DICOM, CorbaMED, XML-based …
- HL7:
  - Messages (unit of transfer, type="purpose")
    - Segments (e.g., header, event-type, pat. id, …)
      - Fields (character string)
- HL7 Details
  - Optional and repeated fields
  - Character encoding:
    - <cr> Segment separator
    - | Field separator
    - ^ Component separator
    - & Subcomponent separator
    - ~ Repetition separator

HL7 Examples

- AD – Address

Components: <street address (ST)> ^ < other designation (ST)> ^ <city (ST)> ^ <state or province (ST)> ^ <zip or postal code (ST)> ^ <country (ID)> ^ <address type (ID)> ^ <other geographic designation (ST)>

|10 ASH LN^#3^LIMA^OH^48132_|

- CE – Coded Element

Components: <identifier (ST)> ^ <text (ST)> ^ <name of coding system (ST)> ^ <alternate identifier (ST)> ^ <alternate text (ST)> ^ <name of alternate coding system (ST)>

|F-11380^CREATININE^I9^2148-5^CREATININE^LN|

HL7 (3.0) Reference Implementation Model

- Attempt at complete specification of health-care related *communications* (and, by derivation, *records*)
- http://www.hl7.org/

What to do with "Free Text"?

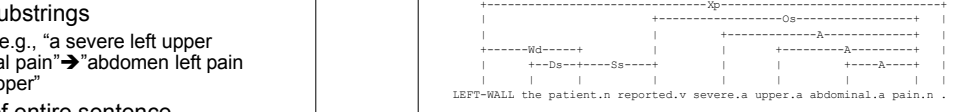
- "Free" is very expensive
  - But, it's what we have
- Approaches:
  - Eliminate it: pick lists, formal vocabularies, …
  - Deal with it: huge thesauri, … full language understanding

1. Treat text as a combination of terms

- "The patient reported a severe left upper quadrant abdominal pain."
  - Try to look up *all* substrings
    - Normalize strings: e.g., "a severe left upper quadrant abdominal pain"➔"abdomen left pain quadrant severe upper"
  - Find best "cover" of entire sentence
    - AI search problem, assuming some metric
  - Hope that combination of concepts is meaningful

2. Parse text according to rules of English grammar

- "The patient reported upper abdominal pain":




```
(S (NP The patient)
   (VP reported
       (NP severe upper abdominal pain))
   .)
```

- Grammar guides semantic composition of meaning

Semantic relations in NN terms

- Family doctor: subtype
- Bile secretion: activity
- Haptoglobin gene: production
- Asthma therapy: purpose
- Aids patient: person afflicted
- Blood analysis: study instrument
- Food allergy: cause
- … [Rosario & Hearst]