



University of Pittsburgh

**The Center for Biomedical Informatics**

February 3, 2006

# Biomedical resources for text mining



*Olivier Bodenreider*

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

# Overview

- ◆ An example
- ◆ Three types of resources
  - Lexical resources
  - Terminological resources
  - Ontological resources
- ◆ Some issues



An example

*Neurofibromatosis 2*

# Neurofibromatosis 2

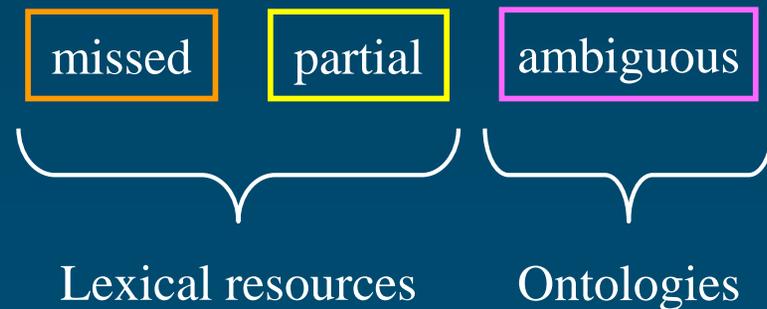
Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

[Uppal, S., and A. P. Coatesworth. "Neurofibromatosis Type 2." *Int J Clin Pract*, 57, no. 8, 2003, pp. 698-703.]



# Entity recognition

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.



# Relation extraction

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

- vestibular schwannomas *manifestation of* neurofibromatosis 2
- neurofibromatosis 2 *associated with* mutation of NF2 gene
- NF2 gene *located on* chromosome 22



# Resources for text mining

# Types of resources

## ◆ Lexical resources

- Collections of lexical items
- Additional information
  - Part of speech
  - Spelling variants
- Useful for entity recognition
- UMLS SPECIALIST Lexicon, WordNet

## ◆ Ontological resources

- Collections of
  - kinds of entities (substances, qualities, processes)
  - relations among them
- Useful for **relation extraction**
- UMLS Semantic Network, SNOMED CT



# Types of resources (revisited)

- ◆ Lexical and terminological resources
  - Mostly collections of names for biomedical entities
  - Often have some kind of hierarchical organization (e.g., relations)
- ◆ Ontological resources
  - Mostly collections of relations among biomedical entities
  - Sometimes also collect names



# Unified Medical Language System



## ◆ SPECIALIST Lexicon

- 200,000 lexical items
- Part of speech and variant information

## ◆ Metathesaurus

- 5M names from over 100 terminologies
- 1M concepts
- 16M relations

## ◆ Semantic Network

- 135 high-level categories
- 7000 relations among them

Lexical  
resources

Terminological  
resources

Ontological  
resources



Lexical resources

*SPECIALIST Lexicon*

# SPECIALIST Lexicon

- ◆ Content
  - English lexicon
  - Many words from the biomedical domain
- ◆ 200,000+ lexical items
- ◆ Word properties
  - morphology
  - orthography
  - syntax
- ◆ Used by the lexical tools



# SPECIALIST Lexicon record

```
{  
  base=hemoglobin      (base form)  
  spelling_variant=haemoglobin  
  entry=E0031208      (identifier)  
  cat=noun            (part of speech)  
  variants=uncount    (no plural)  
  variants=reg        (plural: hemoglobins , haemoglobins)  
}
```

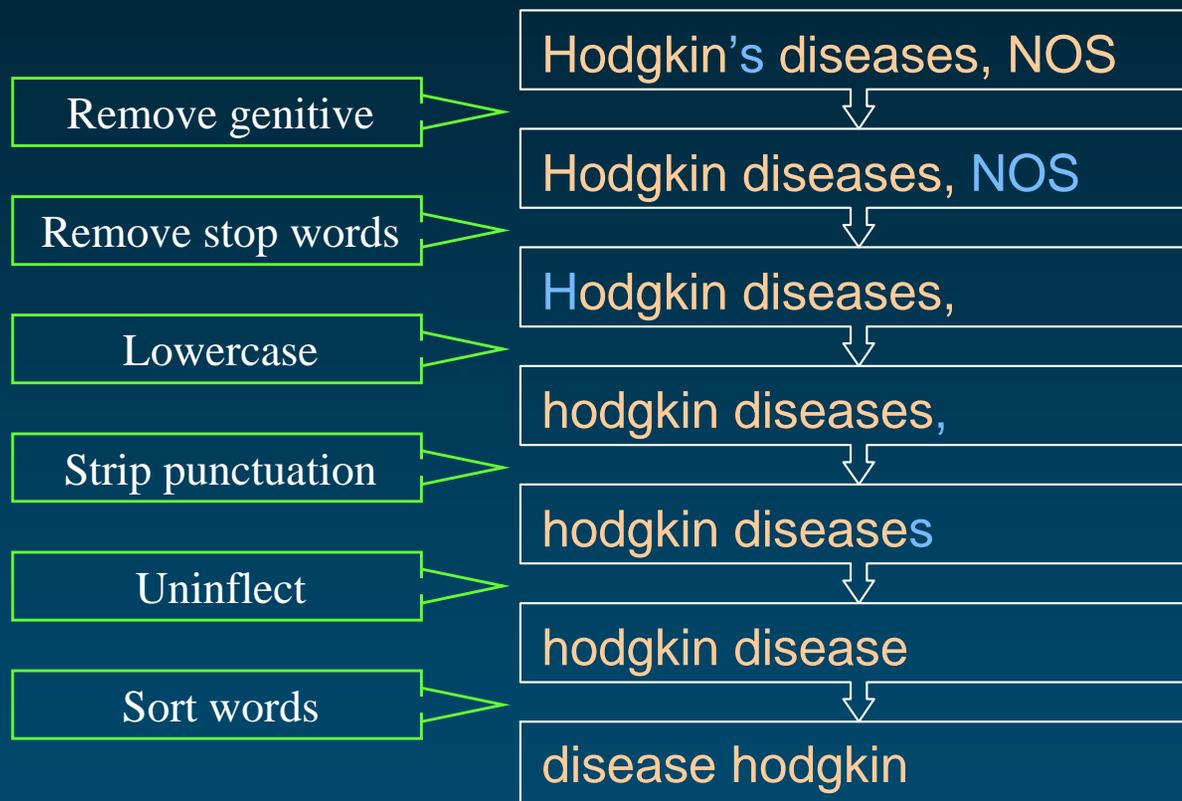


# Lexical tools

- ◆ 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



# Normalization



# Normalization: Example

Hodgkin Disease  
HODGKINS DISEASE  
Hodgkin's Disease  
Disease, Hodgkin's  
Hodgkin's, disease  
HODGKIN'S DISEASE  
Hodgkin'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



# Normalization Applications

- ◆ Model for lexical resemblance
- ◆ Help find lexical variants for a term
  - Terms that normalize the same usually share the same LUI
- ◆ Help find candidates to synonymy among terms
- ◆ Help map input terms to UMLS concepts



# Terminological resources

*UMLS Metathesaurus*

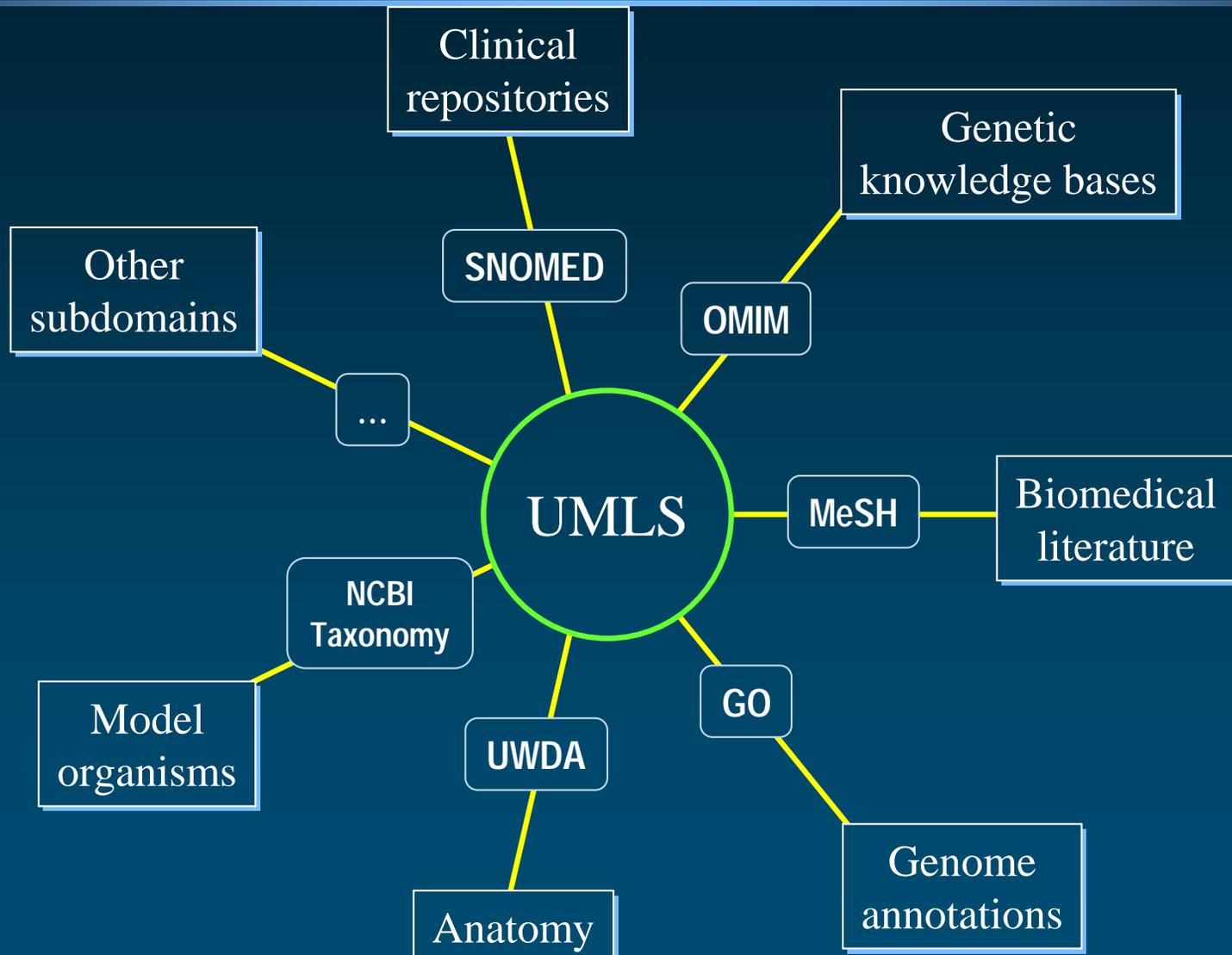
# Source Vocabularies

(2005AA)

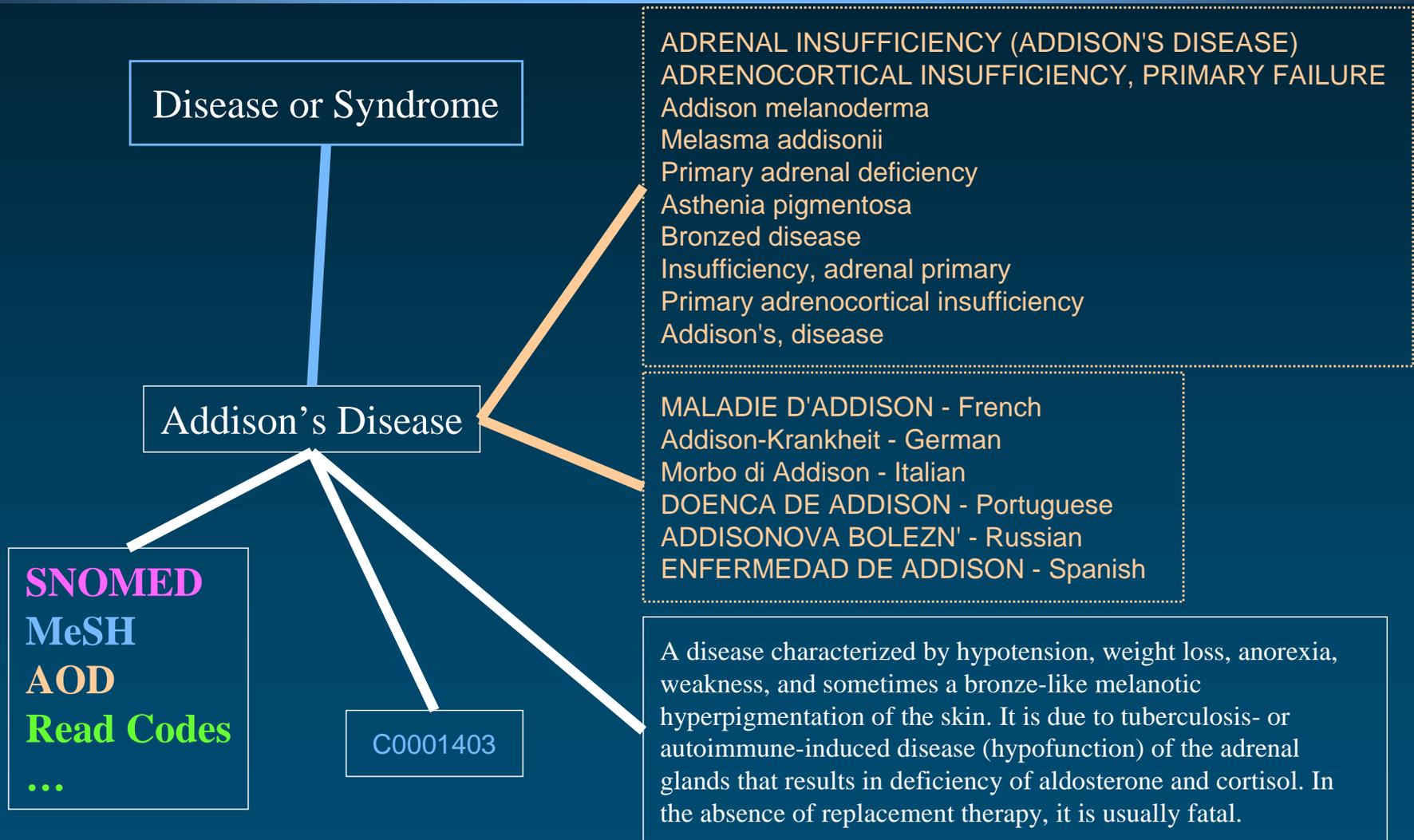
- ◆ 134 source vocabularies
  - 132 contributing concept names
- ◆ Broad coverage of biomedicine
  - 5M names
  - 1M concepts
  - 16M relations
- ◆ Common presentation



# Integrating subdomains

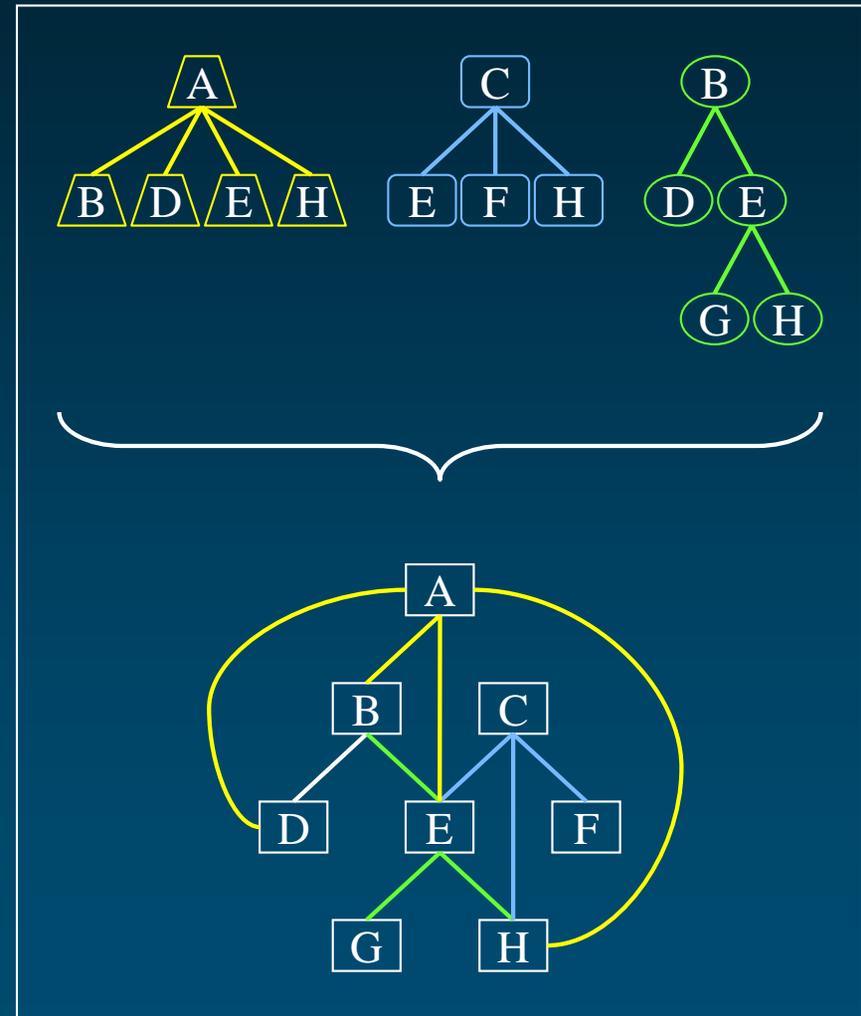


# Addison's Disease: Concept



# Organize concepts

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



# Metathesaurus concepts Examples

Neurofibromatosis type 2	s	C0027832	Neurofibromatosis 2
NF2	s	C0085114	Neurofibromatosis 2 genes
peripheral neurofibromatosis	s	C0027831	Neurofibromatosis 1
[bilateral] vestibular schwannomas	a	C0027859	Neuroma, Acoustic
mutation / mutations	s	C0026882	Mutation
gene	s	C0017337	Genes
merlin	m	C0254123	Neurofibromin 2
chromosome 22	s	C0008665	Chromosomes, Human, Pair 22



# Metaheasaurus relations Examples

## ◆ Neurofibromin 2

- Multiple parent concepts
  - Membrane proteins [MeSH]
  - Tumor suppressor proteins [MeSH]
  - Signaling protein [NCI Thesaurus]
- 1 child concept
  - Merlin, Drosophila [MeSH]
- Co-occurring concepts in MEDLINE
  - Neurofibromatosis 2 [13]
  - Membrane proteins [8]
  - ...



# Ontological resources

*UMLS Semantic Network*

# Semantic Network

## ◆ Semantic types (135)

- tree structure
- 2 major hierarchies
  - Entity
    - Physical Object
    - Conceptual Entity
  - Event
    - Activity
    - Phenomenon or Process

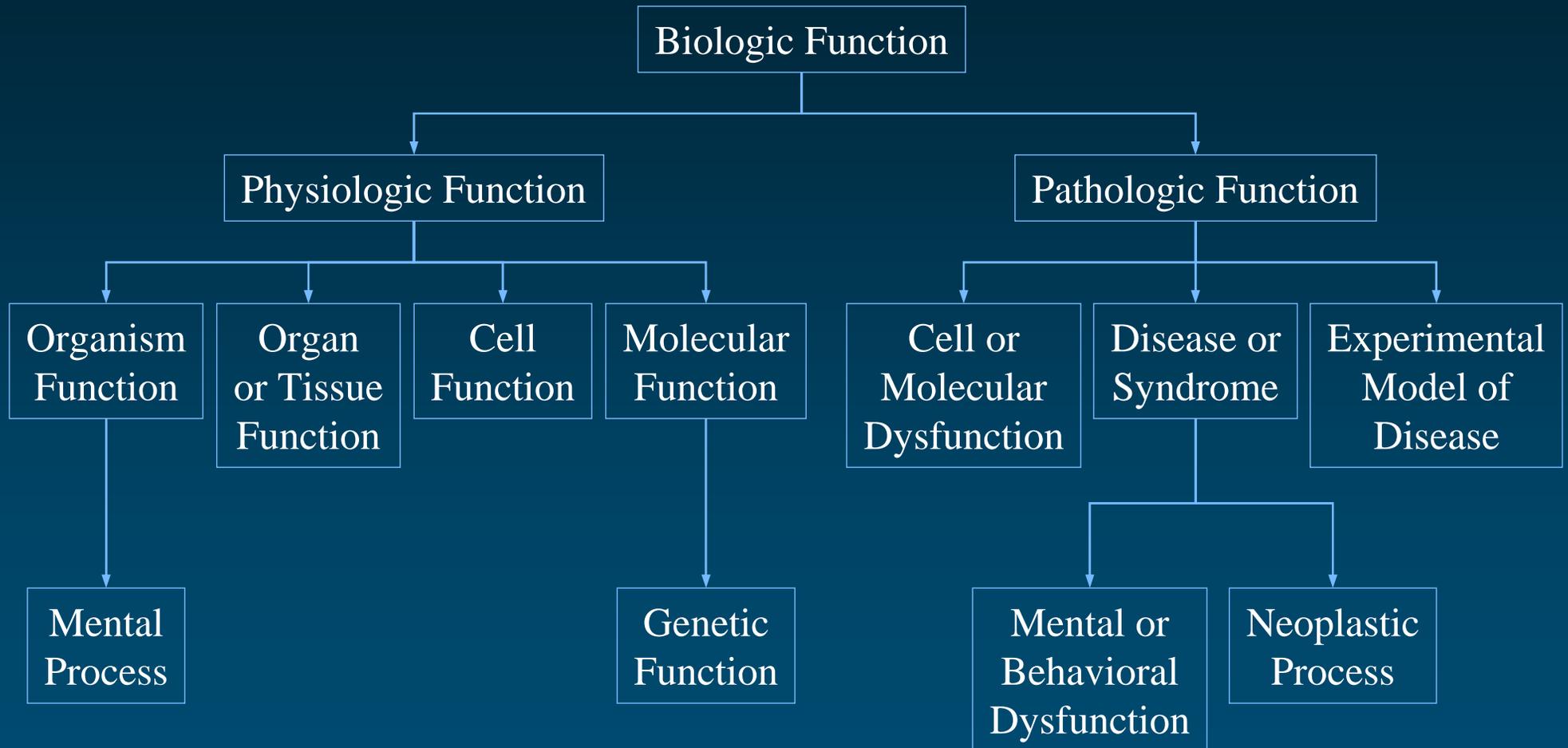


# Semantic Network

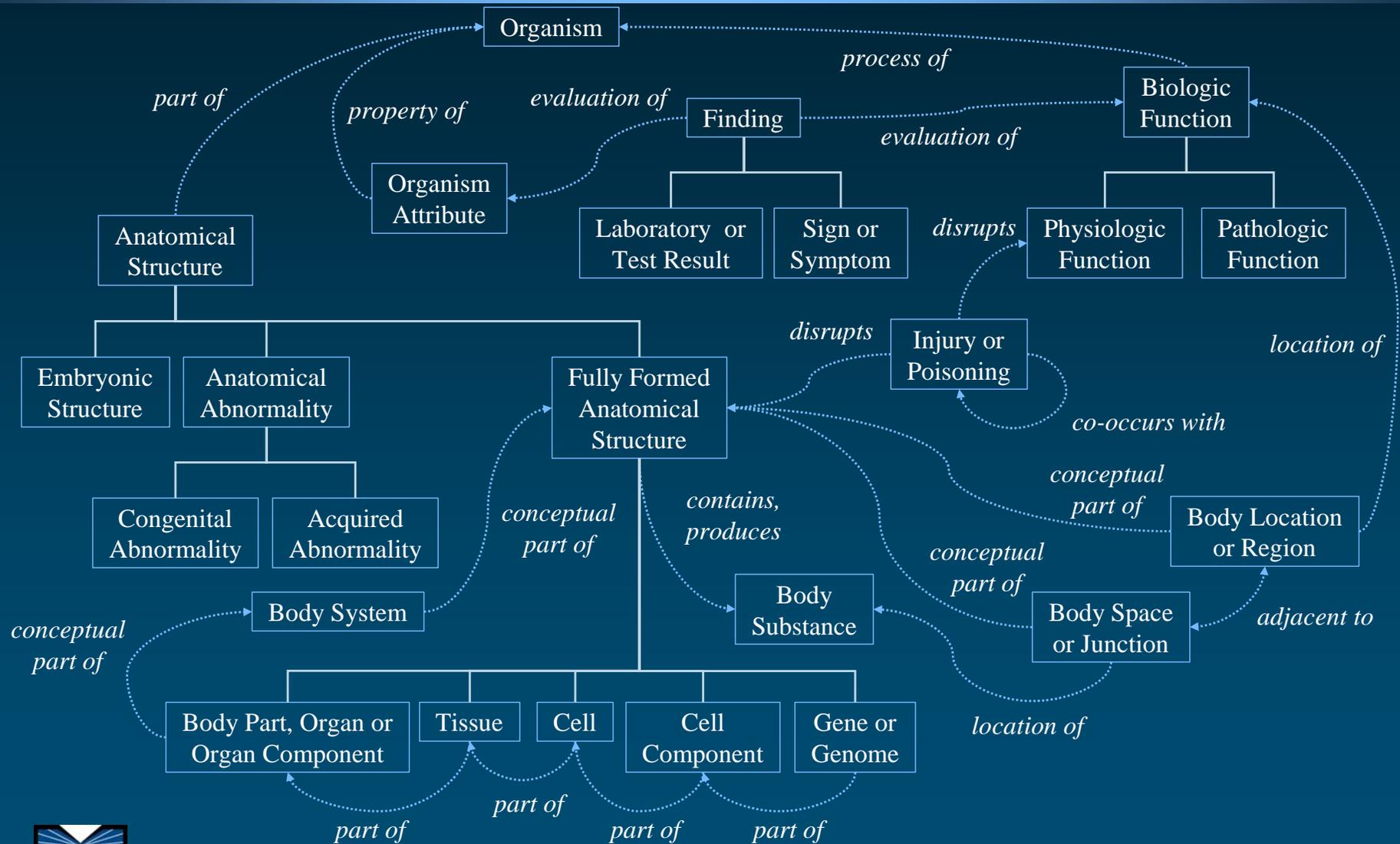
- ◆ Semantic network relationships (54)
  - hierarchical (isa = is a kind of)
    - among types
      - *Animal isa Organism*
      - *Enzyme isa Biologically Active Substance*
    - among relations
      - *treats isa affects*
  - non-hierarchical
    - *Sign or Symptom diagnoses Pathologic Function*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



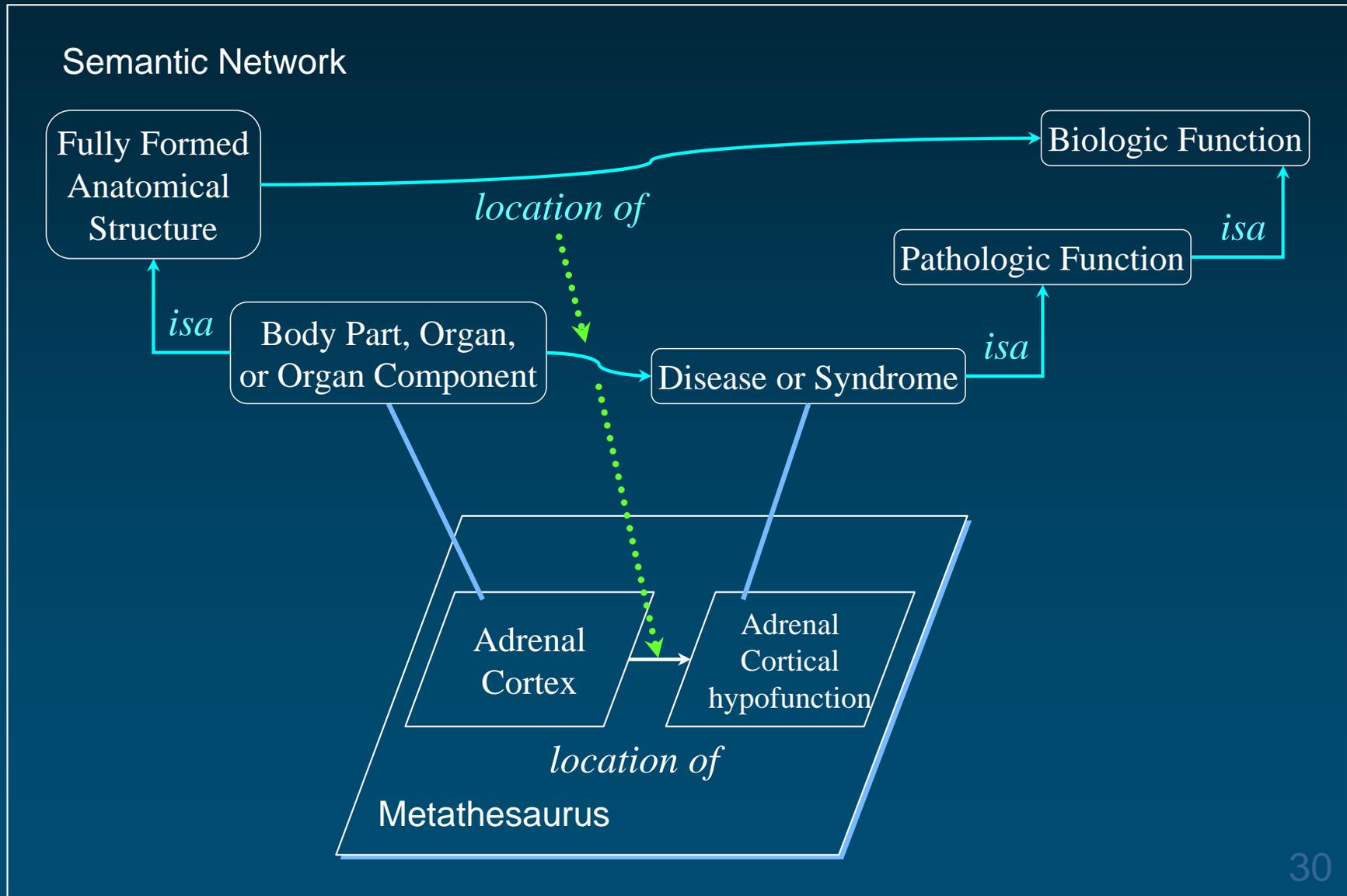
# “Biologic Function” hierarchy (isa)



# Associative (non-isa) relationships



# Relationships can inherit semantics



Semantic Types

Anatomical Structure

Fully Formed Anatomical Structure

Embryonic Structure

Disease or Syndrome

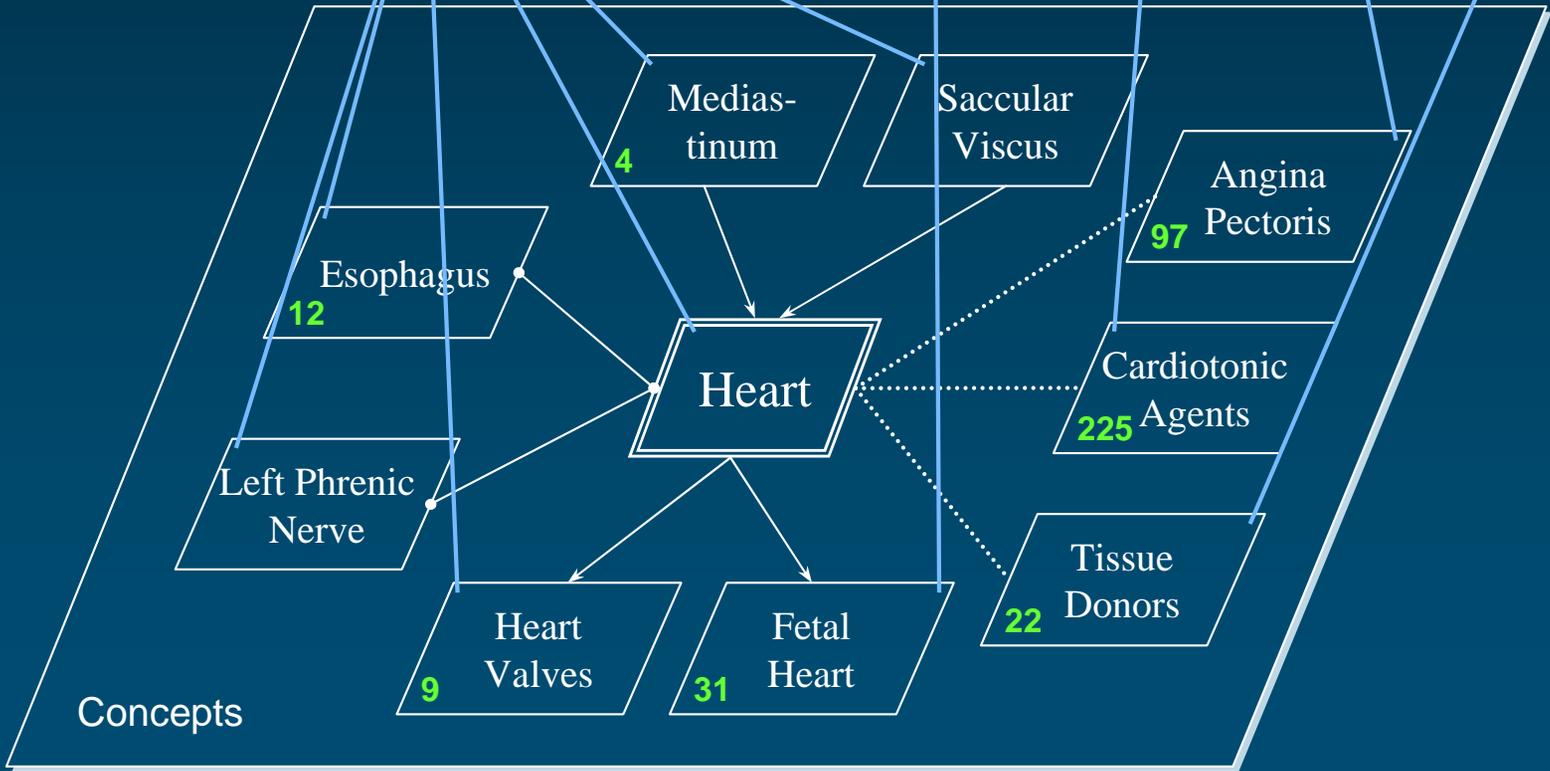
Body Part, Organ or Organ Component

Pharmacologic Substance

Population Group

Semantic Network

Metathesaurus



Concepts

Esophagus  
12

Left Phrenic Nerve

Heart Valves  
9

Fetal Heart  
31

Heart

Mediastinum  
4

Saccular Viscus

Angina Pectoris  
97

Cardiotonic Agents  
225

Tissue Donors  
22

# Other resources

## ◆ Lexical

- WordNet <http://wordnet.princeton.edu/>
- Specialized resources (e.g., for gene names)

## ◆ Terminological

- Gene Ontology <http://geneontology.org/>
- MeSH <http://www.nlm.nih.gov/mesh/>

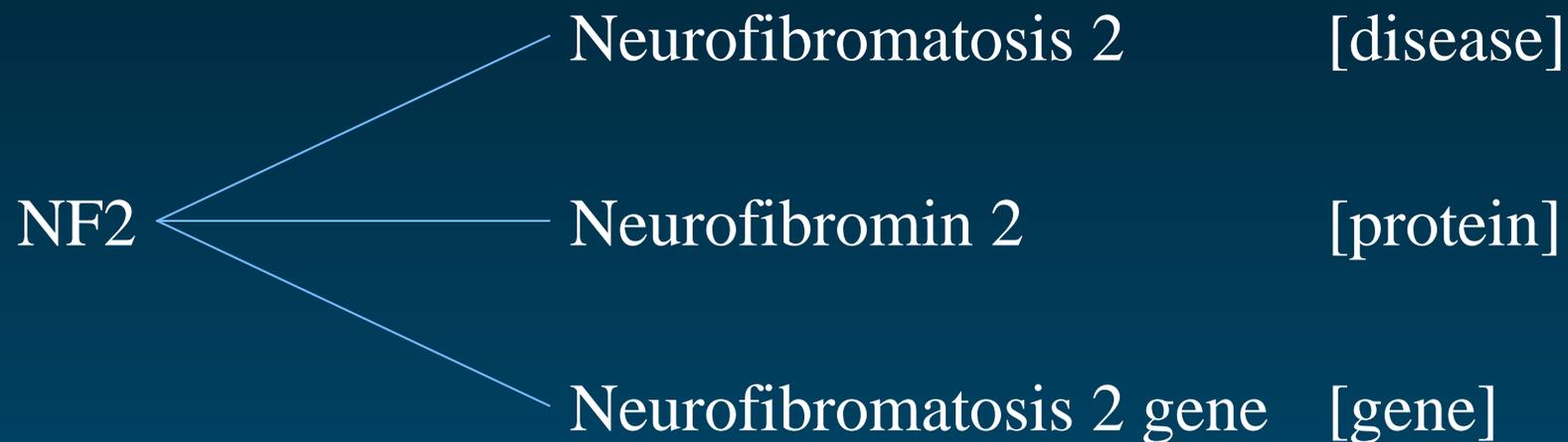
## ◆ Ontological

- SNOMED CT <http://www.snomed.org/>
- FMA <http://fma.biostr.washington.edu/>
- OpenGALEN <http://www.opengalen.org/>



Some issues  
related to these resources

# Ambiguity



# Acronyms and abbreviations

- ◆ Many algorithms
  - For identifying acronyms
  - For extracting the fully specified terms
- ◆ Can be harvested systematically from the biomedical literature and collected in databases
  - *Biomedical Abbreviation Server*  
<http://bionlp.stanford.edu/abbreviation/>
  - *AcroMed*  
<http://medstract.med.tufts.edu/acro1.1/index.htm>
- ◆ Ambiguity issue



# Limited coverage

- ◆ e.g., Gene and protein names
  - Additional sources
  - Additional identification methods

<b>Genew</b>	<a href="http://www.gene.ucl.ac.uk/nomenclature/">http://www.gene.ucl.ac.uk/nomenclature/</a>
<b>Entrez Gene</b>	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene</a>
<b>UniProt</b>	<a href="http://www.ebi.uniprot.org/index.shtml">http://www.ebi.uniprot.org/index.shtml</a>

# Terminological vs. ontological relations

- ◆ Purpose-dependent relations in terminologies
  - *Addison's disease* **isa** *Autoimmune disorder*
  - *Accidents* hierarchy in MeSH
- ◆ Relations used to create hierarchies  
vs. hierarchical relations



# Conclusions

# Conclusions

- ◆ Lexical and terminological resources enable entity recognition
- ◆ Terminological and ontological resources enable relation extraction

But...

- ◆ Text mining techniques can also benefit
  - Terminologies: term extraction
  - Ontologies: ontology population



# References

- ◆ Bodenreider O.

*Lexical, terminological and ontological resources for biological text mining.*

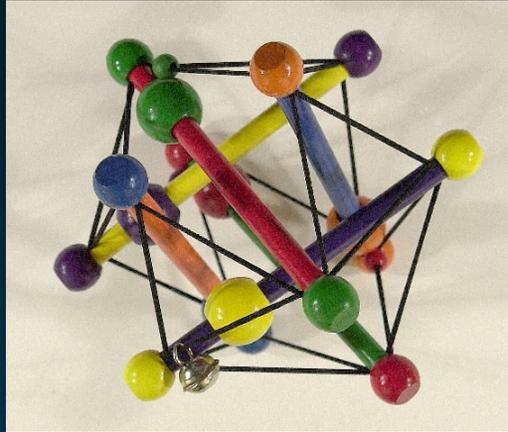
In: Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine: Artech House; 2006. p. 43-66.



# UMLS documentation and support

- ◆ UMLS homepage <http://umlsinfo.nlm.nih.gov/>
  - with links to all other UMLS information
- ◆ UMLSKS homepage <http://umlsks.nlm.nih.gov/>
  - with links to the User's and Developer's guides
- ◆ Email address for support [custserv@nlm.nih.gov](mailto:custserv@nlm.nih.gov)





# 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