Semantic Data Integration Workshop
Amsterdam, The Netherlands
May 18, 2009

UMLS and semantic integration

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

◆ Unified Medical Language System overview
  ● UMLS Metathesaurus
  ● UMLS Semantic Network

◆ Data integration questions
Uses of biomedical ontologies

- **Knowledge management**
  - Annotating data and resources
  - Accessing biomedical information
  - Mapping across biomedical ontologies

- **Data integration, exchange and semantic interoperability**

- **Decision support**
  - Data selection and aggregation
  - Decision support
  - NLP applications
  - Knowledge discovery

[Bodenreider, YBMI 2008]
Unified Medical Language System

Overview
Motivation

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

«[...] 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.»
The UMLS in practice

◆ Database
  ● Series of relational files

◆ Interfaces
  ● Web interface: Knowledge Source Server (UMLSKS)
  ● Application programming interfaces
    (Java and XML-based)

◆ Applications
  ● lvg (lexical programs)
  ● MetamorphoSys (installation and customization)
  ● RRF browser (browsing subsets)

The UMLS is not an end-user application
UMLS 3 components

- **Lexical resources**
  - SPECIALIST Lexicon
  - Lexical tools

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

- **Semantic Network**
  - Semantic types
  - Semantic network relationships
UMLS Knowledge Sources

UMLS Metathesaurus
Metathesaurus Basic organization

◆ Concepts
  ● Synonymous terms are clustered into a concept
  ● Properties are attached to concepts, e.g.,
    ■ Unique identifier
    ■ Definition

◆ Relations
  ● Concepts are related to other concepts
  ● Properties are attached to relations, e.g.,
    ■ Type of relationship
    ■ Source
Source Vocabularies

- 152 source vocabularies
  - 19 languages
- Broad coverage of biomedicine
  - 9.7M names
  - 2.1M concepts
  - >10M relations
- 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)
  - oncology (PDQ)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART)
  - primary care (ICPC)

- Terminology of knowledge bases (AI/Rheum, DXplain, QMR)
Integrating subdomains

Clinical repositories

Genetic knowledge bases

Other subdomains

SNOMED CT

OMIM

Biomedical literature

Model organisms

NCBI Taxonomy

FMA

MeSH

Genome annotations

Anatomy
Integrating subdomains

- Clinical repositories
- Genetic knowledge bases
- Biomedical literature
- Genome annotations
- Model organisms
- Anatomy
- Other subdomains
Trans-namespaces integration

- **Addison's disease** (363732003)
- **Other subdomains**
- **Clinical repositories**
- **Genetic knowledge bases**
- **Other subdomains**
- **Snomed CT**
- **UMLS**
- **OMIM**
- **MeSH**
- **Biomedical literature**
- **Addison Disease** (D000224)
- **Model organisms**
- **NCBI Taxonomy**
- **FMA**
- **GO**
- **Genome annotations**
- **Anatomy**
- **Biomedical literature**
- **Clinical repositories**
Addison’s Disease: Concept

An adrenal disease characterized by the progressive destruction of the adrenal cortex, resulting in insufficient production of aldosterone and hydrocortisone. Clinical symptoms include anorexia; nausea; weight loss; muscle weakness; and hyperpigmentation of the skin due to increase in circulating levels of ACTH precursor hormone which stimulates melanocytes.
Metathesaurus Relationships

- Symbolic relations: ~8 M pairs of concepts
- Statistical relations: ~6 M pairs of concepts (co-occurring concepts)
- Mapping relations: ~150,000

- Categorization: Relationships between concepts and semantic types from the Semantic Network
Heart

Concepts

Metathesaurus

Semantic Types

Anatomical Structure

Fully Formed Anatomical Structure

Embryonic Structure

Body Part, Organ or Organ Component

Disease or Syndrome

Pharmacologic Substance

Population Group

Semantic Network

Anatomical Structure

Mediastinum

Saccular Viscus

Esophagus

16

Left Phrenic Nerve

Heart Valves

5

Heart

13

Fetal Heart

22

Angina Pectoris

49

Cardiotonic Agents

237

Tissue Donors

38

Population Group
UMLS Knowledge Sources

UMLS Semantic Network
Semantic Network

- Semantic network relationships (54)
  - hierarchical (isa = is a kind of)
    - among types
      - Animal isa Organism
      - Enzyme isa Biologically Active Substance
    - among relations
      - treats isa affects
  - non-hierarchical
    - Sign or Symptom diagnoses Pathologic Function
    - Pharmacologic Substance treats Pathologic Function
“Biologic Function” hierarchy (isa)

- Biologic Function
  - Physiologic Function
    - Organ Function
    - Organ or Tissue Function
    - Cell Function
    - Molecular Function
  - Pathologic Function
    - Cell or Molecular Dysfunction
    - Disease or Syndrome
    - Experimental Model of Disease

- Mental Process
- Genetic Function
- Mental or Behavioral Dysfunction
- Neoplastic Process
Why a semantic network?

- Semantic Types serve as high level categories assigned to Metathesaurus concepts, independently of their position in a hierarchy.

- A relationship between 2 Semantic Types (ST) is a possible link between 2 concepts that have been assigned to those STs.
  - The relationship may or may not hold at the concept level.
  - Other relationships may apply at the concept level.
Relationships can inherit semantics

Semantic Network

- Fully Formed Anatomical Structure
- Body Part, Organ, or Organ Component
- Disease or Syndrome
- Adrenal Cortex
- Adrenal Cortical hypofunction
- Metathesaurus

 isa

location of

isa

isa

isa
UMLS and semantic integration

Data integration questions
Semantic interoperability through the UMLS

- **Metathesaurus:**
  Terminology/ontology integration
  - Terms from various terminologies linked through UMLS

- **Semantic Network:**
  Top domain ontology
  - Framework for semantic categorization of concepts
  - Template for potential relations among concepts
Potential contribution of UMLS to integration

◆ Data consistency
  ● SN as a source of domain and range constraints for relations

◆ Data query
  ● Resolve terms into concepts
  ● Source of synonymy
  ● [Lexical variants, normalization]

◆ Service query

◆ Service interoperability
Potential contribution of UMLS to integration

- **Provenance**
  - Rich source of metadata about terms

- **Data integration**
  - Map terms/concepts across vocabularies
  - Data integration through terminology integration

- **Semantic mediation**
  - UMLS as a the global schema

- **Reasoning**
  - Limited

[Mougin, DILS 2008]
Data, metadata and semantics

- Not specifically in UMLS
- caBIG
  - Cancer Biomedical Informatics Grid
  - National Cancer Institute
  - Cancer Data Standards Registry and Repository (caDSR)
    - Common data elements
    - Metadata repository
Use of the Metathesaurus in applications

- Indexing, semantic annotation, coding
- Mapping across vocabularies
- Aggregation
- Support for Natural Language Processing applications (entity recognition)
- Source of value sets for information models
Use of the Semantic Network in applications

- Partition concepts into subdomains
  - Aggregation
- Support for Natural Language Processing applications (language understanding)
- Consistency checking of relations
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
References

◆ UMLS
  umlsinfo.nlm.nih.gov

◆ UMLS browsers
  (free, but UMLS license required)
  • Knowledge Source Server: umlsks.nlm.nih.gov
  • Semantic Navigator:
  • RRF browser
    (standalone application distributed with the UMLS)
References

◆ Recent overviews

