Mapping New Vocabularies to the UMLS

Experience with ICF

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

◆ Terminology integration
  • *The Unified Medical Language System*

◆ Methods
  • Normalizing terms
  • Categorizing terms
  • Recording relations
  • Editing and auditing

◆ Experience with ICF
Terminology integration

The Unified Medical Language System
Motivation

- Started in 1986
- National Library of Medicine
- “Long-term R&D project”
- Complementary to IAIMS

“[...] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
- The second is the distribution of useful information among many disparate databases and systems.”
Source Vocabularies

- 134 source vocabularies
  - 132 contributing concept names
- ~80 families of vocabularies
  - multiple translations (e.g., MeSH, ICPC, ICD-10)
  - variants (American-English equivalents, Australian extension/adaptation)
  - subsequent editions usually considered distinct families (ICD: 9-10; DSM: IIIR-IV)
- Broad coverage of biomedicine
- Common presentation
Biomedical terminologies

◆ General vocabularies
  ● anatomy (UWDA, Neuronames)
  ● drugs (RxNorm, First DataBank, Micromedex)
  ● medical devices (UMD, SPN)

◆ Several perspectives
  ● clinical terms (SNOMED CT)
  ● information sciences (MeSH, CRISP)
  ● administrative terminologies (ICD-9-CM, CPT-4)
  ● data exchange terminologies (HL7, LOINC)
Biomedical terminologies (cont’d)

◆ Specialized vocabularies
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART)
  - primary care (ICPC)
  - genomics (GO, OMIM, HUGO)

◆ Terminology of knowledge bases (AI/Rheum, DXplain, QMR)

The UMLS serves as a vehicle for the regulatory standards (HIPAA, CHI)
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Other subdomains
- SNOMED
- OMIM
- MeSH
- Model organisms
- NCBI Taxonomy
- UWDA
- GO
- Anatomy
- OMIM
- UMLS
- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Other subdomains
- SNOMED
- OMIM
- MeSH
- Model organisms
- NCBI Taxonomy
- UWDA
- GO
- Anatomy
- OMIM
- UMLS
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Anatomy
- Model organisms
- Other subdomains
UMLS: 3 components

- **Metathesaurus**
  - Concepts
  - Inter-concept relationships

- **Semantic Network**
  - Semantic types
  - Semantic network relationships

- **Lexical resources**
  - SPECIALIST Lexicon
  - Lexical tools
Addison’s Disease in medical vocabularies

◆ Synonyms: different terms
  - Addisonian syndrome
  - Bronzed disease
  - Addison melanoderma
  - Asthenia pigmentosa
  - Primary adrenal deficiency
  - Primary adrenal insufficiency
  - Primary adrenocortical insufficiency
  - Chronic adrenocortical insufficiency

◆ Contexts: different hierarchies
  - eponym
  - symptoms
  - clinical
  - variants

- Symptomatology of Addison’s Disease
- Medical vocabularies
Addison’s Disease: Concept

A disease characterized by hypotension, weight loss, anorexia, weakness, and sometimes a bronze-like melanotic hyperpigmentation of the skin. It is due to tuberculosis- or autoimmune-induced disease (hypofunction) of the adrenal glands that results in deficiency of aldosterone and cortisol. In the absence of replacement therapy, it is usually fatal.
Metathesaurus Concepts (2005AA)

- **Concept** (~ 1.2M) CUI
  - Set of synonymous concept names

- **Term** (~ 4.2M) LUI
  - Set of normalized names

- **String** (~ 4.7M) SUI
  - Distinct concept name

- **Atom** (~ 5.5M) AUI
  - Concept name in a given source

---

<table>
<thead>
<tr>
<th>CUI</th>
<th>Term</th>
<th>String</th>
<th>Atom</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0000001</td>
<td>headache (source 1)</td>
<td>S0000003</td>
<td>L0000002</td>
</tr>
<tr>
<td>A0000002</td>
<td>headache (source 2)</td>
<td>S0000003</td>
<td>L0000002</td>
</tr>
<tr>
<td>A0000003</td>
<td>Headache (source 1)</td>
<td>S0000003</td>
<td>L0000002</td>
</tr>
<tr>
<td>A0000004</td>
<td>Headache (source 2)</td>
<td>S0000003</td>
<td>L0000002</td>
</tr>
<tr>
<td>A0000005</td>
<td>Cephalgia (source 1)</td>
<td>S0000003</td>
<td>L0000002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Cluster of synonymous terms

<table>
<thead>
<tr>
<th>Concept</th>
<th>Term</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C0001621</td>
<td>S0011232 Adrenal Gland Diseases</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0011231 Adrenal Gland Disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0000441 Disease of adrenal gland</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0481705 Disease of adrenal gland, NOS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0220090 Disease, adrenal gland</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0044801 Gland Disease, Adrenal</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0860744 Disorder of adrenal gland, unspecified</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0217833 Unspecified disorder of adrenal glands</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0225481 ADRENAL DISORDER</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0627685 DISORDER ADRENAL (NOS)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0632950 Disorder of adrenal gland</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0354509 Adrenal Gland Disorders</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0586222 Adrenal disease</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0466921 ADRENAL DISEASE, NOS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1520972 Nebennierenkrankheiten</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S0226798 SURRENALE, MALADIES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*GER* GER

*FRE* FRE
**Metathesaurus Evolution over time**

- **Concepts never die (in principle)**
  - CUIs are permanent identifiers

- **What happens when they do die (in reality)?**
  - Concepts can merge or split
  - Resulting in new concepts and deletions
Metathesaurus Relationships

- Symbolic relations: ~9 M pairs of concepts
- Statistical relations: ~7 M pairs of concepts (co-occurring concepts)
- Mapping relations: 100,000 pairs of concepts

Categorization: Relationships between concepts and semantic types from the Semantic Network
Diseases of the endocrine system

Diseases of the Adrenal Glands

Addison’s Disease
Endocrine Diseases

Adrenal Gland Diseases

Adrenal Gland Hypofunction

Addison’s Disease

Diseases

Endocrine Diseases
Primary adrenocortical insufficiency

Other disorders of adrenal gland

Disorders of other endocrine gland

ICD-10
Organize concepts

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

Hypoadrenalism

Adrenal Gland Hypofunction

Adrenal cortical hypofunction

Addison’s Disease

Endocrine Diseases

Adrenal Gland Diseases

Adrenal Cortex Diseases

Hypoadrenalism

Adrenal Gland Hypofunction

Adrenal cortical hypofunction

Addison’s Disease

organize concepts
Symbolic relations

- Relation
  - Pair of “atom” identifiers
  - Type
  - Attribute (if any)
  - List of sources (for type and attribute)

- Semantics of the relationship:
  defined by its type [and attribute]

Source transparency: the information is recorded at the “atom” level
Symbolic relationships

- Hierarchical
  - Parent / Child
  - Broader / Narrower than

- Derived from hierarchies
  - Siblings (children of parents)

- Associative
  - Other

- Various flavors of near-synonymy
  - Similar
  - Source asserted synonymy
  - Possible synonymy

Type:

- PAR/CHD
- RB/RN
- SIB
- RO
- RL
- SY
- RQ
Symbolic relationships  Attribute

◆ Hierarchical
  • isa (is-a-kind-of)
  • part-of

◆ Associative
  • location-of
  • caused-by
  • treats
  • ...

◆ Cross-references (mapping)
Terminology integration methods
How do they do that?

- Integrating terms
  *Lexical knowledge*
- Categorizing concepts
  *Semantic pre-processing*
- Integrating relations
  *Recording relations*
- Editing and auditing
  *UMLS editors*
Terminology integration methods

Lexical knowledge
Lexical knowledge

Adrenal gland diseases
Adrenal disorder
Disorder of adrenal gland
Diseases of the adrenal glands
C0001621
Normalization

- Remove genitive: Hodgkin’s diseases, NOS
- Remove stop words: Hodgkin diseases, NOS
- Lowercase: Hodgkin diseases,
- Strip punctuation: hodgkin diseases,
- Uninflect: hodgkin diseases
- Sort words: hodgkin disease
- Result: disease hodgkin
Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodkin's Disease
Disease, Hodgkin's
Hodkin's, disease
HODGKIN'S DISEASE
Hodkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
disease, Hodgkin

normalize

disease hodgkin
Lexical tools

◆ To manage lexical variation in biomedical terminologies

◆ Major tools
  ● Normalization
  ● Indexes
  ● Lexical Variant Generation program (lvg)

◆ Based on the SPECIALIST Lexicon

◆ Used by noun phrase extractors, search engines
Integrating terms

Examples

◆ Exact match
  ● Original term: Pain in back (b28013)
  ● Concept mapped to: Back Pain (C0003862)

  *Pain in back present in the Metathesaurus (from the Read Codes)*

◆ Match after normalization
  ● Original term: Pain in joints (b28016)
  ● Normalized term: joint pain
  ● Concept mapped to: Arthralgia (C0003862)

  *Joint pain is a synonym for Arthralgia*
Integrating terms

No match found

- Radiating pain in body part (b2801) Too general
- Radiating pain in a dermatome (b2803) Too specific
- Pain in stomach or abdomen (b28012) Coordination

*Population* (→ C0032659)
*Demographic change* (→ C0681668)
*Population density* (→ C0032665)
*other specified* Population
*unspecified* Population
## Integrating terms

### Multiple matches

- **Impulse control**
  - Impulse control (C0150632)
  - Impulse control training (C0262701)
  - Ability to control impulses (C0517616)

- **Frontal lobe**
  - frontal lobe (C0016733)
  - Entire frontal lobe (C1268977)

- **Bites**
  - Biting (C0005658)
  - 2-(4-ethoxybenzyl)-1-diethylaminoethyl-5-isothiocyanatobenzimidazole (C0045724)

*synonym for BIT alkylating agent*
Terminology integration methods

Semantic pre-processing
Semantic pre-processing

- Metadata in the source vocabularies
- Tentative categorization
- Positive (or negative) evidence for tentative synonymy relations based on lexical features
Semantic pre-processing in practice

- **Mapping between**
  - Semantic types (UMLS Semantic Network)
  - Semantics of a given subset of a terminology

- **Semantic Network**
  - 135 semantic types (high-level categories)
  - 2 hierarchies for Entity and Event
  - Examples
    - Disease or Syndrome
    - Body Part, Organ, or Organ Component
    - Mental Process
### UMLS Semantic Groups

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTI</td>
<td>Activities &amp; Behaviors</td>
</tr>
<tr>
<td>ANAT</td>
<td>Anatomy</td>
</tr>
<tr>
<td>CHEM</td>
<td>Chemicals &amp; Drugs</td>
</tr>
<tr>
<td>CONC</td>
<td>Concepts &amp; Ideas</td>
</tr>
<tr>
<td>DEVI</td>
<td>Devices</td>
</tr>
<tr>
<td>DISO</td>
<td>Disorders</td>
</tr>
<tr>
<td>GENE</td>
<td>Genes &amp; Molecular Sequences</td>
</tr>
<tr>
<td>GEOG</td>
<td>Geographic Areas</td>
</tr>
<tr>
<td>LIVB</td>
<td>Living Beings</td>
</tr>
<tr>
<td>OBJC</td>
<td>Objects</td>
</tr>
<tr>
<td>OCCU</td>
<td>Occupations</td>
</tr>
<tr>
<td>ORGA</td>
<td>Organizations</td>
</tr>
<tr>
<td>PHEN</td>
<td>Phenomena</td>
</tr>
<tr>
<td>PHYS</td>
<td>Physiology</td>
</tr>
<tr>
<td>PROC</td>
<td>Procedures</td>
</tr>
</tbody>
</table>

- **Acquired Abnormality**
- **Anatomical Abnormality**
- **Cell or Molecular Dysfunction**
- **Congenital Abnormality**
- **Disease or Syndrome**
- **Experimental Model of Disease**
- **Finding**
- **Injury or Poisoning**
- **Mental or Behavioral Dysfunction**
- **Neoplastic Process**
- **Pathologic Function**
- **Sign or Symptom**
Semantic areas in ICF

**b** BODY FUNCTIONS
- Physiology
- Sign or Symptom
- Finding
- Biologic Function
- Individual Behavior

**s** BODY STRUCTURES
- Anatomy

**d** ACTIVITIES AND PARTICIPATION
- Physiology
- Activities & Behaviors
- Machine Activity
- Sign or Symptom
- Finding
- Educational Activity
Semantic areas in ICF

- **e1** PRODUCTS AND TECHNOLOGY
  - ????
- **e2** NATURAL ENVIRONMENT AND HUMAN-MADE CHANGES TO ENVIRONMENT
  - Phenomena
- **e3** SUPPORT AND RELATIONSHIPS
  - Family Group
  - Population Group
  - Professional or Occupational Group
- **e4** ATTITUDES
  - ????
- **e5** SERVICES, SYSTEMS AND POLICIES
  - Governmental or Regulatory Activity
  - Regulation or Law
Semantic pre-processing Examples

- **BODY FUNCTIONS**
  - Physiology
  - Sign or Symptom
  - Finding
  - Biologic Function
  - Individual Behavior

- **Appetite (b1302)**

- **Desire for food (C0003618)**

**Validation**
Semantic pre-processing

Examples

- Pharmacologic Substance
- Organic Chemical

b BODY FUNCTIONS
- Physiology
- Sign or Symptom
- Finding
- Biologic Function
- Individual Behavior

BIT (C0045724)

Biting (b5101)

Bites (C0005658)

No match

no semantic match

no semantic match
Semantic pre-processing

Examples

- Mental Process
- Auditory Perception (C0004309)
  - exact match

- Listening (d115)
  - exact match
  - semantic match

- Auscultation (C0004339)
  - no semantic match

ACTIVITIES AND PARTICIPATION
- Physiology
- Activities & Behaviors
- Machine Activity
- Sign or Symptom
- Finding
- Educational Activity

Disambiguation
Terminology integration methods

Recording relations
Recording relations

◆ Relations
  ● Recorded at the term (atom) level
  ● Aggregated at the concept level

◆ Once integrated into the UMLS, ICF relations participate to the Metathesaurus graph

◆ Possibly redundant with relations from other sources
ICF relations in the Metathesaurus

- ICF hierarchical relations in the UMLS
  - REL: parent/child
  - RELA: none
  - SAB: ICF

- Other relations?
ICF relations in the Metathesaurus

- Pain (b280-b289)
  - Sensation of pain (b280)
    - Pain in back (b28013)
  - Back Pain (b28013)
- Pain (C0030193)
  - Pain finding at anatomical site (C0458087)
    - Pain of truncal structure (C0578056)
  - Back Pain (C0004604)
ICF relations in the Metathesaurus

Pain (C0030193)

- Sensation of pain (b280)
  - Pain finding at anatomical site (C0458087)
  - Pain of truncal structure (C0578056)
- Back Pain (C0004604)
Terminology integration methods

Editing and auditing
Additional knowledge: UMLS editors

- Adrenal Gland Diseases
  - Adrenal Cortex Diseases
    - Adrenal Cortex Dysfunction
      - Hypoadrenalism
      - Adrenal Gland Hypofunction
        - Adrenal cortical hypofunction
          - Addison’s Disease
  - Other disorders of adrenal gland

Experience with ICF
Acknowledgments

- Marcy Harris
- Guergana Savova
Materials

◆ ICF: 1495 terms
  ● 478 terms filtered out
    ■ 218 terms with other specified
    ■ 217 terms with unspecified
    ■ 37 terms with other specified and unspecified
    ■ 2 terms with specified (alone)
    ■ 1 term with other specified (alone)
  ● 1017 terms remaining

◆ UMLS: version 2004AA
Mapping to UMLS Metathesaurus

◆ Methods
  ● Exact match first
  ● Normalized match, if necessary

◆ Results
  ● 359 ICF terms mapped (35%)
  ● 658 terms without mapping
Mapping by category

- b: 32% BODY FUNCTIONS
- d: 26% ACTIVITIES AND PARTICIPATION
- e: 21% ENVIRONMENTAL FACTORS
- s: 60% BODY STRUCTURES
Issues with mapping

- Phenomena preventing the terms form being mapped:
  - coordination with *and* alone: 147
    - Education *and* training policies (e5852)
  - coordination with *or* alone: 7
    - Pain in stomach *or* abdomen (b28012)
  - coordination with both *and* and *or*: 2
    - Assistive products *and* technology for the practice of religion *or* spirituality (e1451)
Semantic validation

◆ Method
  ● Correspondence between ICF chapters and UMLS semantic types/groups

◆ Issues
  ● Correspondence difficult to establish for some subgroups in ENVIRONMENTAL FACTORS
    ■ PRODUCTS AND TECHNOLOGY
    ■ ATTITUDES
Semantic validation Results

- single, valid
- multiple, one valid
- multiple, multiple valid
- one match, rejected
- multiple, all rejected
Issues with semantic validation

- Multiple “valid” matches must be reviewed by experts and disambiguated
- Rejected mappings
  - Semantically invalid UMLS concepts
  - or
  - Missing correspondence (ICF chapter/UMLS ST-SG)
Conclusions
Conclusions (1)

◆ Integrating ICF into the UMLS
  - Should not be too difficult
    - Relatively small
    - Many concepts already present in UMLS
  - Challenges
    - Underspecified terms
    - Coordination
    - Specific perspective
Conclusions (2)

◆ Integrating ICF into the UMLS
  ● Benefit for ICF
    ▪ Links to other vocabularies
    ▪ Facilitate downward extension
  ● Benefit for UMLS
    ▪ Adds specific perspective
Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

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