



## *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

Lexical resources

Terminological resources

Ontological resources

# 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

(2009AA)

- ◆ 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

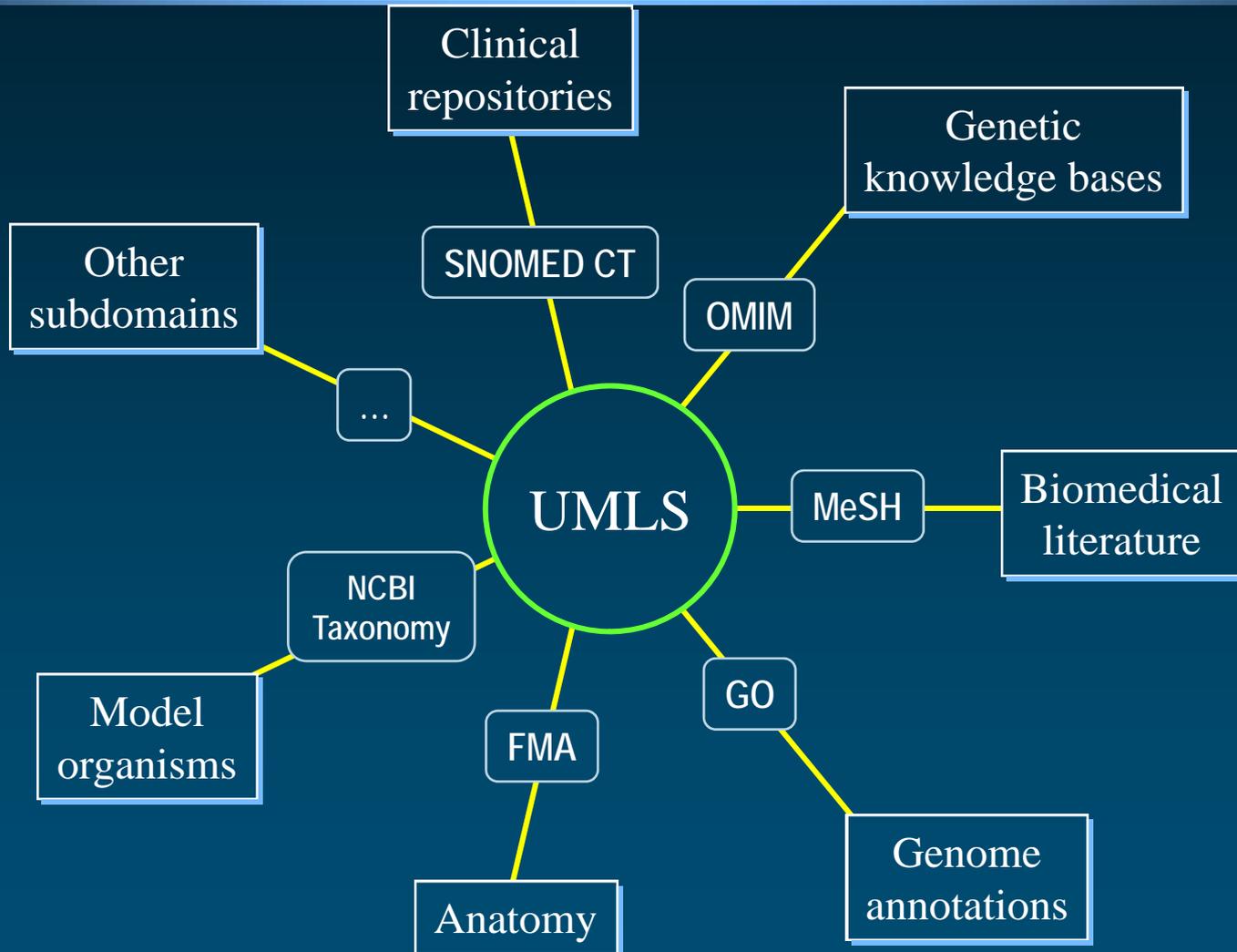
- clinical terms (SNOMED CT)
- information sciences (MeSH, CRISP)
- administrative terminologies (ICD-9-CM, CPT-4)
- data exchange terminologies (HL7, LOINC)



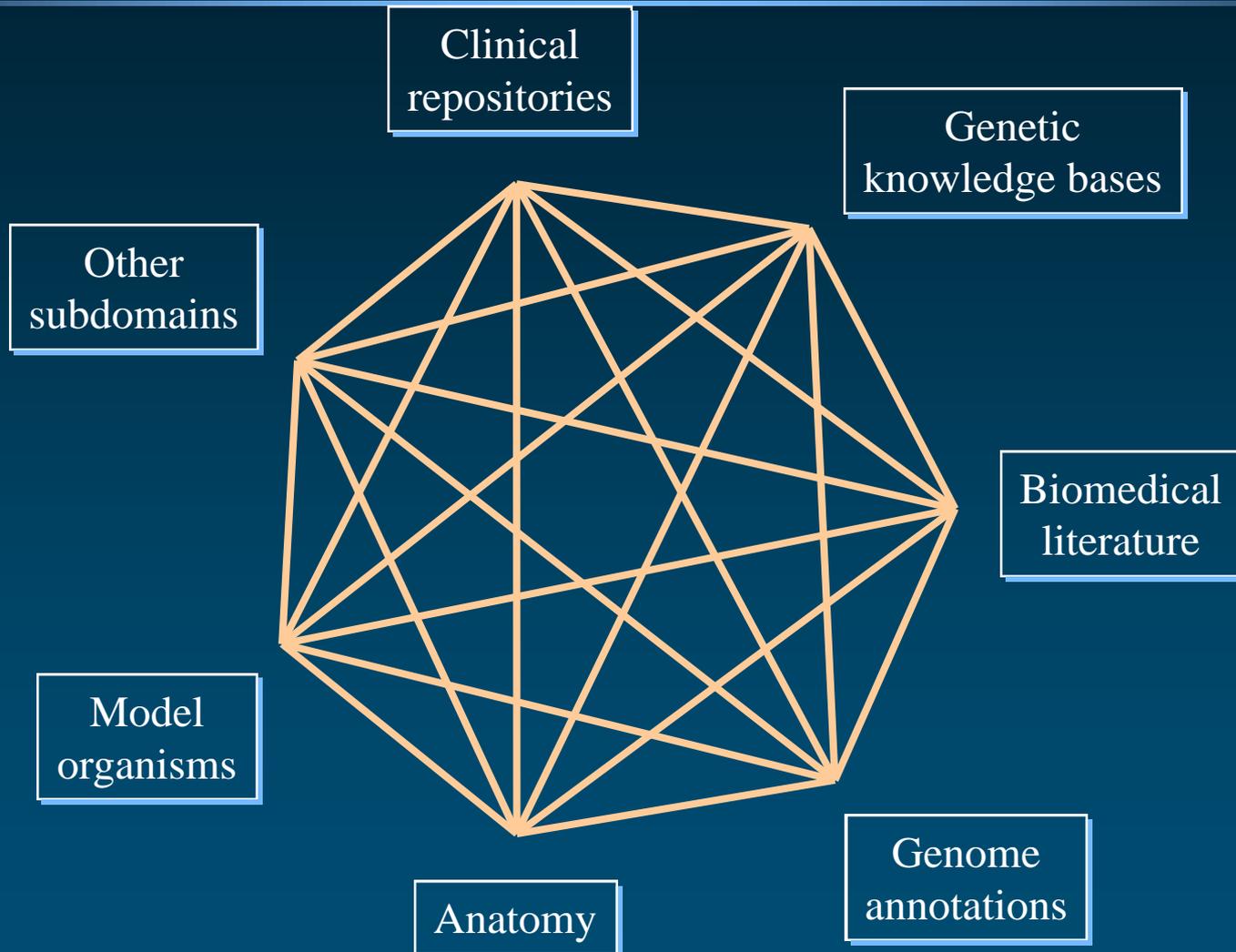
# Biomedical terminologies (cont'd)

- ◆ 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)

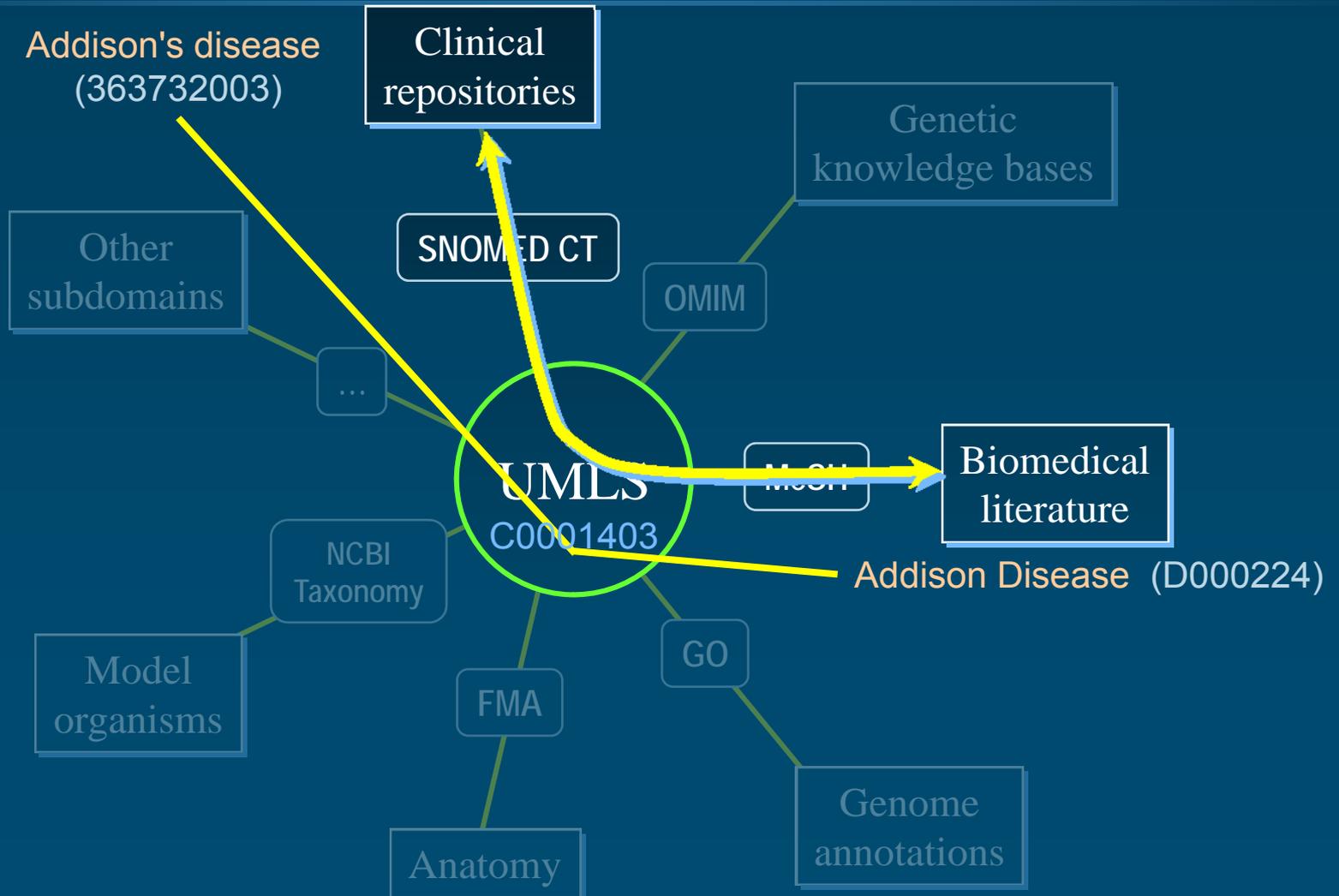
# Integrating subdomains



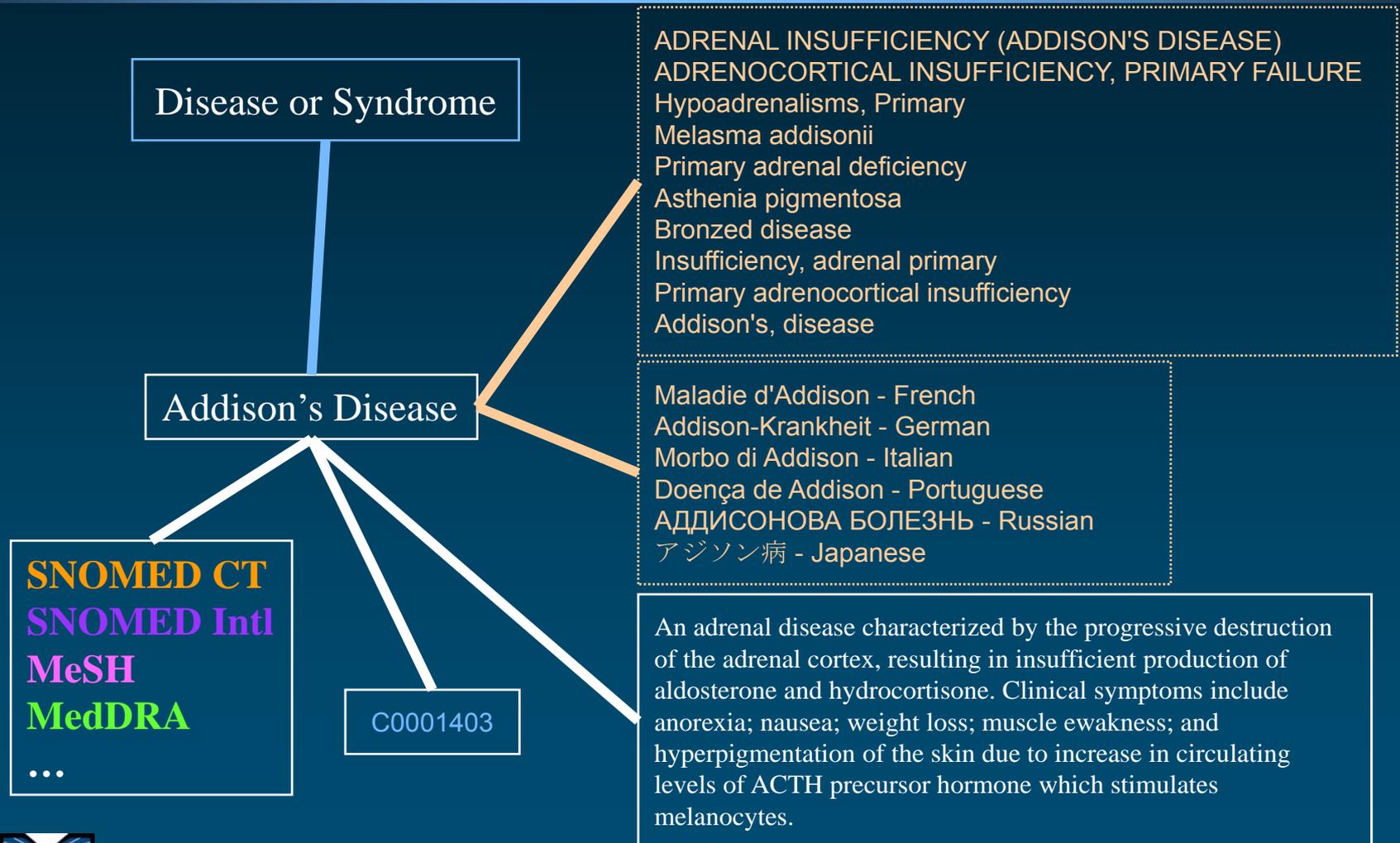
# Integrating subdomains



# Trans-namespace integration



# Addison's Disease: Concept



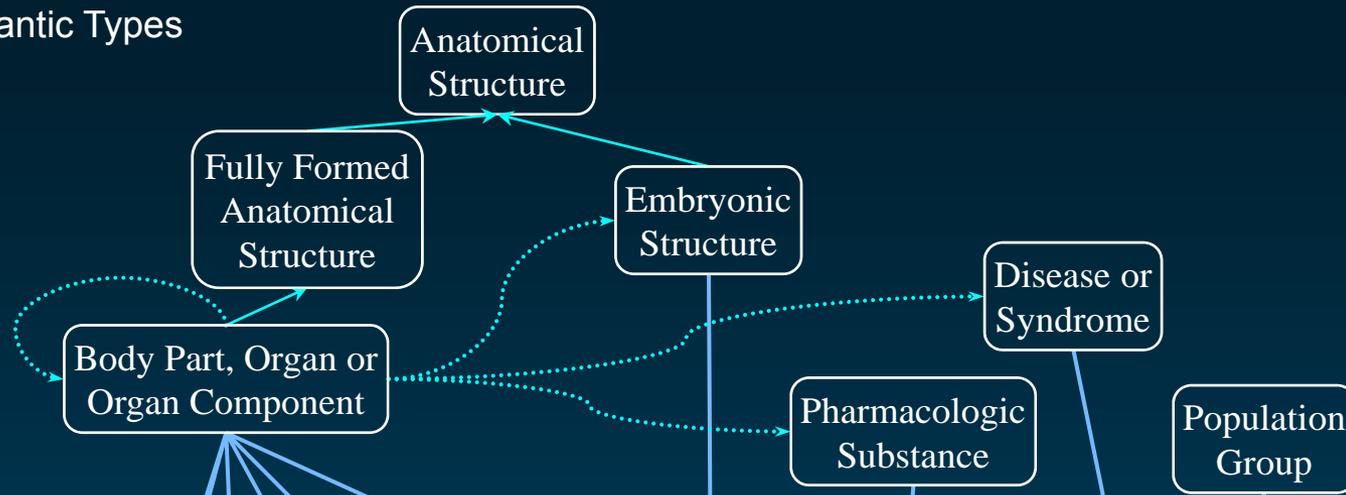
# Metathesaurus Relationships

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

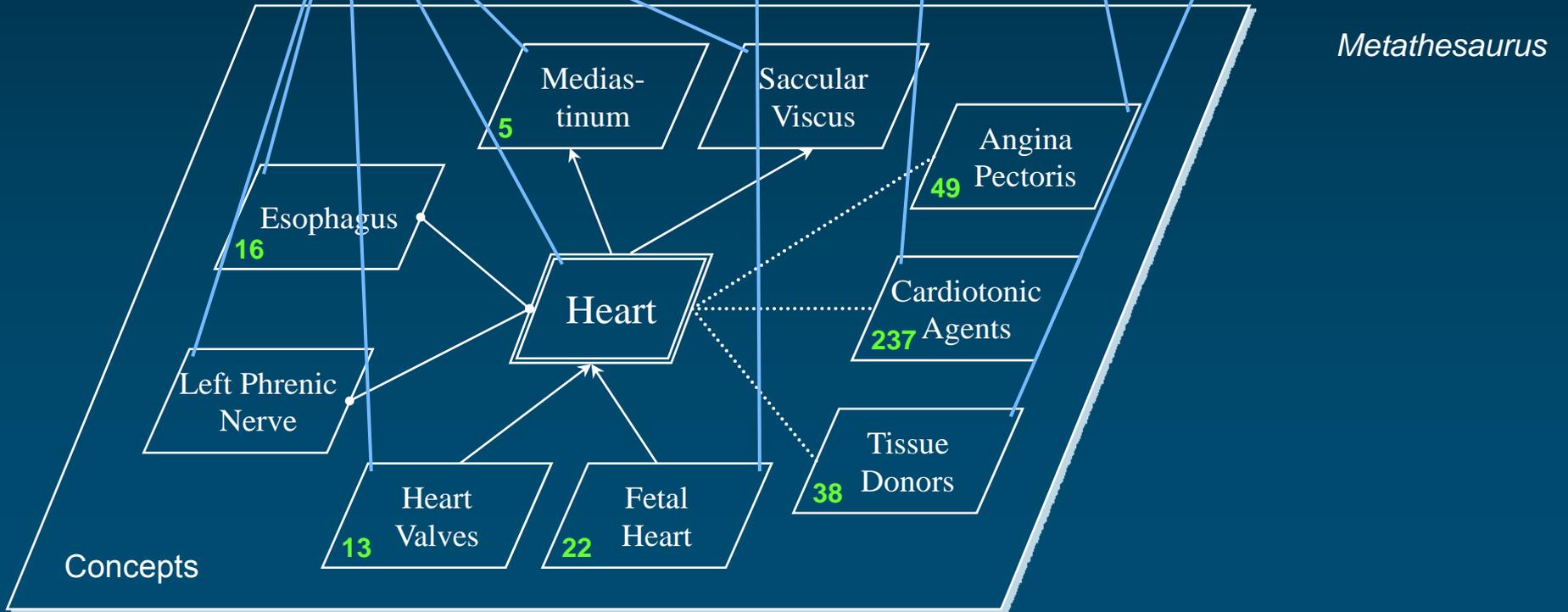
- 
- ◆ Categorization: Relationships between concepts and semantic types from the Semantic Network



Semantic Types



*Semantic Network*



*Metathesaurus*

# UMLS Knowledge Sources

*UMLS Semantic Network*

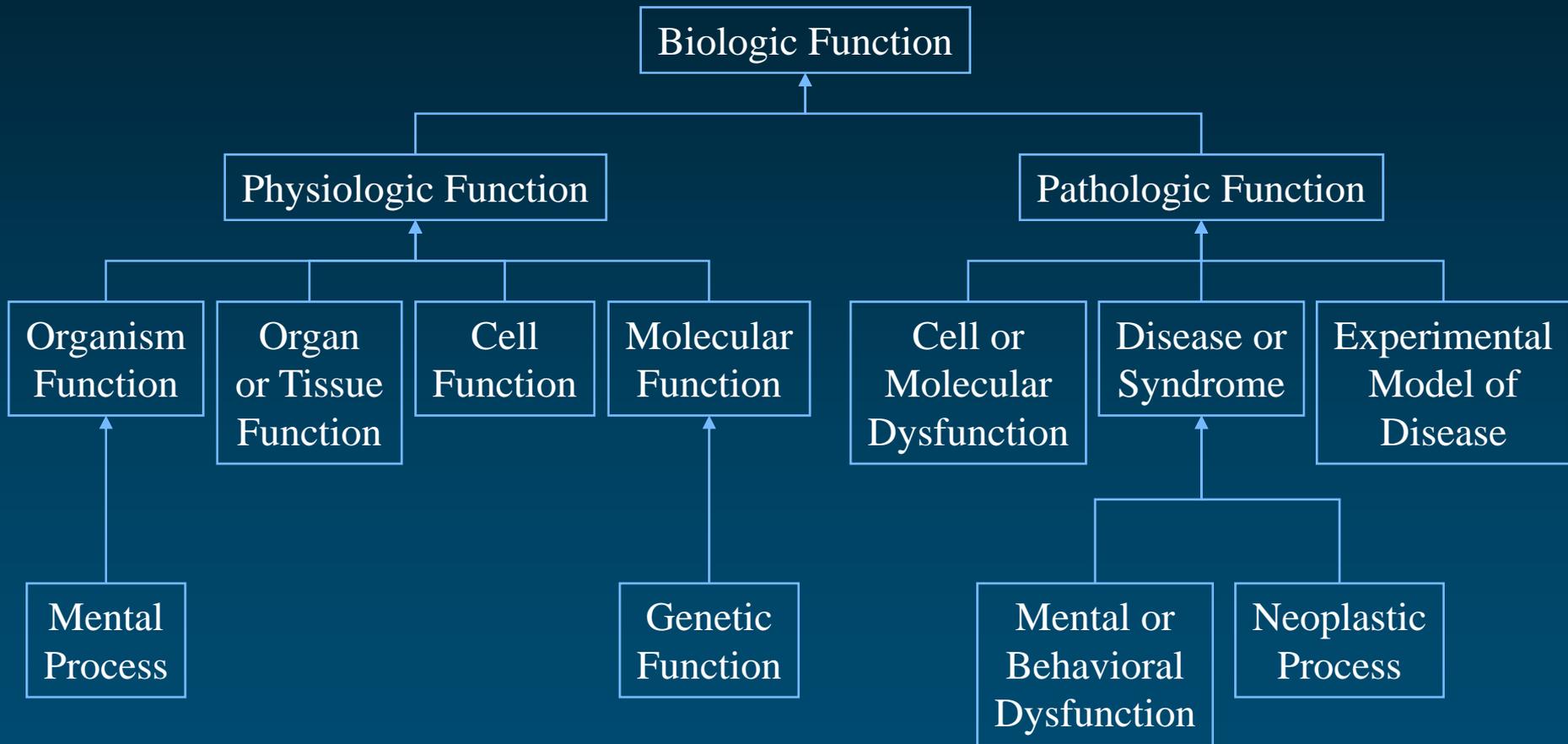
# Semantic Network

## ◆ Semantic network relationships (54)

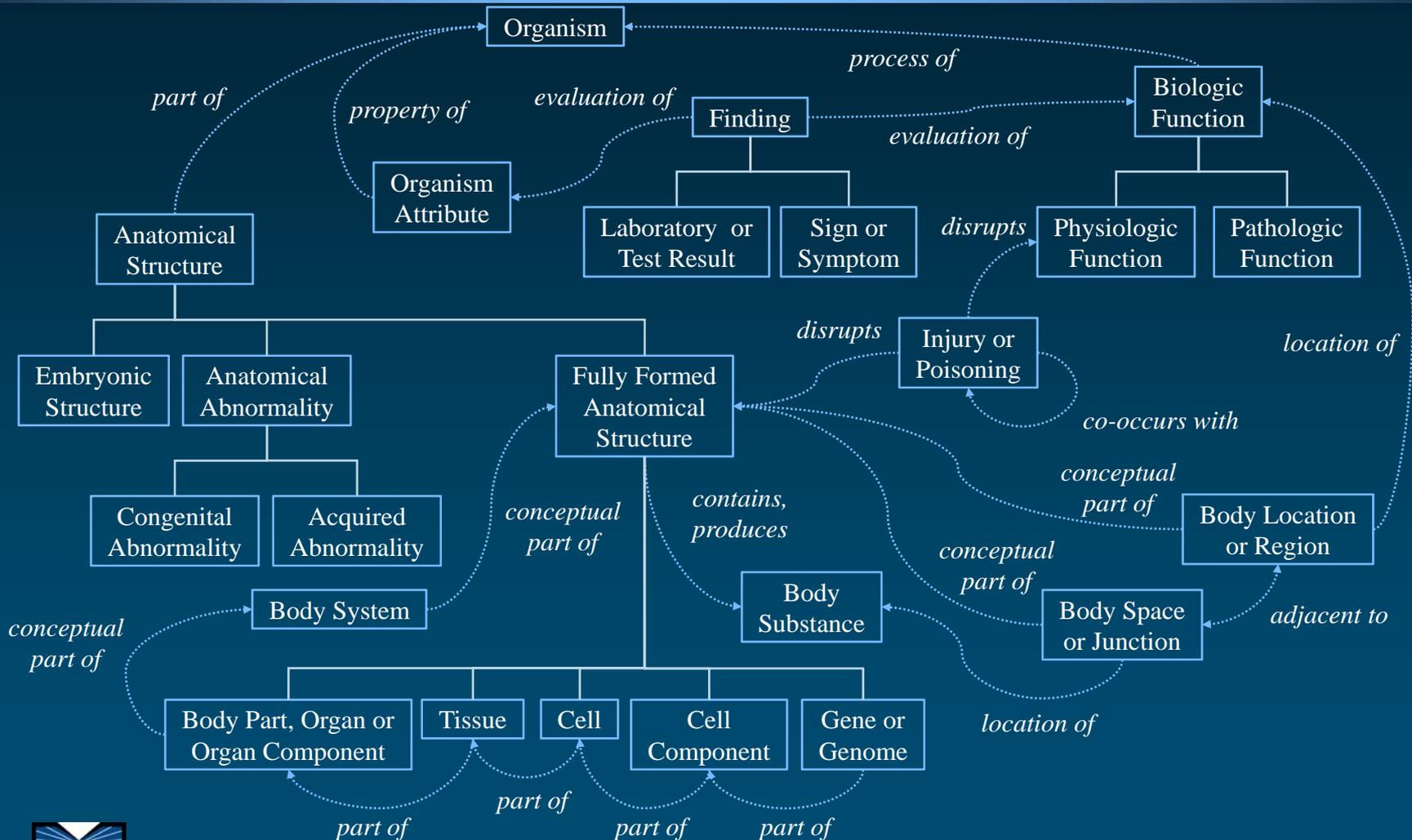
- hierarchical (isa = is a kind of)
  - among types
    - *Animal isa Organism*
    - *Enzyme isa Biologically Active Substance*
  - among relations
    - *treats isa affects*
- non-hierarchical
  - *Sign or Symptom diagnoses Pathologic Function*
  - *Pharmacologic Substance treats Pathologic Function*



# “Biologic Function” hierarchy (isa)



# Associative (non-isa) relationships

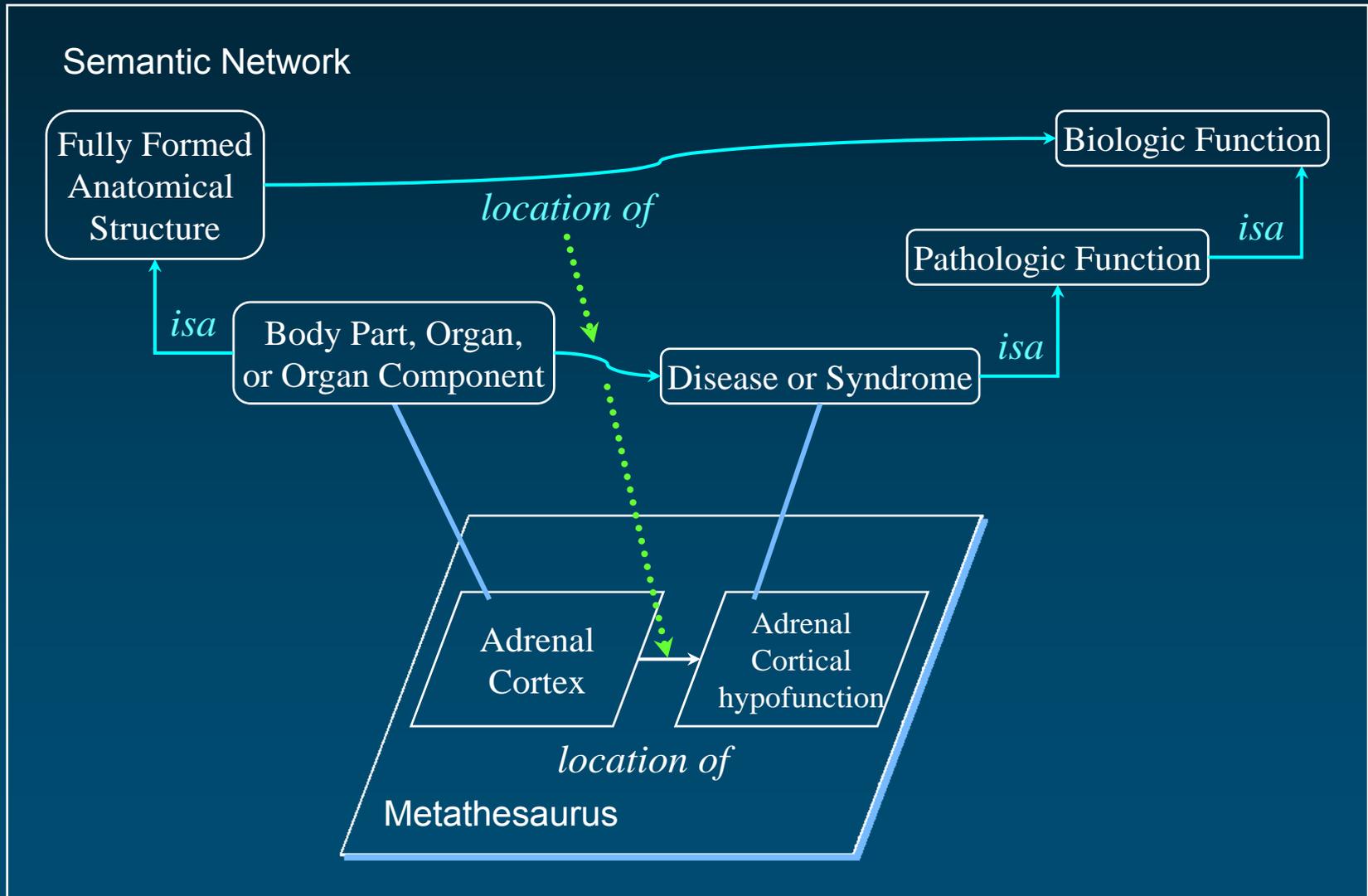


# Why a semantic network?

- ◆ 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



# 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

[Mougin, DILS 2008]

## ◆ Reasoning

- Limited



# Data, metadata and semantics

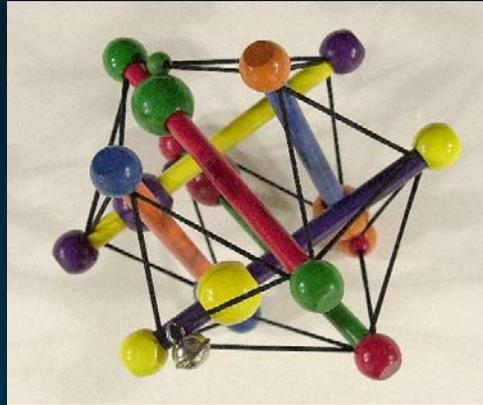
- ◆ Not specifically in UMLS
- ◆ caBIG
  - Cancer Biomedical Informatics Grid  
<http://cabig.cancer.gov/>
  - National Cancer Institute
  - Cancer Data Standards Registry and Repository (caDSR)  
[http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore\\_overview/cadsr](http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr)
    - 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](mailto:olivier@nlm.nih.gov)

Web: [mor.nlm.nih.gov](http://mor.nlm.nih.gov)



*Olivier Bodenreider*

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

# References

## ◆ UMLS

[umlsinfo.nlm.nih.gov](http://umlsinfo.nlm.nih.gov)

## ◆ UMLS browsers

(free, but UMLS license required)

- Knowledge Source Server: [umlsks.nlm.nih.gov](http://umlsks.nlm.nih.gov)
- Semantic Navigator:  
<http://mor.nlm.nih.gov/perl/semnav.pl>
- RRF browser  
(standalone application distributed with the UMLS)



# References

## ◆ Recent overviews

- Bodenreider O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*; D267-D270.
- Bodenreider O. From terminology integration to information integration: Unified Medical Language System (UMLS). BioRDF Teleconference, W3C Semantic Web Health Care and Life Sciences Interest Group, June 5, 2006.  
<http://mor.nlm.nih.gov/pubs/pres/060605-BioRDF.pdf>



# Biodiversity in the UMLS

**Siblings**

**Concepts & Ideas**

- Biomass →
- Habitat →
- water environment →

**Geographic Areas**

- Forests →

**Phenomena**

- Ecological Systems, Closed →
- Food Chain →
- Wetlands →

(7 siblings)

[direct children

Environment

MSH

Ecosystem

MSH

Biodiversity

**Other Related Concepts**

**Concepts & Ideas**

- Biota →

(1 other related concept)

**Co-occurring Concepts**

**Activities & Behaviors**

- Aquaculture [6
- Automatic Data Processing [10
- Behavior, Animal [10]

**BCI** **Biodiversity** **LEGEND \***

Start again Apply new parameters

Restrict to vocabulary: Show all

Highlight vocabulary: Nothing

**Similar Concepts**

(none)

**Closest MeSH Terms**

Main Headings