

Training Course

Schloß Dagstuhl

in

Biomedical Ontology

May 24, 2006

Ontologies for Data Integration: A Semantic Web Perspective



Olivier Bodenreider

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

Outline

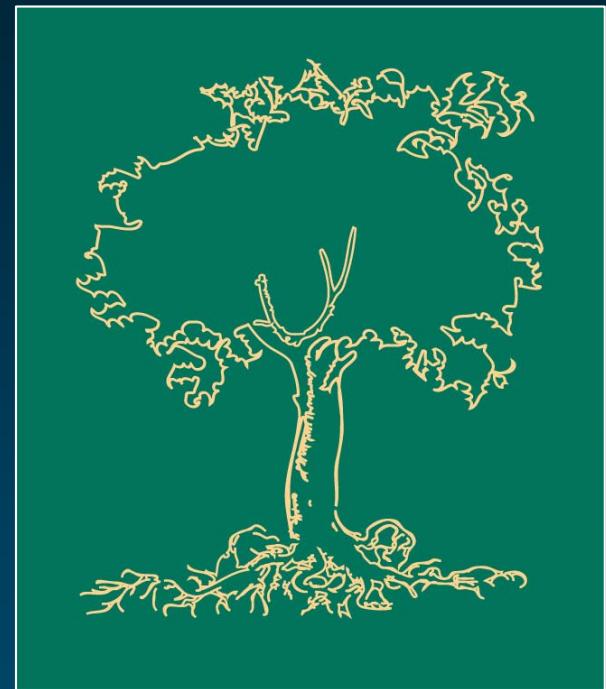
- ◆ From terminology integration
to information integration
Unified Medical Language System (UMLS)
- ◆ UMLS in use:
Mapping across terminologies
- ◆ Beyond UMLS:
Towards a *Biomedical Knowledge Repository*



From terminology integration
to information integration
Unified Medical Language System (UMLS)

What does UMLS stand for?

- ◆ Unified
- ◆ Medical
- ◆ Language
- ◆ System



UMLS®
Unified Medical Language System®
UMLS Metathesaurus®



Lister Hill National Center for Biomedical Communications

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.

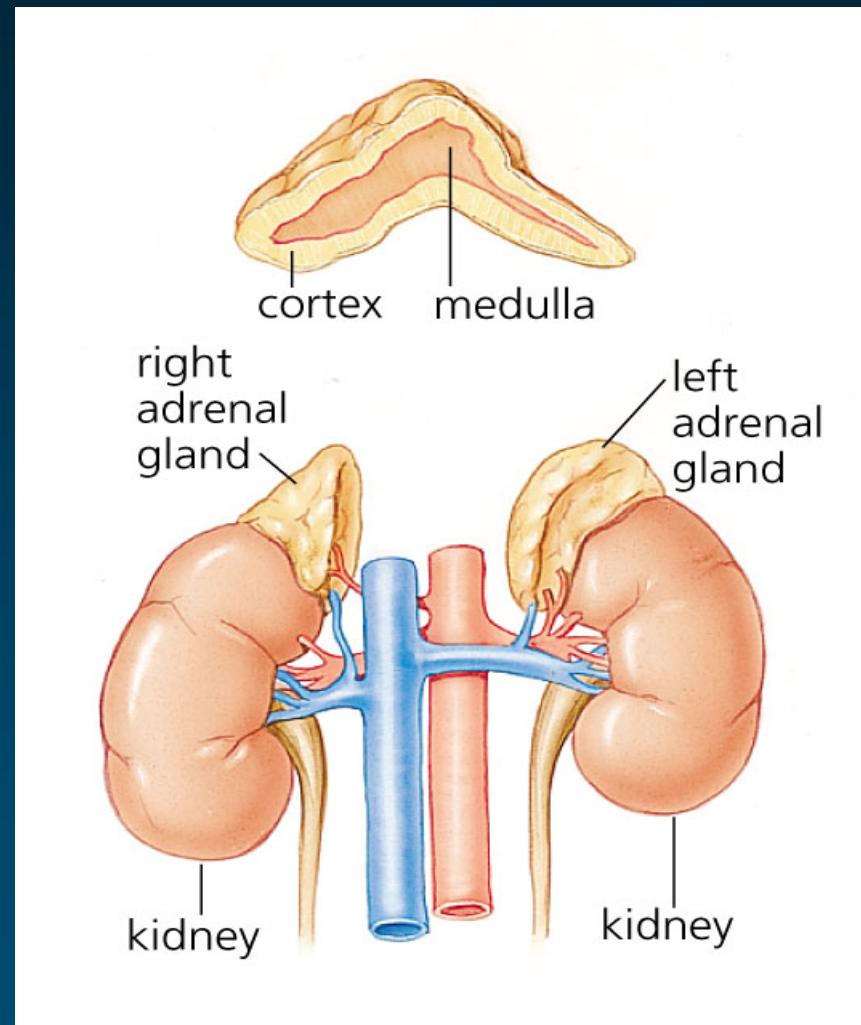
- The first is **the variety of ways the same concepts are expressed** in different machine-readable sources and by different people.
- The second is the **distribution** of useful information among many disparate databases and systems.»



Overview through an example

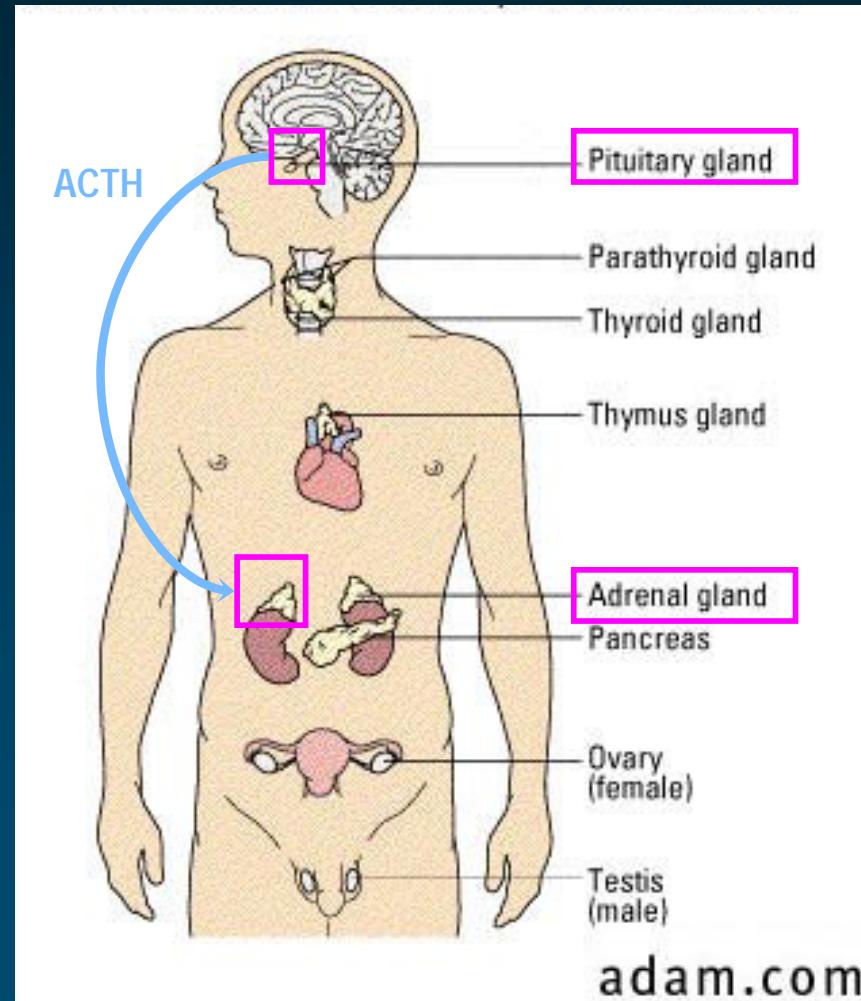
Addison's disease

- ◆ Addison's disease is a rare endocrine disorder
- ◆ Addison's disease occurs when the adrenal glands do not produce enough of the hormone cortisol
- ◆ For this reason, the disease is sometimes called chronic adrenal insufficiency, or hypocortisolism



Adrenal insufficiency Clinical variants

- ◆ Primary / Secondary
 - Primary: lesion of the adrenal glands themselves
 - Secondary: inadequate secretion of ACTH by the pituitary gland
- ◆ Acute / Chronic
- ◆ Isolated / Polyendocrine deficiency syndrome



adam.com

Addison's disease: Symptoms

- ◆ Fatigue
- ◆ Weakness
- ◆ Low blood pressure
- ◆ Pigmentation of the skin (exposed and non-exposed parts of the body)
- ◆ ...



AD in medical vocabularies

◆ Synonyms: different terms

- Addisonian syndrome } eponym
- Bronzed disease }
- Addison melanoderma } symptoms
- Asthenia pigmentosa }
- Primary adrenal deficiency } clinical
- Primary adrenal insufficiency }
- Primary adrenocortical insufficiency }
- Chronic adrenocortical insufficiency }

◆ Contexts: different hierarchies

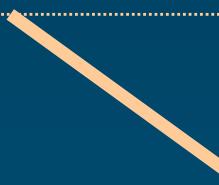


Organize terms

- ◆ Synonymous terms clustered into a concept
- ◆ Preferred term
- ◆ Unique identifier (CUI)

Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

C0001403



Addison's disease



SNOMED International

Diseases/Diagnoses

Diseases of the endocrine system

Diseases of the Adrenal Glands

Addison's Disease



MeSH

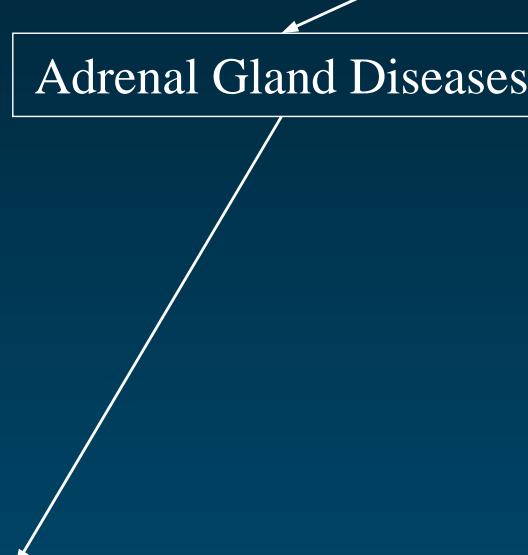
Diseases

Endocrine Diseases

Adrenal Gland Diseases

Adrenal Gland Hypofunction

Addison's Disease



AOD

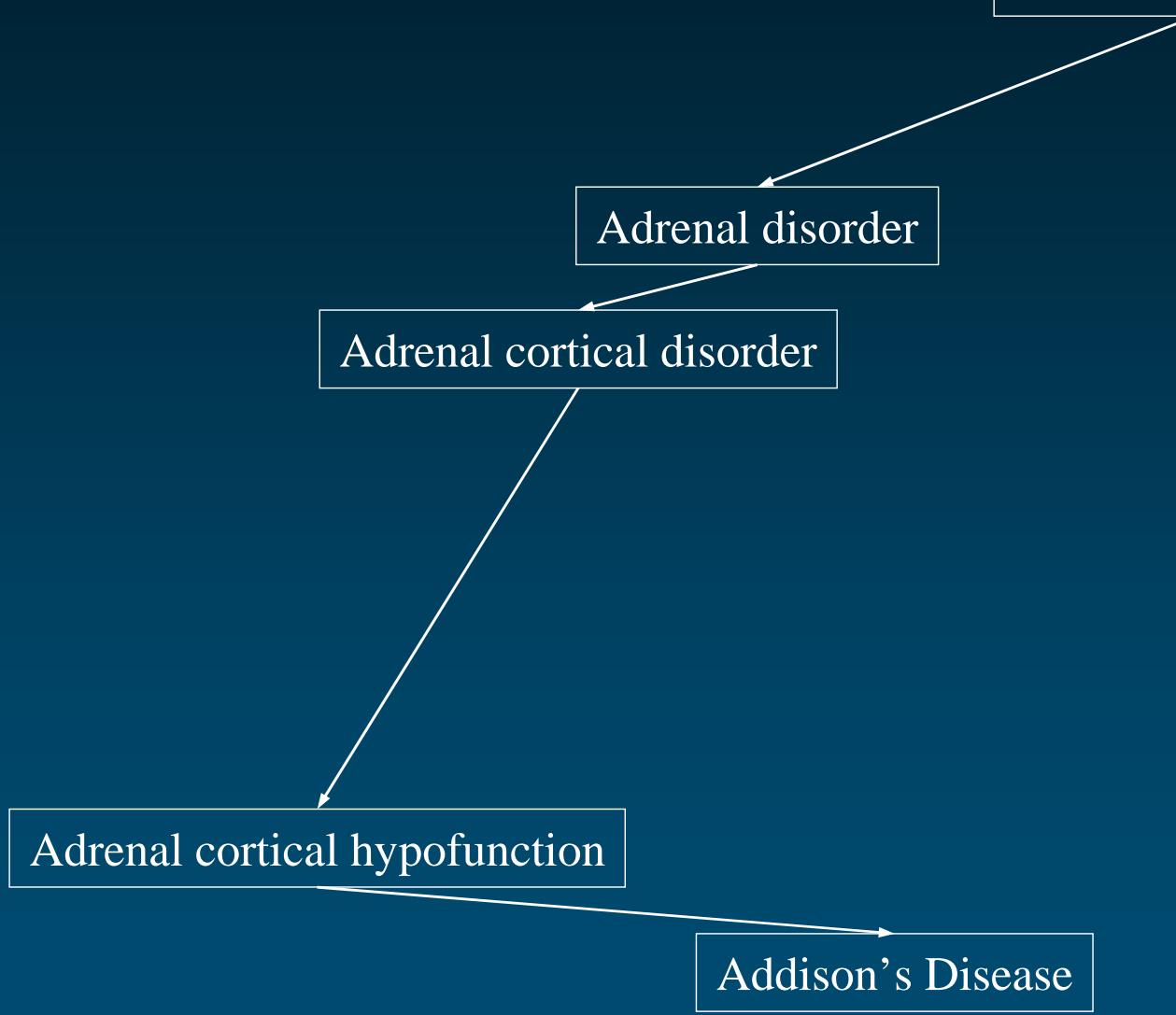
Endocrine disorder

Adrenal disorder

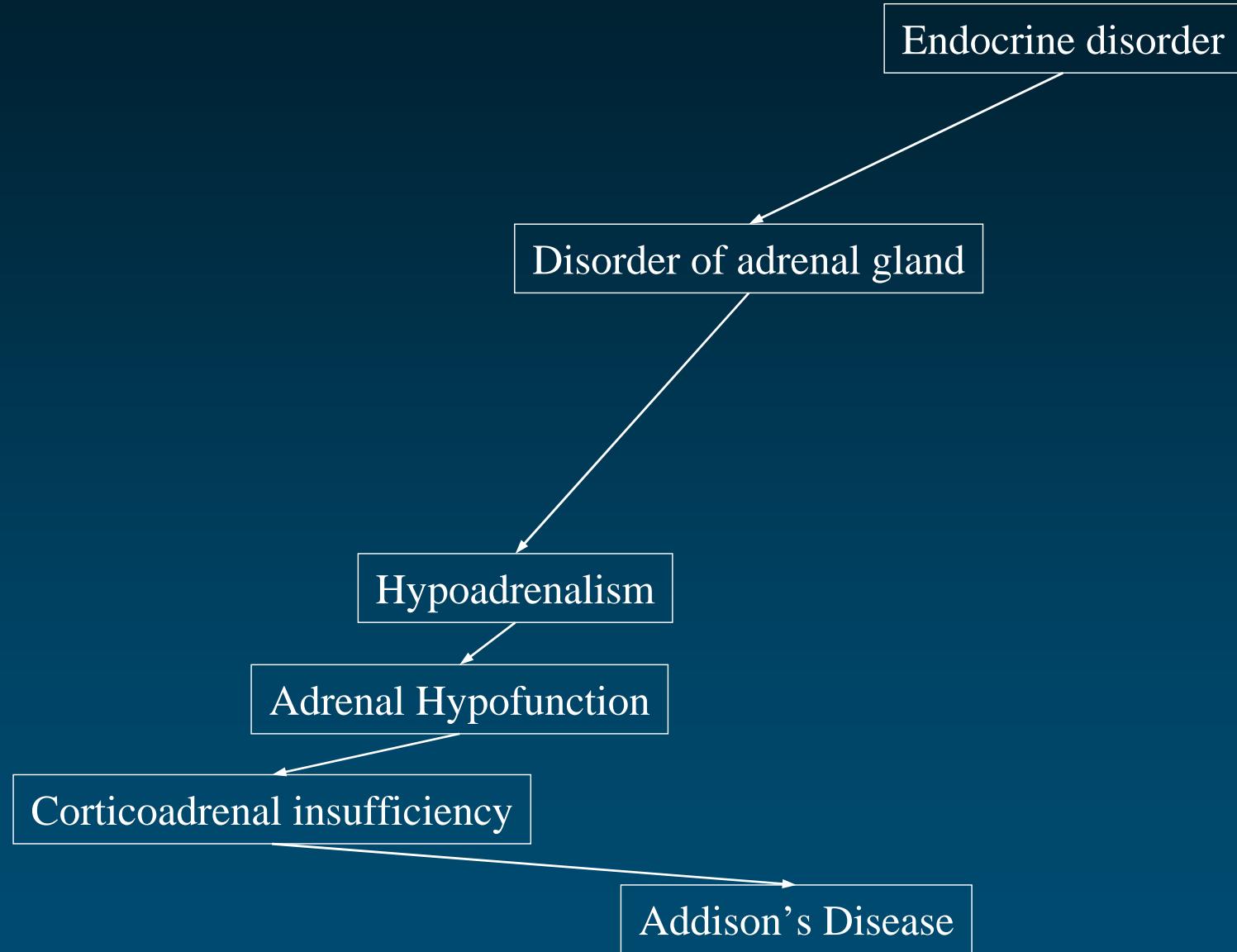
Adrenal cortical disorder

Adrenal cortical hypofunction

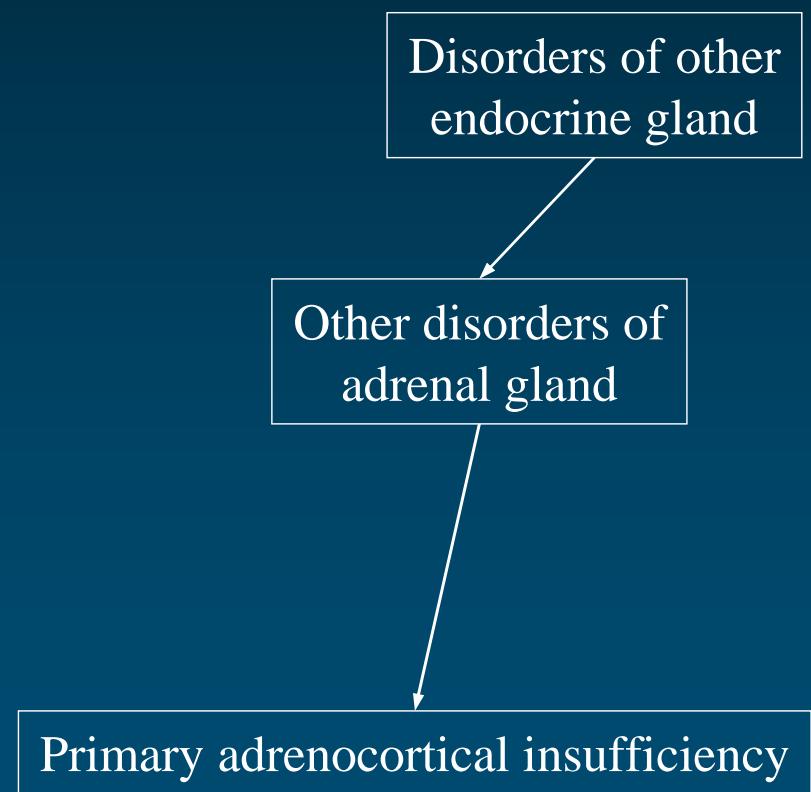
Addison's Disease



Read Codes

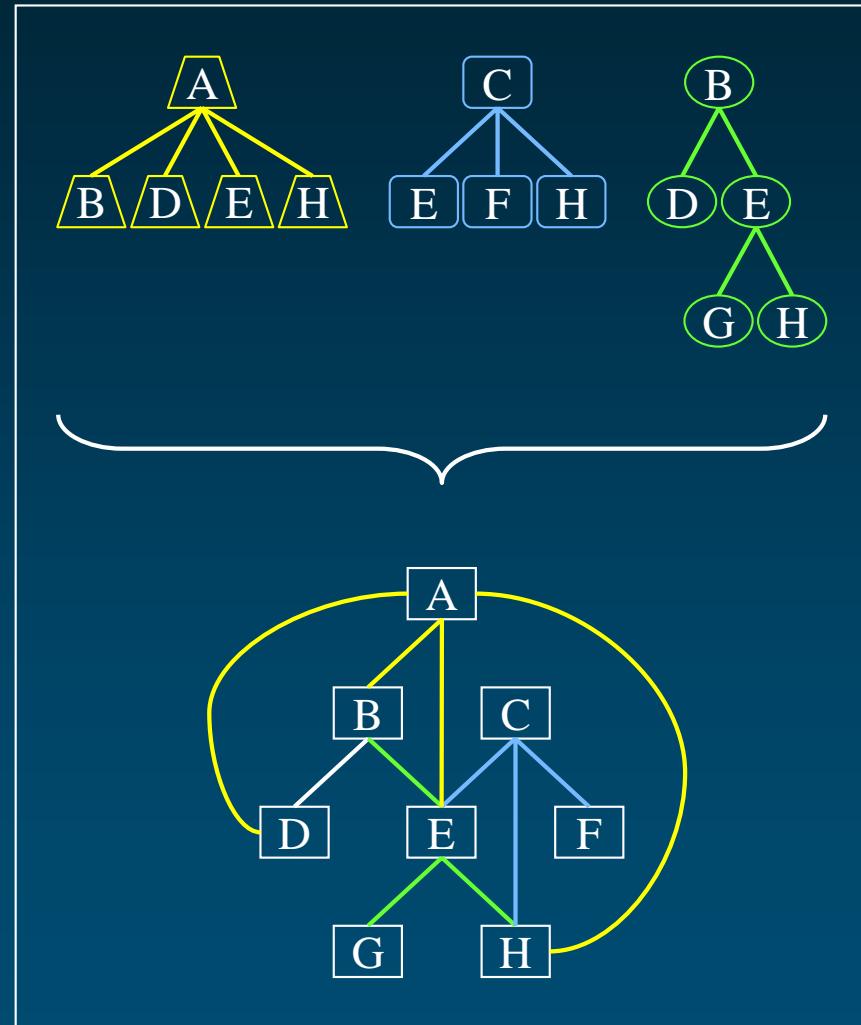


ICD-10



Organize concepts

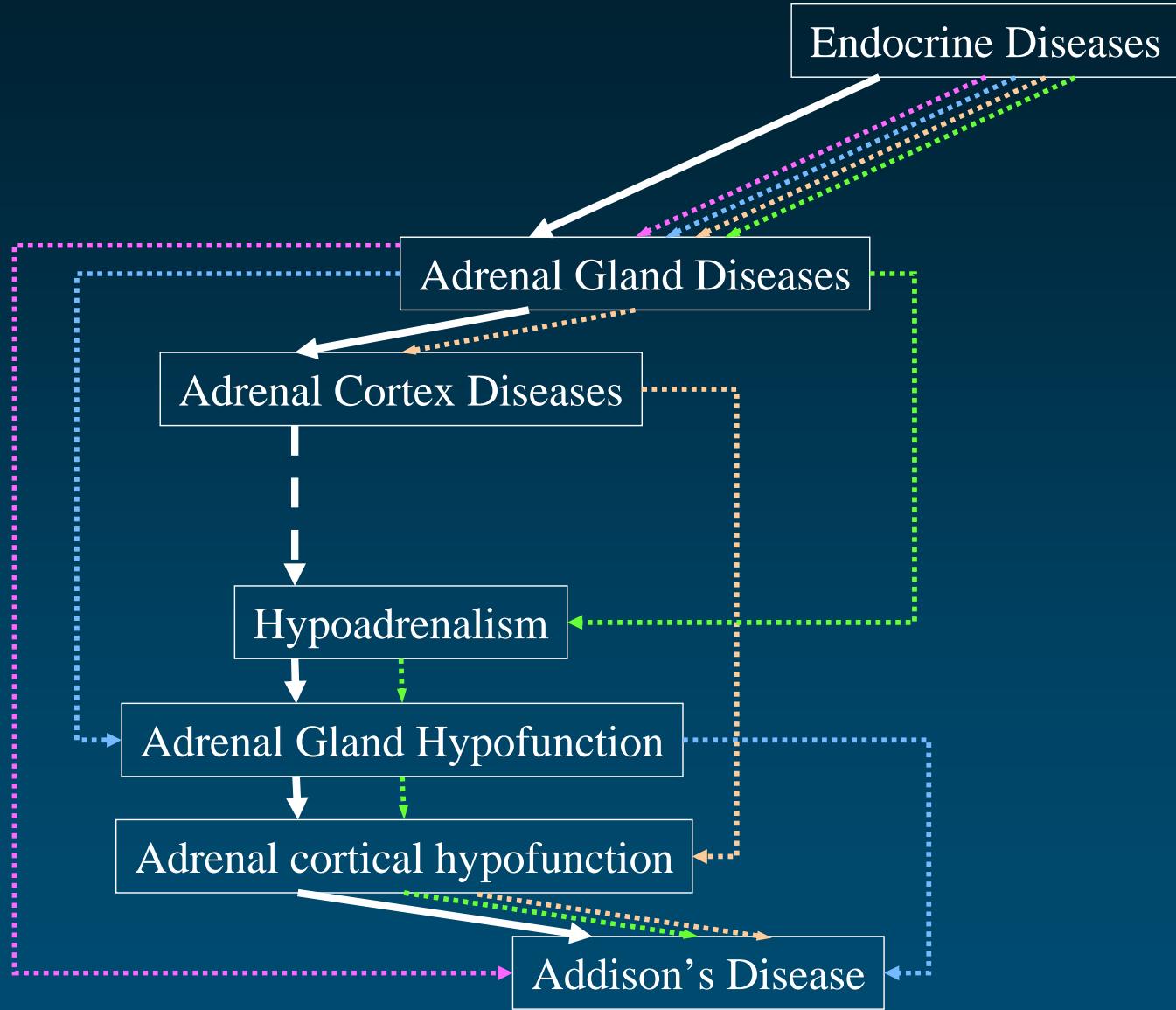
- ◆ Inter-concept relationships: hierarchies from the source vocabularies
- ◆ Redundancy: multiple paths
- ◆ One graph instead of multiple trees (multiple inheritance)



organize concepts

SNOMED
MeSH
AOD
Read Codes

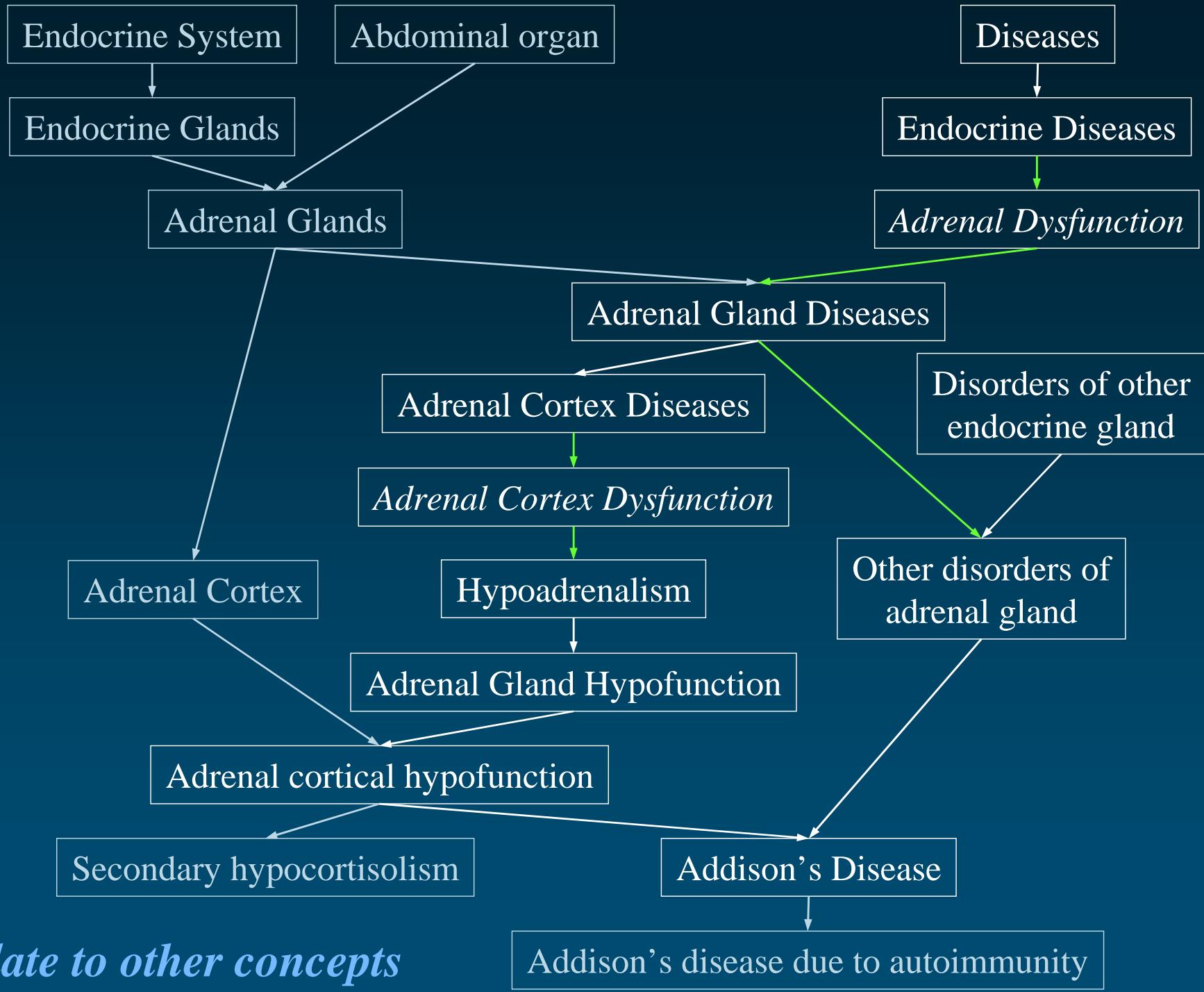
UMLS



Relate to other concepts

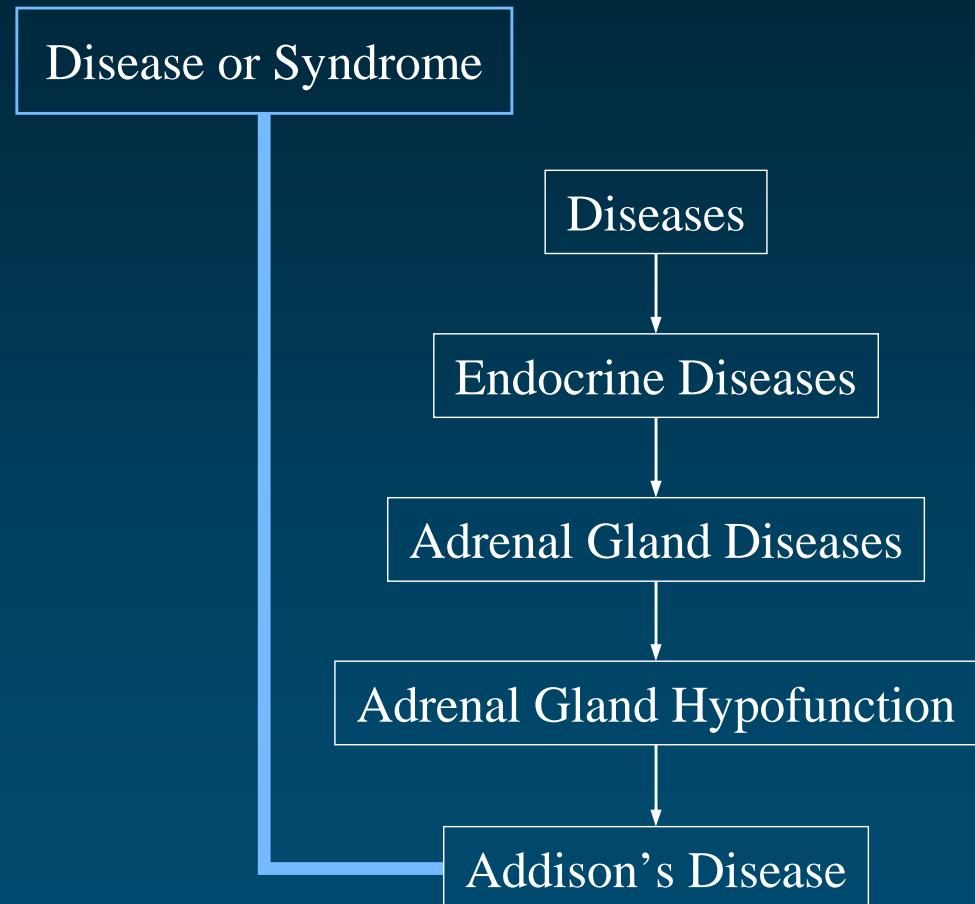
- ◆ Additional hierarchical relationships
 - link to other trees
 - make relationships explicit
- ◆ Non-hierarchical relationships
- ◆ Co-occurring concepts
- ◆ Mapping relationships





Categorize concepts

- ◆ High-level categories (semantic types)
- ◆ Assigned by the Metathesaurus editors
- ◆ Independently of the hierarchies in which these concepts are located



How do they do that?

- ◆ Lexical knowledge
- ◆ Semantic pre-processing
- ◆ UMLS editors



Lexical knowledge

Adrenal gland diseases

Adrenal disorder

Disorder of adrenal gland

Diseases of the adrenal glands

C0001621

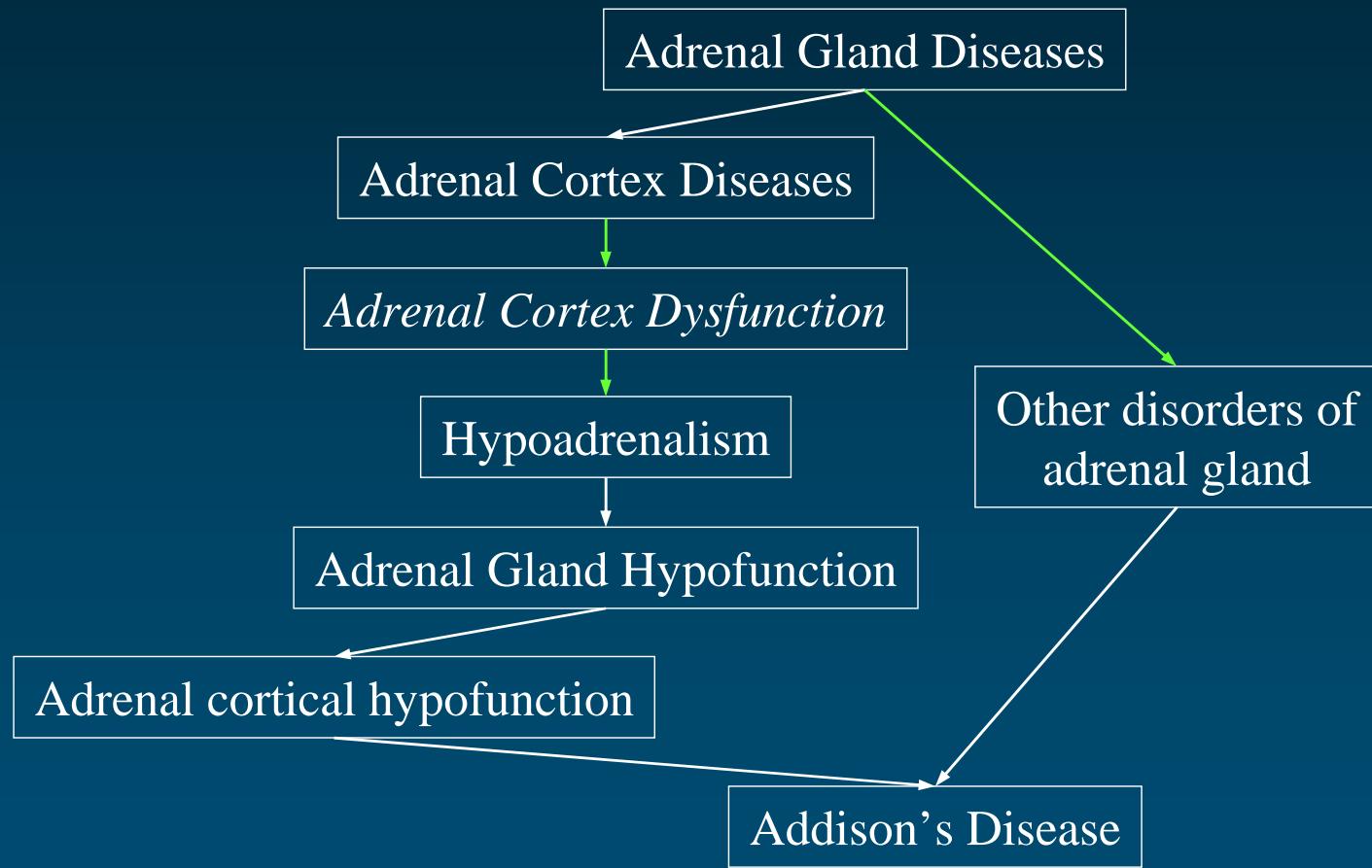


Semantic pre-processing

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



Additional knowledge: UMLS editors



UMLS: 3 components



◆ 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



UMLS Metathesaurus



Source Vocabularies

(2006AB)

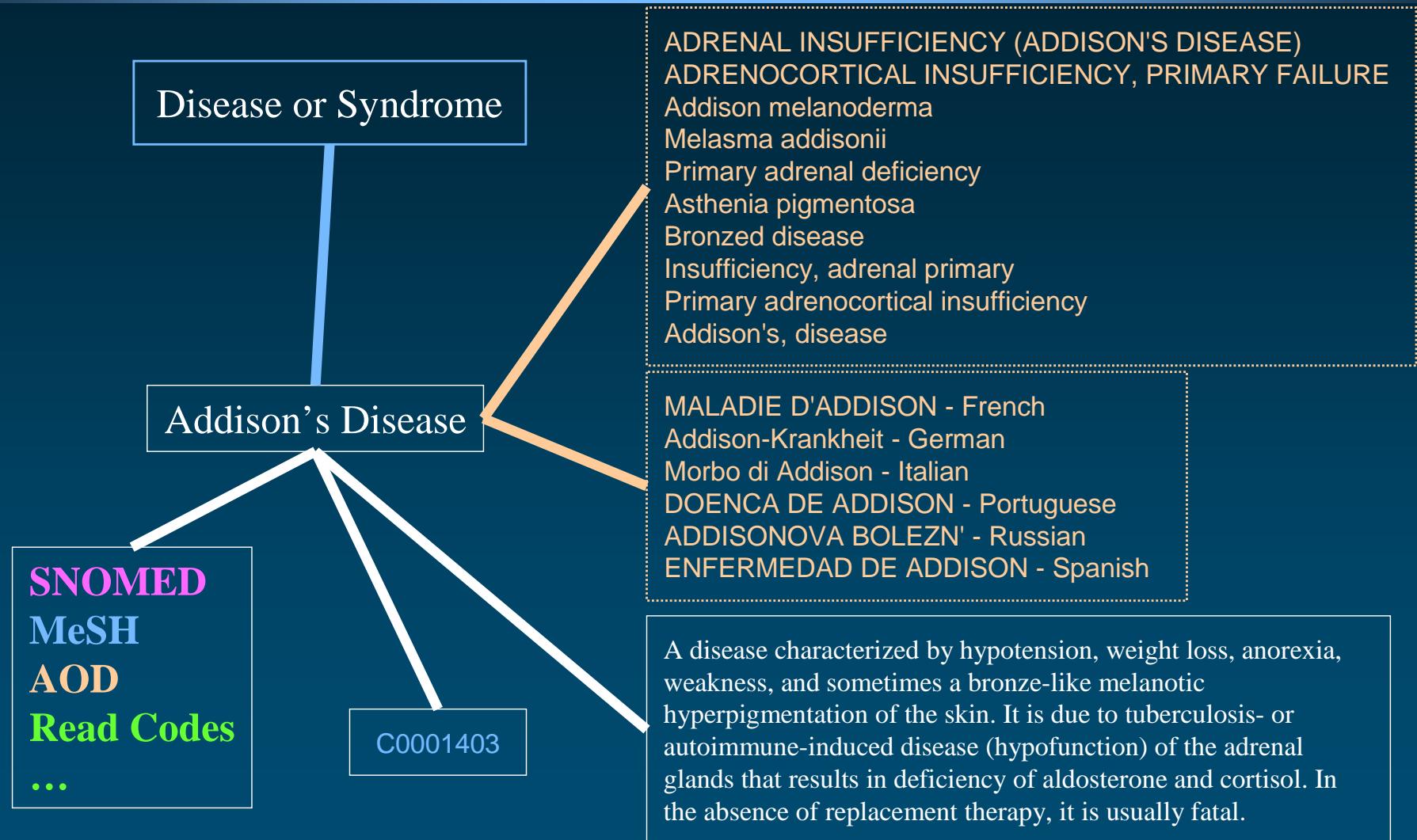
- ◆ 139 source vocabularies
 - 17 languages
- ◆ Broad coverage of biomedicine
 - 5.1M names
 - 1.3M concepts
 - 16M relations
- ◆ Common presentation



Lister Hill National Center for Biomedical Communications

28

Addison's Disease: Concept



Metathesaurus Concepts (2006AB)

- ◆ Concept (> 1.3M) CUI
 - Set of synonymous concept names
- ◆ Term (> 4.6M) LUI
 - Set of normalized names
- ◆ String (> 5.1M) SUI
 - Distinct concept name
- ◆ Atom (> 6.2M) AUI
 - Concept name in a given source

A0000001	headache	(source 1)
A0000002	headache	(source 2)
S0000001		
A0000003	Headache	(source 1)
A0000004	Headache	(source 2)
S0000002		
L0000001		
A0000005	Cephalgia	(source 1)
S0000003		
L0000002		
C0000001		



Cluster of synonymous terms

Concept C0001621	Term L0001621	S0011232 Adrenal Gland Diseases S0011231 Adrenal Gland Disease S0000441 Disease of adrenal gland S0481705 Disease of adrenal gland, NOS S0220090 Disease, adrenal gland S0044801 Gland Disease, Adrenal	[...]
	Term L0041793	S0860744 <i>Disorder of adrenal gland, unspecified</i> S0217833 Unspecified disorder of adrenal glands	
	Term L0161347	S0225481 ADRENAL DISORDER S0627685 DISORDER ADRENAL (NOS)	[...]
	Term L0181041	S0632950 <i>Disorder of adrenal gland</i> S0354509 Adrenal Gland Disorders	[...]
	Term L0368399	S0586222 <i>Adrenal disease</i> S0466921 ADRENAL DISEASE, NOS	[...]
	Term L1279026	S1520972 <i>Nebennierenkrankheiten</i>	GER
	Term L0162317	S0226798 <i>SURRENALE, MALADIES</i>	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



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



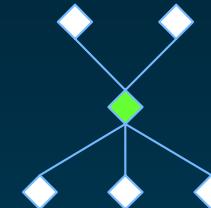
Symbolic relationships Type

◆ Hierarchical

- Parent / Child
- Broader / Narrower than

PAR/CHD

RB/RN



◆ Derived from hierarchies

- Siblings (children of parents)

SIB



◆ Associative

- Other

RO



◆ Various flavors of near-synonymy

- Similar
- Source asserted synonymy
- Possible synonymy

RL

SY

RQ

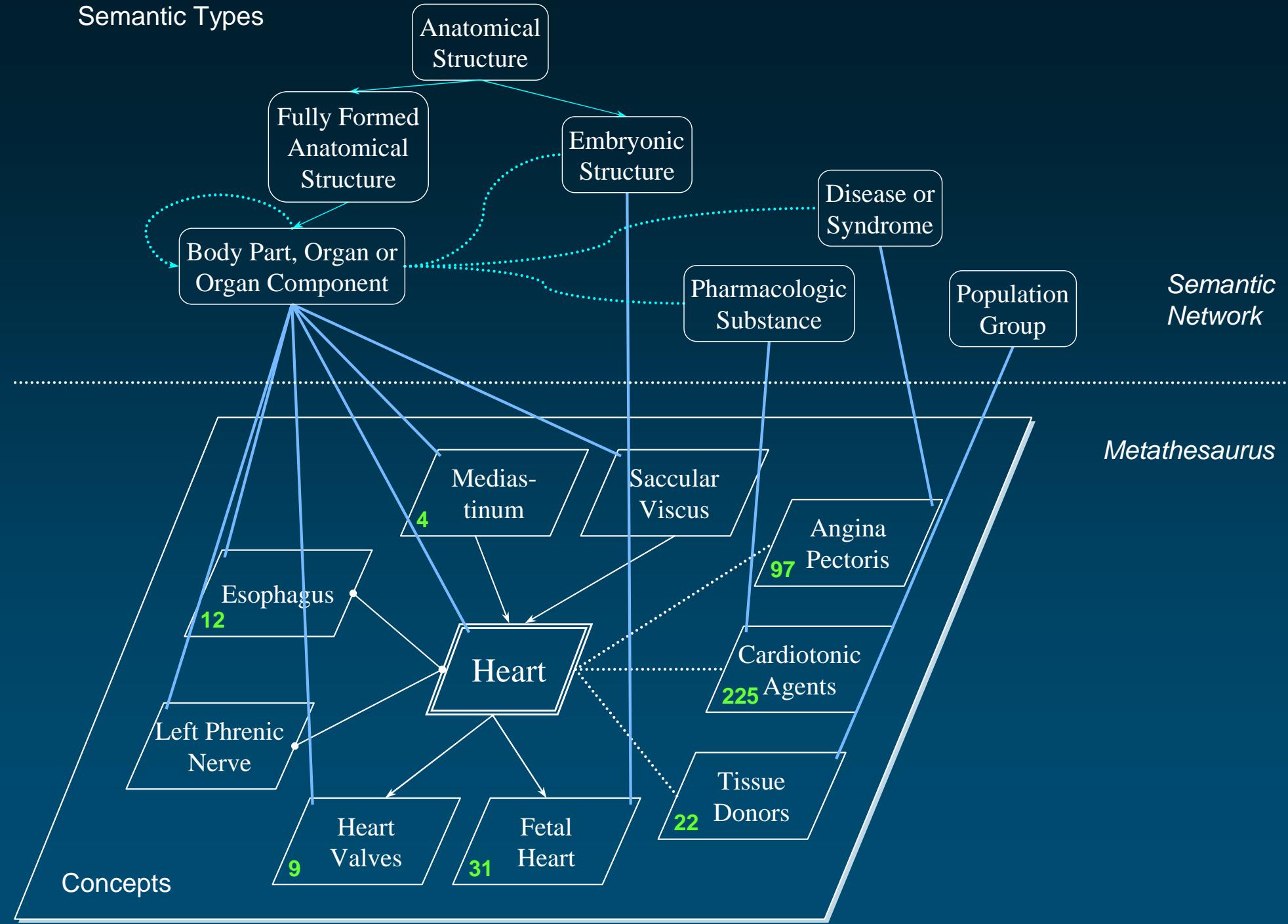


Symbolic relationships Attribute

- ◆ Hierarchical
 - isa (is-a-kind-of)
 - part-of
- ◆ Associative
 - location-of
 - caused-by
 - treats
 - ...
- ◆ Cross-references (mapping)



Semantic Types



UMLS Semantic Network

Semantic Network

◆ Semantic types (135)

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



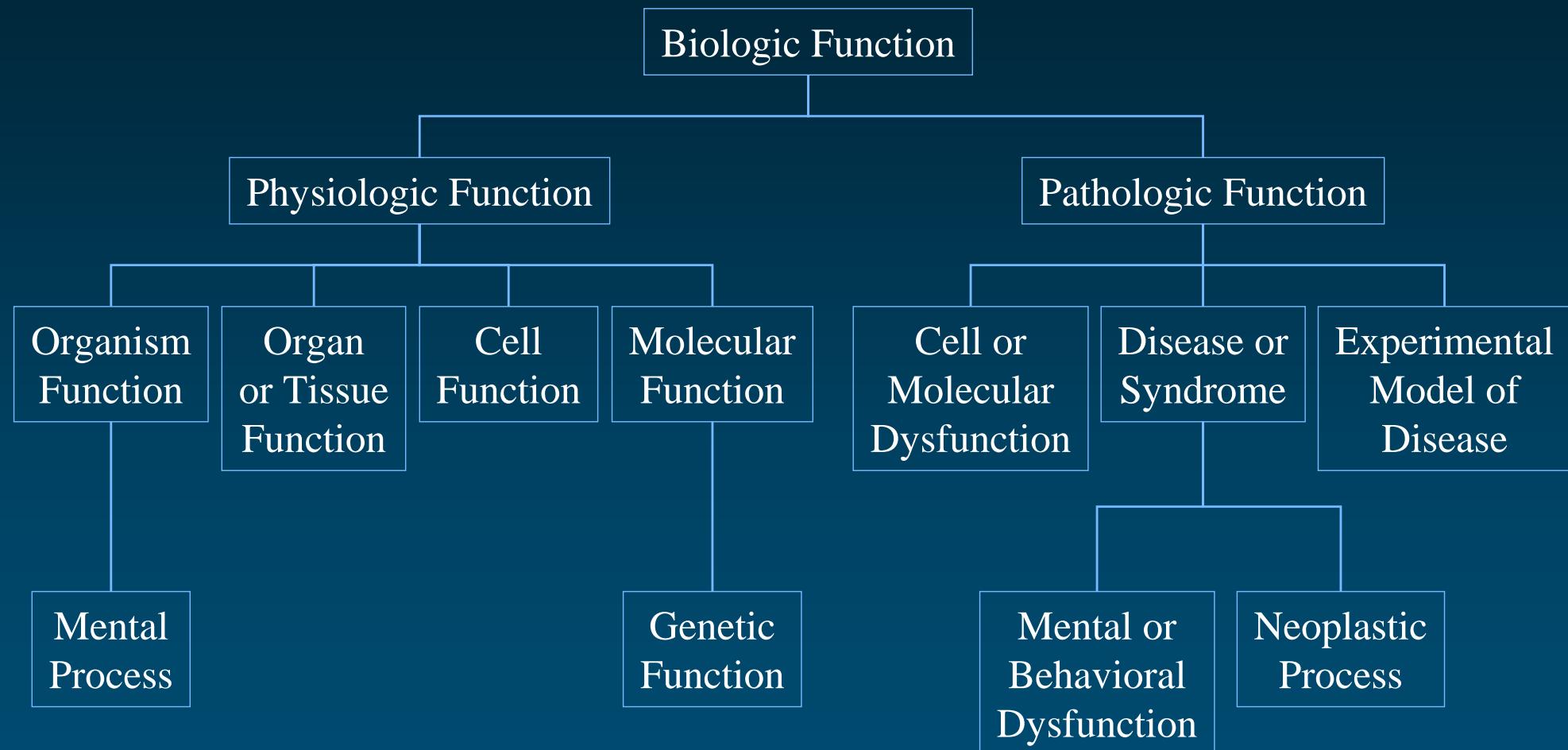
Semantic Network

◆ Semantic network relationships (54)

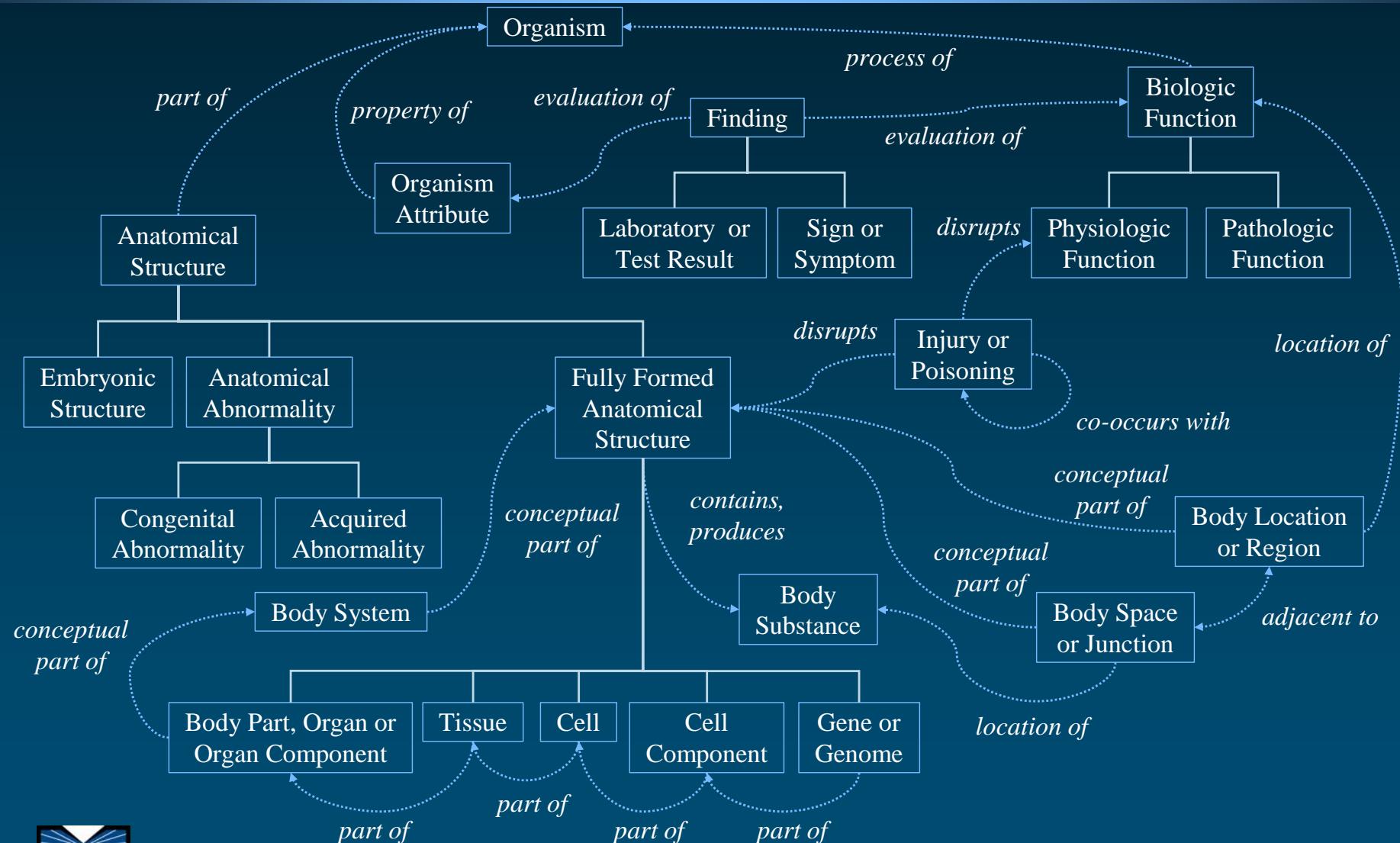
- hierarchical (*isa* = is a kind of)
 - among types
 - Animal *isa* Organism
 - Enzyme *isa* Biologically Active Substance
 - among relations
 - treats *isa* affects
- non-hierarchical
 - Sign or Symptom *diagnoses* Pathologic Function
 - Pharmacologic Substance *treats* Pathologic Function



“Biologic Function” hierarchy (isa)



Associative (non-isa) relationships

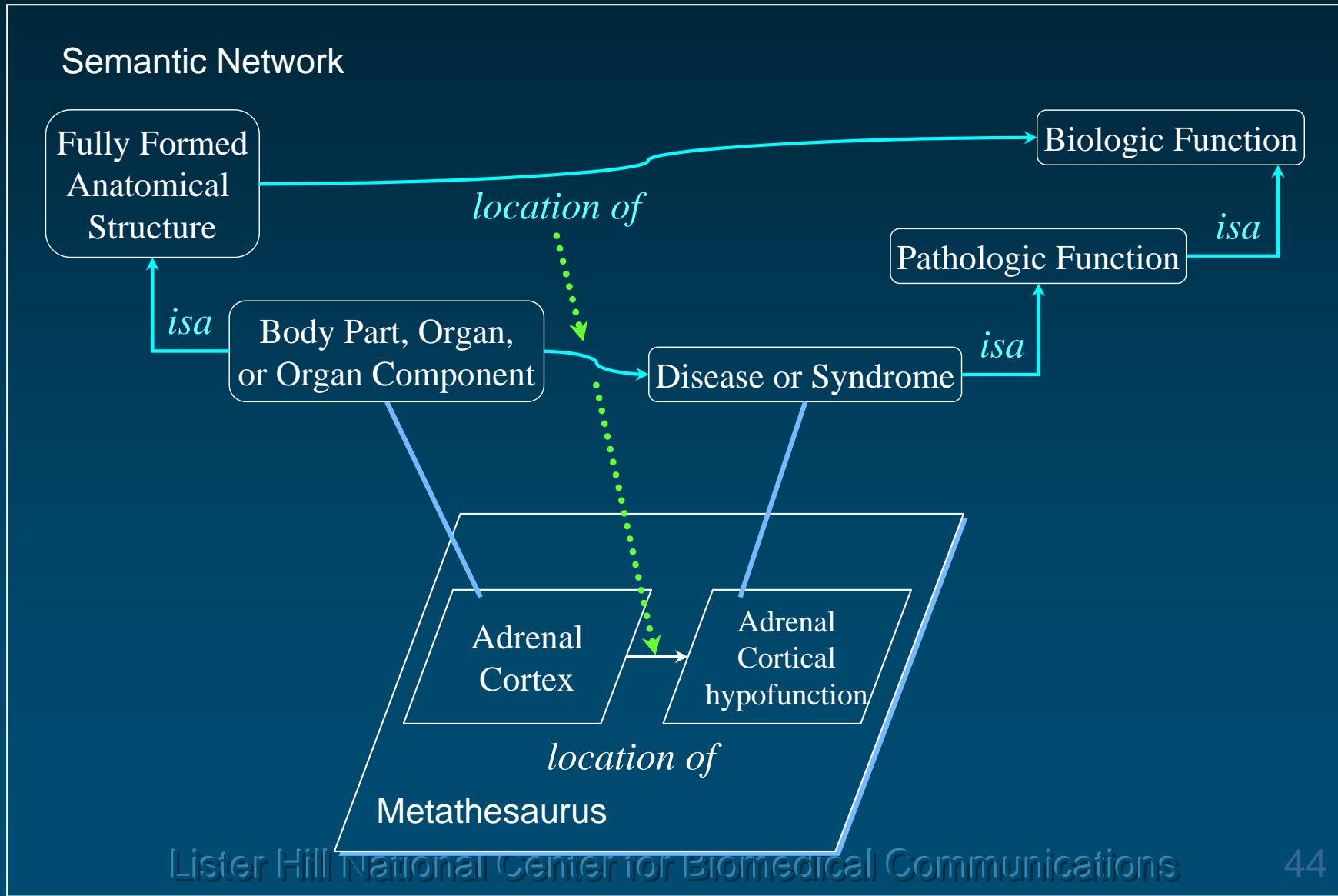


Why a semantic network?

- ◆ 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



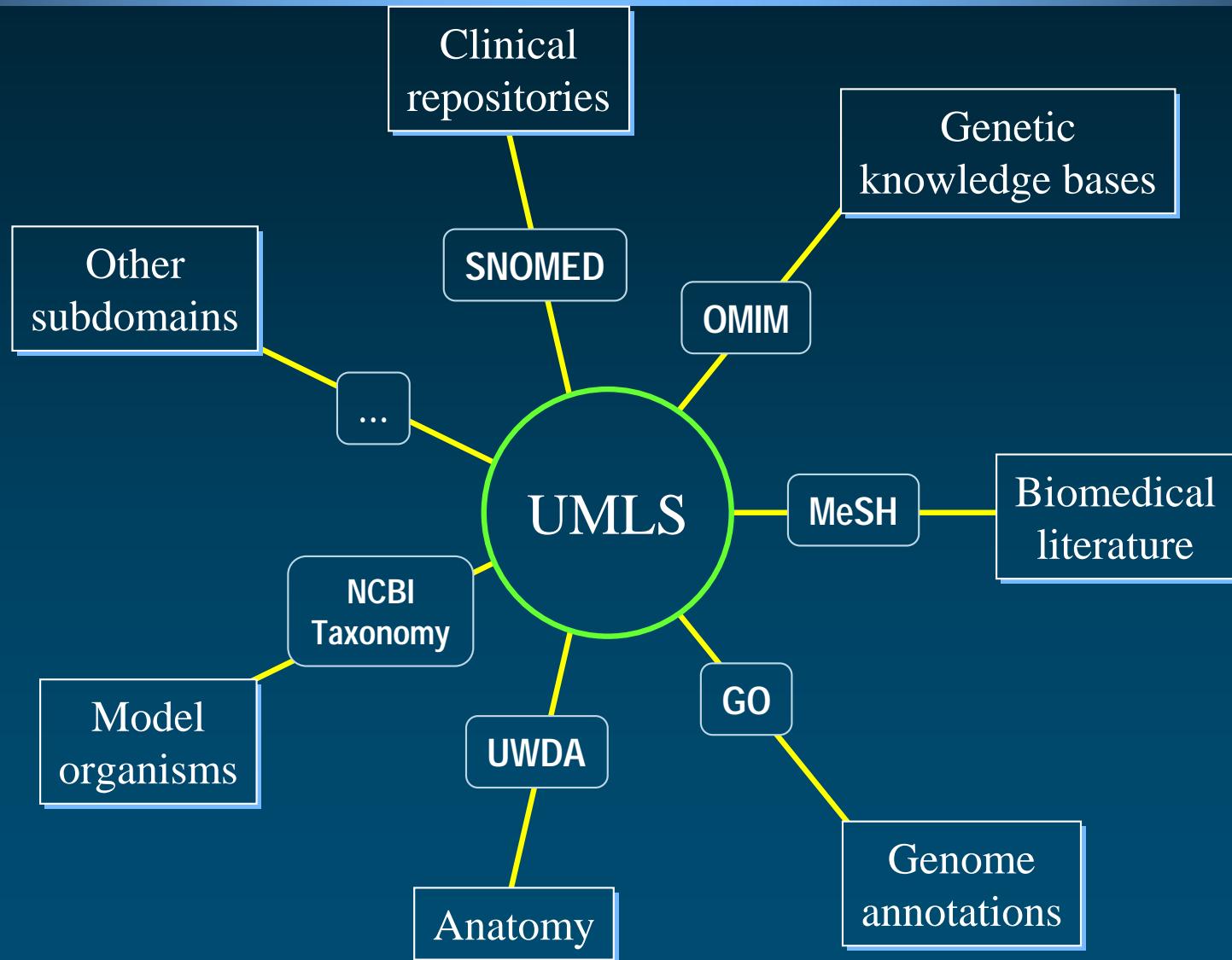
UMLS Summary

- ◆ Synonymous terms clustered into concepts
- ◆ Unique identifier

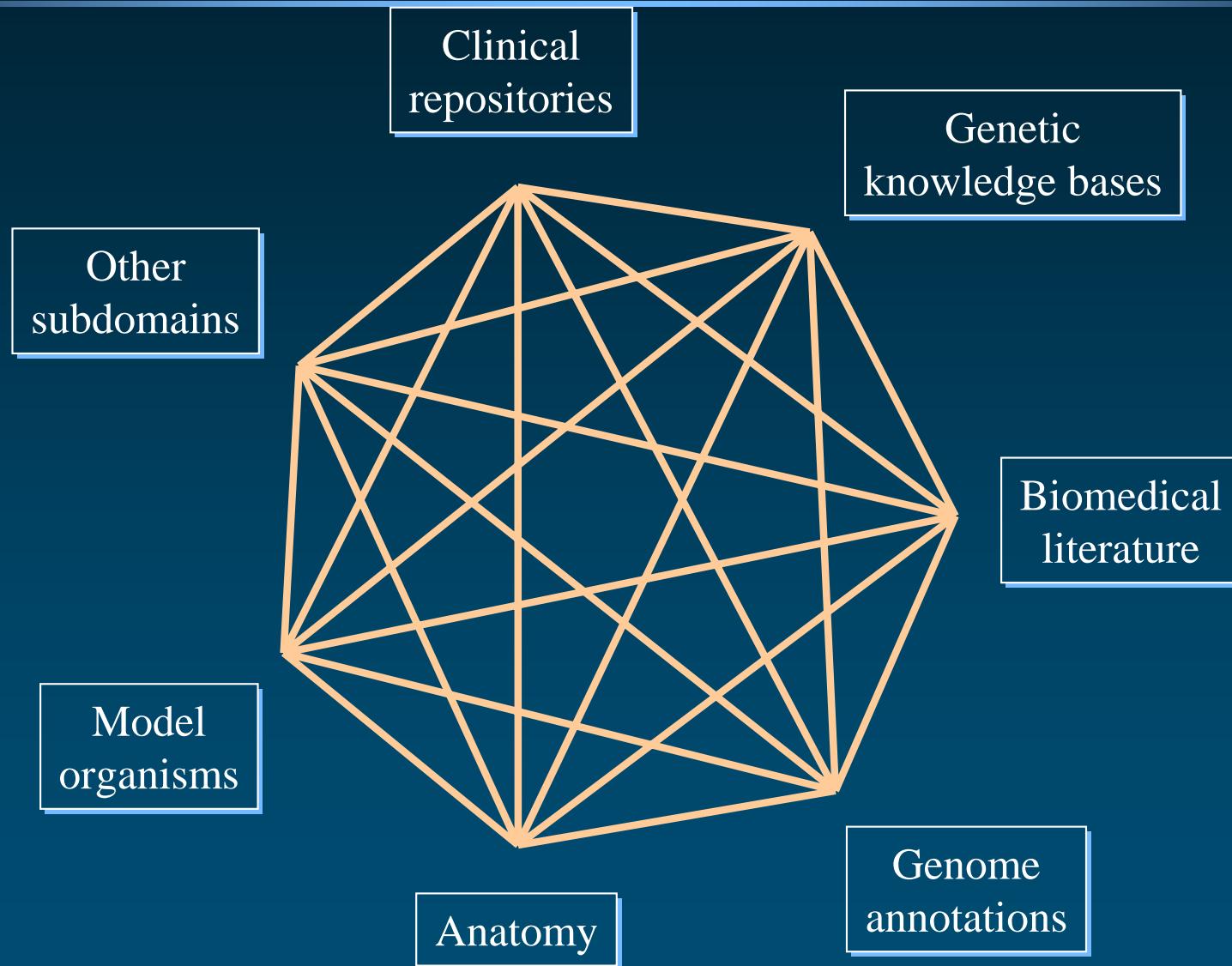
- ◆ Finer granularity
- ◆ Broader scope
- ◆ Additional hierarchical relationships
- ◆ Semantic categorization



Integrating subdomains



Integrating subdomains



Information integration

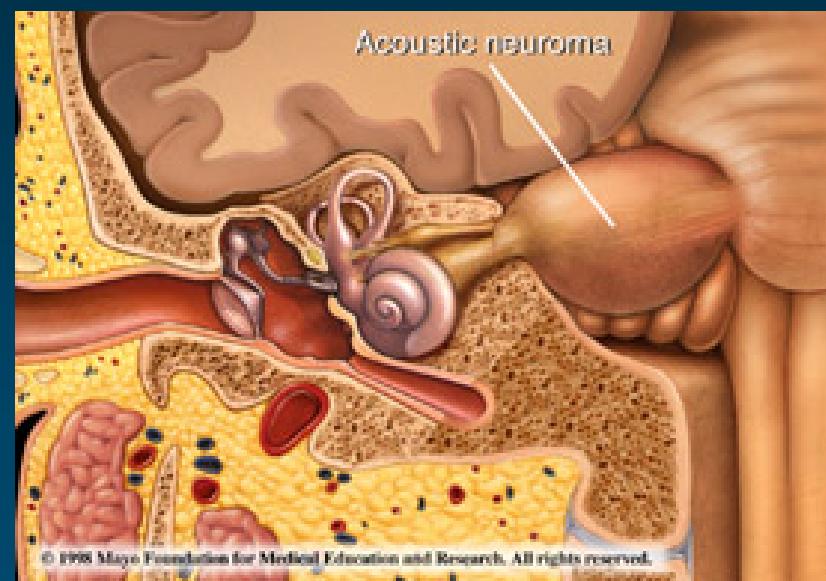
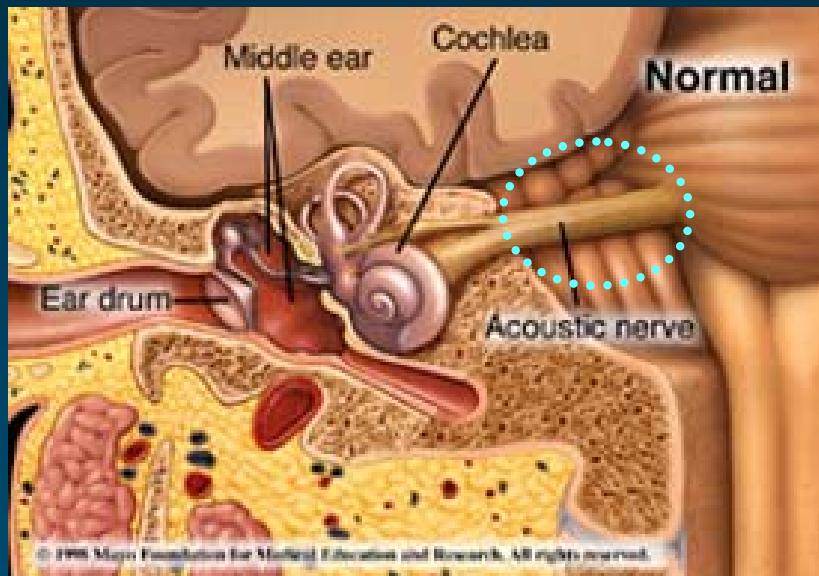
Genomics as an example

NF2 Gene, protein, and disease

Neurofibromatosis 2 is an autosomal dominant disease characterized by tumors called schwannomas involving the acoustic nerve, as well as other features. The disorder is caused by mutations of the *NF2 gene* resulting in absence or inactivation of the protein product. The protein product of NF2 is commonly called *merlin* (but also neurofibromin 2 and schwannomin) and functions as a tumor suppressor.



Schwannoma (acoustic neuroma)



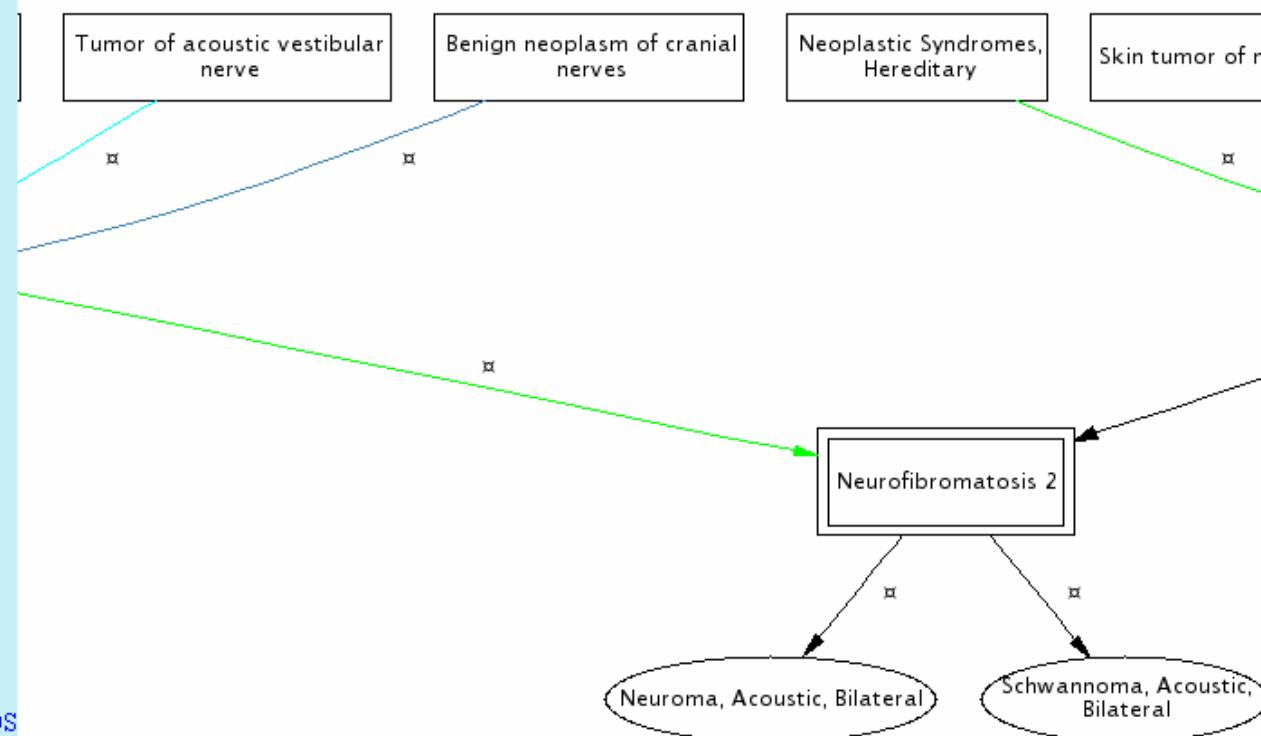
<http://www.mayoclinic.com>

Siblings**Disorders**

- Cerebellopontine Angle Acoustic Neuroma ☞
- Diffuse neurofibroma ☞
- Melanocytic Vestibular Schwannoma ☞
- Neurofibromatosis (nonmalignant) ☞
- Neurofibromatosis 1 ☞
- neurofibromatosis 1 and 2 (NF1 and NF2) ☞
- Neurofibromatosis 3 ☞
- Neurofibromatosis type 3 ☞
- NEUROFIBROMATOSIS TYPE IV, OF RICCARDI ☞
- Neuroma, Acoustic, Unilateral ☞
- Segmental neurofibromatosis ☞

(11 siblings)

[direct children and narrower concepts of direct parents and broader concepts]



BCI

Neurofibromatosis 2

LEGEND *

Similar Concepts
(none)

Allegedly Synonyms

- Neurofibromatosis

Closest MeSH Terms

Main Headings

- Neurofibromatosis 2

Subheadings

Start again **Apply new parameters**

Restrict to vocabulary: Show all

Highlight vocabulary: Nothing

UMLS data: UMLS_2003

Type of hierarchical rel: All Parent/Child only Broader/Narrower only

Other Related Concepts**Anatomy**

- Acoustic Nerve ☞

Chemicals & Drugs

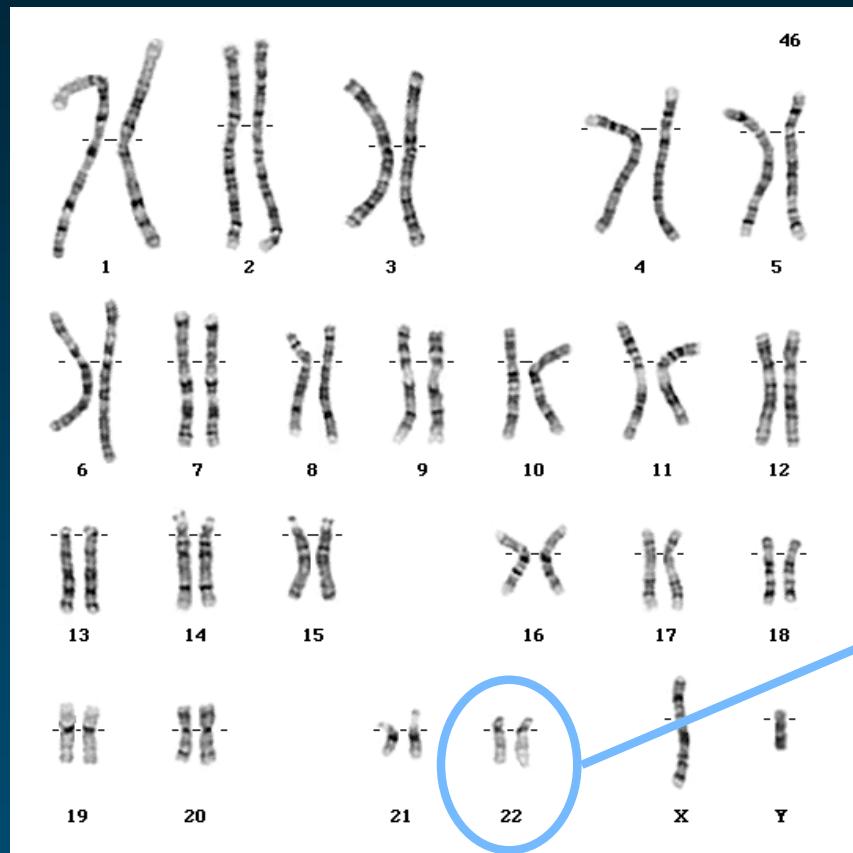
- Neurofibromin 2 ☞

Disorders

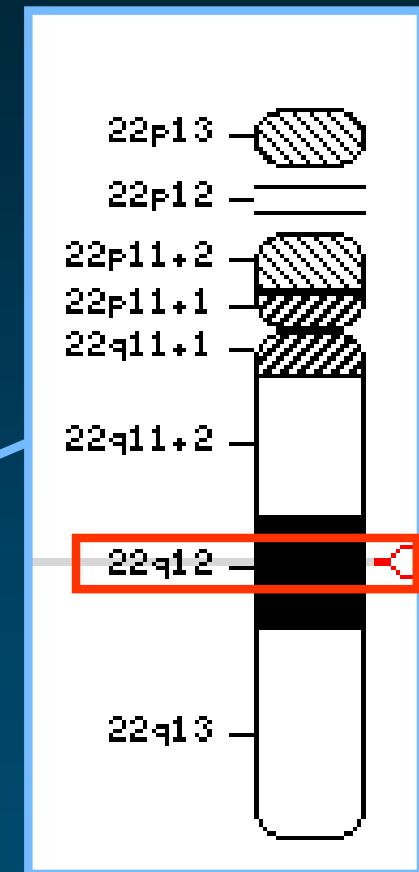
- Familial Acoustic Neuromas ☞
- Neoplasm of uncertain behavior NOS ☞
- Neurofibromatoses ☞
- Neurofibromatosis

- Nerve Sheath Tumors [4] ☞
- Nervous System Neoplasms [6] ☞
- Neurilemmoma [35]
- Neurofibromatosis 1 [38] ☞
- Neuroma, Acoustic [26] ☞
- Peripheral Nervous System Diseases [3] ☞
- Peripheral Nervous System Neoplasms [6] ☞
- Postoperative Complications [9] ☞
- Retinal Diseases [6] ☞
- Skin Neoplasms [9] ☞

NF2 gene



<http://staff.washington.edu/timk/cyto/human/>



<http://www.ncbi.nlm.nih.gov/mapview/>



Lister Hill National Center for Biomedical Communications

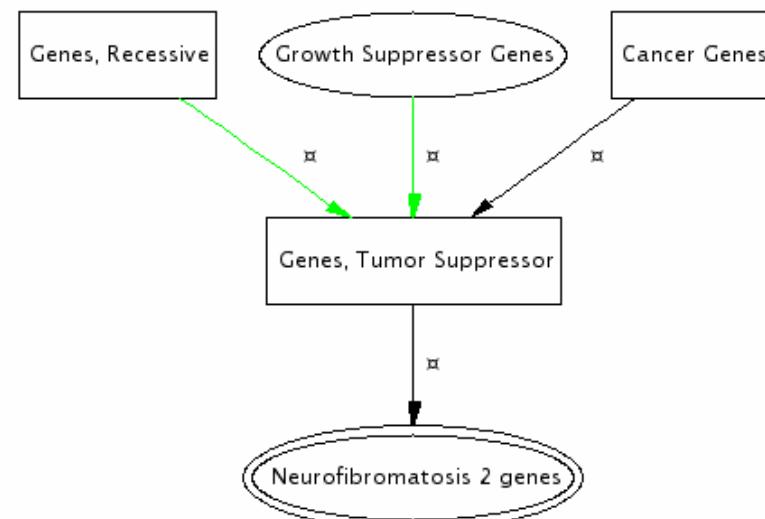
52

Siblings**Chemicals & Drugs**

- ADAM11 protein, human ☈
- DLG5 protein, human ☈
- DPM3 protein, human ☈
- HCCS-1protein, human ☈
- hssh3bp1 protein, human ☈
- HUGL protein, human ☈
- LAPSER1 protein, human ☈
- mitochondria proteolipid-like protein, human ☈
- MRG protein, human ☈
- p53 gene/protein ☈
- PLAGL1 protein, human ☈
- RARRES3 protein, human ☈
- SEZ6L protein, human ☈
- TES protein, human ☈

Genes & Molecular Sequences

- APC Gene ☈
- BAX Gene ☈
- brca gene ☈
- CDH1 gene ☈
- CHES1 Gene ☈
- cyclin-dependent kinase inhibitor 2A



BCI

Neurofibromatosis 2 genes

LEGEND *

Start again

Restrict to vocabulary:

Highlight vocabulary:

UMLS data:

Type of hierarchical rel: All Parent/Child only Broader/Narrower only

Similar Concepts

(none)

Allegedly Synonyms

(none)

Closest MeSH Terms**Main Headings**

- Genes, Neurofibromatosis 2

Subheadings**Other Related Concepts****Chemicals & Drugs**

- Neurofibromin 2 ☈

Disorders

- Neurofibromatosis 2 ☈

(2 other related concepts)

- Chromosome Deletion [7] ☈
- Ependymoma [4] ☈
- Glioma [4] ☈
- Loss of Heterozygosity [7] ☈
- Meningeal Neoplasms [25] ☈
- Meningioma [30] ☈
- mesothelioma <1> [4] ☈
- Neoplasms [4] ☈
- Neurilemmoma [20] ☈
- Neurofibromatoses
- Neurofibromatosis 2 [64] ☈
- Neuroma, Acoustic [5] ☈
- Spinal Cord Neoplasms [3] ☈

Merlin

◆ Synonyms

- Neurofibromin 2
- Schwannomin
- Schwannomerlin
- Neurofibromatosis-2

◆ 10 isoforms

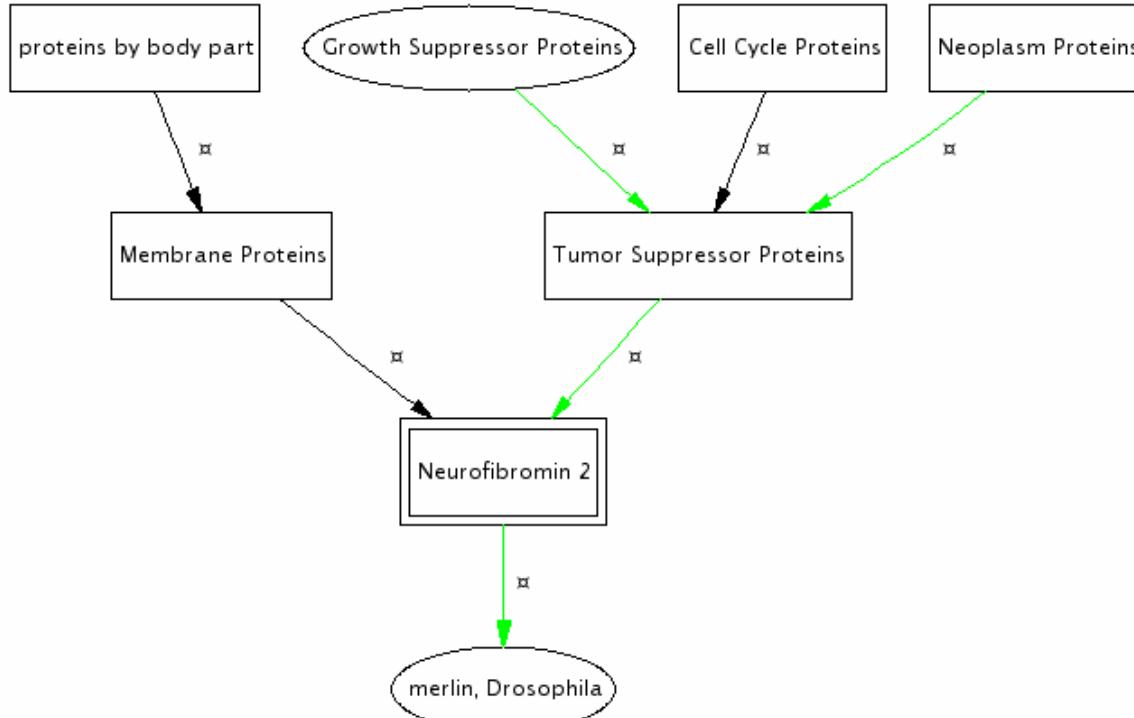
◆ Annotations

- Negative regulation of cell proliferation
- Cytoskeleton
- Plasma membrane



Siblings**Chemicals & Drugs**

- (LA)12 peptide ☐
- (methyl)ammonium uptake carrier, *Corynebacterium* ☐
- 120-kDa hemocyte-specific membrane protein, flesh fly ☐
- 15a protein, *Aedes aegypti* ☐
- 22.6-kDa antigen, *Schistosoma japonicum* ☐
- 36-kDa vesicular integral membrane protein ☐
- 38L protein ☐
- 5-lipoxygenase-activating protein ☐
- 59 kDa dystrophin-associated protein ☐
- A-1 antigen ☐
- A-kinase anchor protein 149 ☐
- A-kinase anchor protein 15 ☐
- A-kinase anchor protein 200 ☐
- A-kinase anchor protein KL ☐
- A14.5L protein ☐
- A15 protein ☐
- ABC-me protein ☐
- ABU-1 protein, *C elegans* ☐
- AcfB protein ☐
- ACR3 protein ☐

**Other Related Concepts****Disorders**

- Neurofibromatosis 2 ☐

Genes & Molecular Sequences

- Neurofibromatosis 2 genes ☐

(2 other related concepts)

Co-occurring Concepts**Anatomy**

- Arachnoid [1] ☐
- Cell Membrane [1] ☐
- Cerebellum [1] ☐
- Chromosomes, Human, Pair 22 [1] ☐
- Cytoplasm [1] ☐
- Cytoskeleton [2] ☐
- Microfilaments [1] ☐
- Purkinje Cells [1] ☐
- Schwann Cells [1] ☐
- Stem Cells [1] ☐

BCI		Neurofibromin 2		LEGEND *	
<input type="button" value="Start again"/>	<input type="button" value="Apply new parameters"/>			<input type="button" value="Similar Concepts"/>	<input type="button" value="Closest MeSH Terms"/>
Restrict to vocabulary:	Show all			(none)	Main Headings
Highlight vocabulary:	Nothing			Allegedly Synonyms	Subheadings
UMLS data:	UMLS_2003			(none)	
Type of hierarchical rel:	All	<input checked="" type="radio"/> Parent/Child only	<input type="radio"/> Broader/Narrower only		
<input type="button" value="T"/> <input type="button" value="C"/> <input type="button" value="G"/> <input type="button" value="S"/> <input type="button" value="E"/> <input type="button" value="A"/>					

UMLS Semantic Network (Semantic Types)

Amino Acid,
Peptide, or Protein

Neoplastic Process

Gene or Genome

Biologically Active Substance

Neurofibromatoses

Benign neoplasms
of cranial nerves

Tumor suppressor genes

Tumor suppressor proteins

Neurofibromatosis 2
(Type II neurofibromatosis,
Bilateral acoustic neurofibromatosis)
C0027832

NF2
(Neurofibromin 2 gene)
C0085114

Merlin
(Schwannomin,
Neurofibromin 2)
C0254123

UMLS Metathesaurus
(Concepts and relations)

Merlin, Drosophila

NEUROFIBROMATOSIS,
TYPE II; NF2
#101000
OMIM

Drosophila melanogaster merlin
(Dmerlin) mRNA, complete cds.
U49724
Genbank

External resources

Limitations

- ◆ Genes not systematically represented
 - Most gene products and diseases are
- ◆ Gene/Gene product-Disease relations
 - Not systematically represented
 - Not explicitly represented (e.g., co-occurrence)
- ◆ Cross-references not systematically represented
- ◆ Naming conventions (genes)



References

- ◆ UMLS
umlsinfo.nlm.nih.gov
- ◆ UMLS browsers
(free, but UMLS license required)
 - Knowledge Source Server: umlsks.nlm.nih.gov
 - Semantic Navigator:
<http://mor.nlm.nih.gov/perl/sempnav.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.
- Nelson, S. J., Powell, T. & Humphreys, B. L. (2002). The Unified Medical Language System (UMLS) Project. In: Kent, Allen; Hall, Carolyn M., editors. *Encyclopedia of Library and Information Science*. New York: Marcel Dekker. p.369-378.



References

- ◆ UMLS as a research project
 - Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, 32(4), 281-91.
 - Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc*, 5(1), 1-11.



References

◆ Technical papers

- McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods Inf Med*, 34(1-2), 193-201.
- Bodenreider O. & McCray A. T. (2003). Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36(6), 414-432.

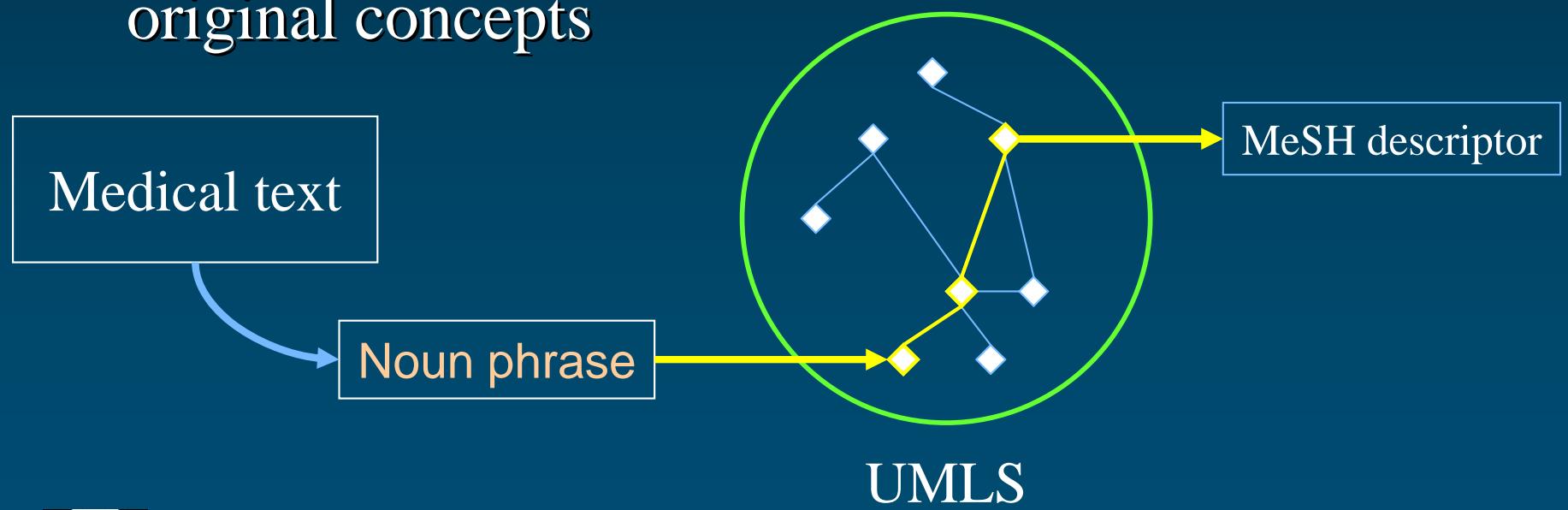


UMLS in Use

Mapping across Vocabularies

The problem

- ◆ For noun phrases extracted from medical texts, map to UMLS concepts
- ◆ Then, select from the MeSH vocabulary the concepts that are the most closely related to the original concepts



Map noun phrases to UMLS

◆ Normalization

- normalize noun phrases
- use the normalized string index

◆ MetaMap

- approximate matching
- more aggressive approach
 - use derivational variants
 - allow partial matches



Restrict to MeSH

- ◆ Based on the principle of semantic locality
- ◆ Use different components of the UMLS
- ◆ 4 techniques of increasing aggressiveness
 - Use Synonymy MRCON + MRSO
 - Use Associated expressions (ATXs) MRATX
 - Explore the Ancestors MRREL + SN
 - Explore the Other related concepts MRREL + SN



Restrict to MeSH: Synonymy

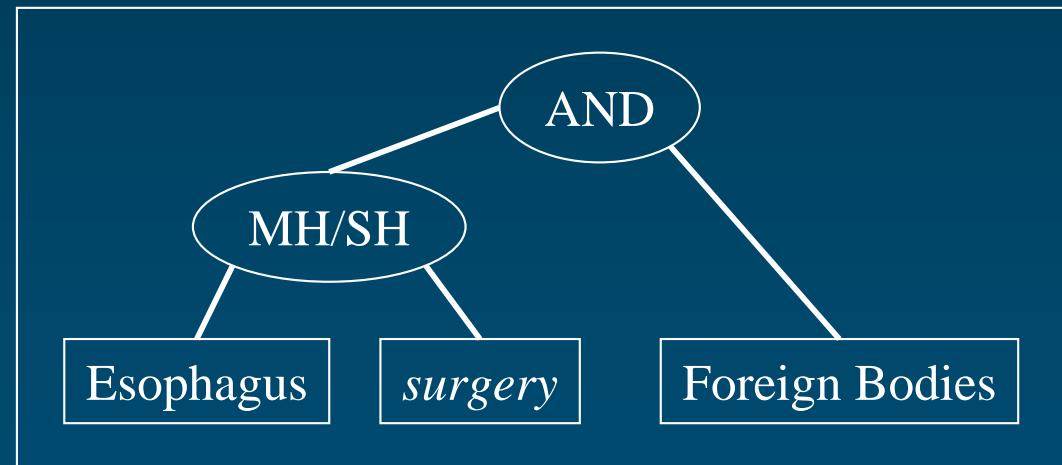
- ◆ Term mapped to Source concept
- ◆ For this concept, is there a synonym term that comes from MeSH? (MRSO)



Restrict to MeSH: Assoc. expressions

- ◆ If not,
- ◆ Is there an associated expression (ATX) that describes this concept using a combination of MeSH descriptors? (**MRATX**)

Endoscopic removal of
intraluminal foreign body
from oesophagus without
incision



Restrict to MeSH: Ancestors

- ◆ If not, let us build the graph of the ancestors of this concept
 - using parents and broader concepts (**MRREL**)
 - all the way to the top
 - excluding ancestors whose semantic types are not compatible with those of the source concept (**MRSTY**)
- ◆ From the graph, select the concepts that come from MeSH (**MRCONSO**)
- ◆ Remove those that are ancestors of another concept coming from MeSH



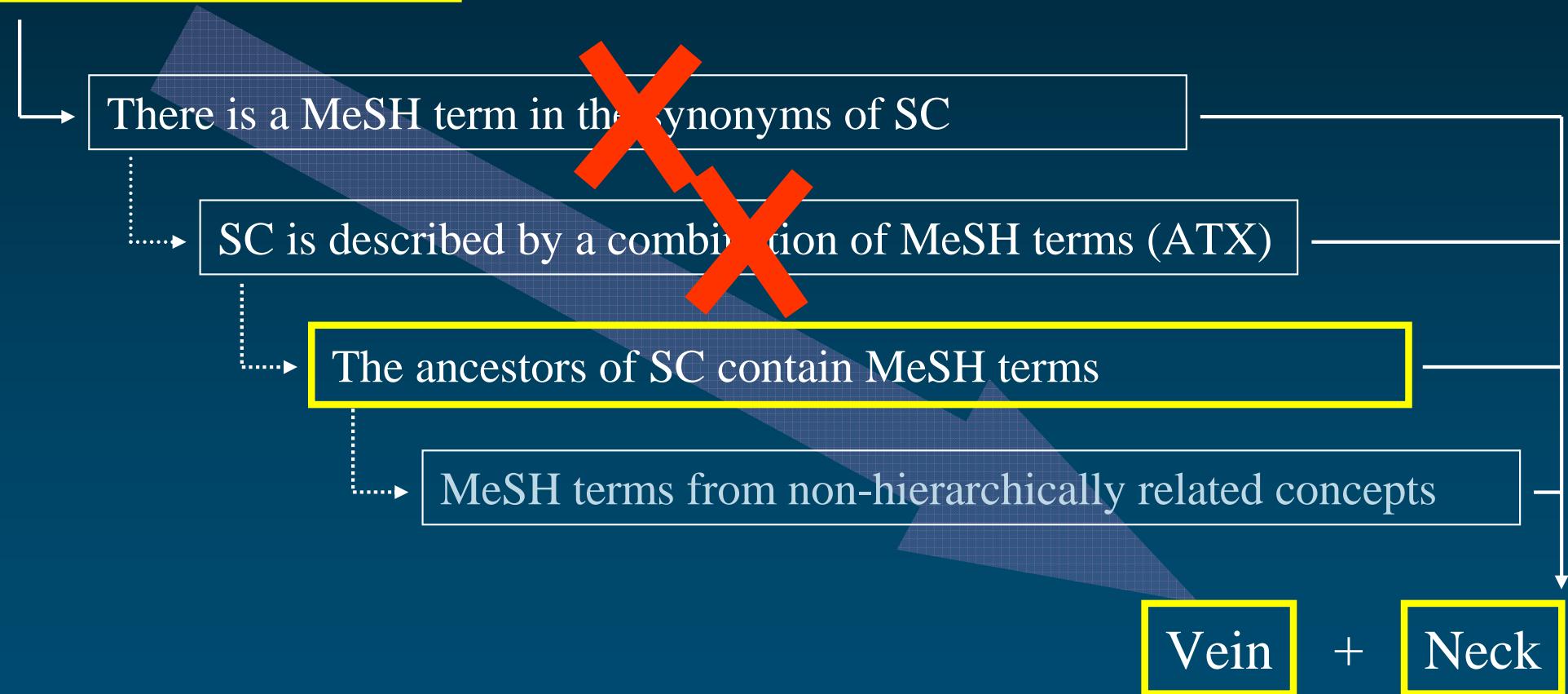
Restrict to MeSH: Other related concepts

- ◆ If not, explore the other related concepts (**MRREL**) whose semantic types are compatible with those of the source concept (**MRSTY**)
- ◆ From those, select the concepts that come from MeSH (**MRCONSO**)

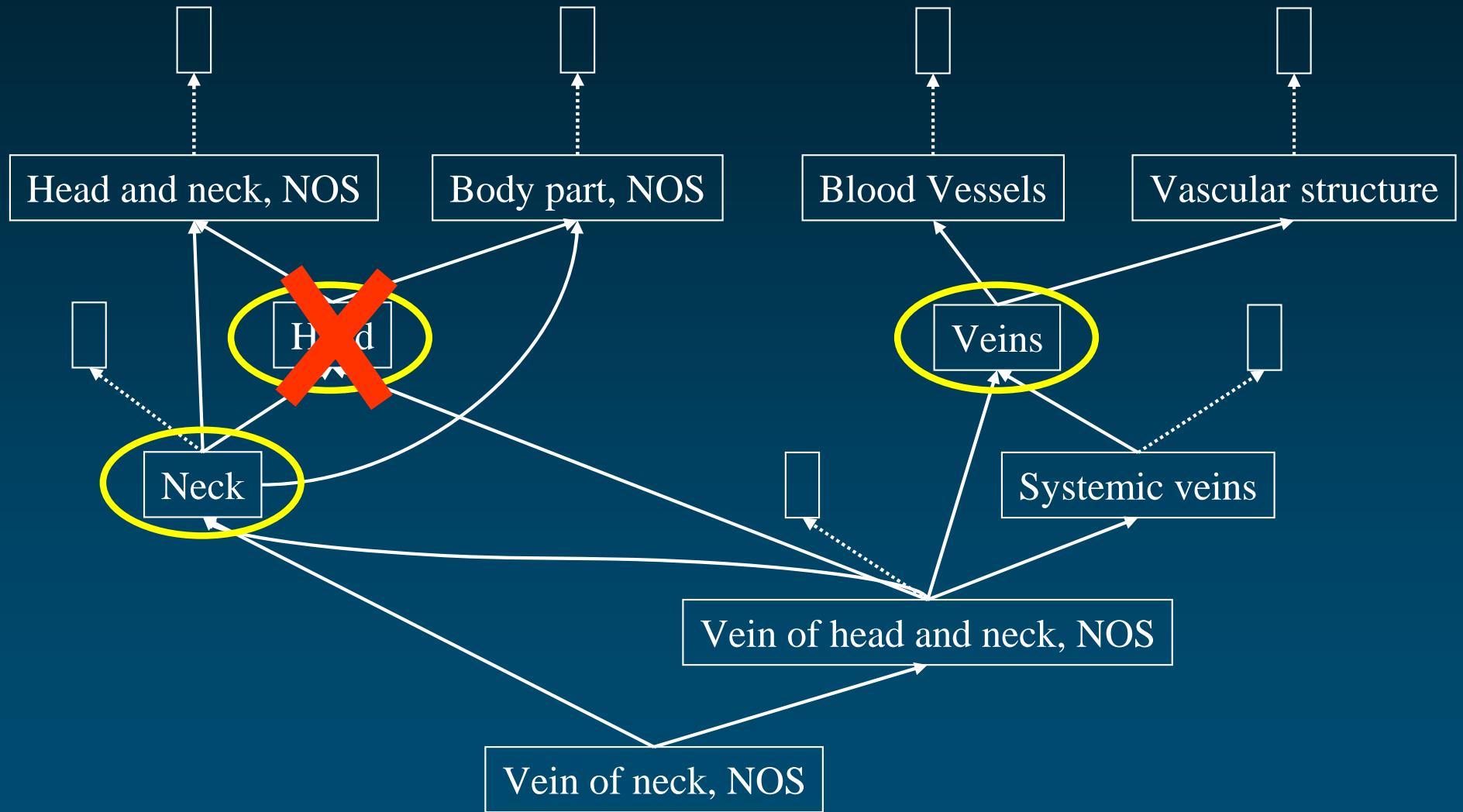


Restrict to MeSH: Example

Vein of neck, NOS



Restrict to MeSH: Example



Overall results

- ◆ Synonymy: 24%
- ◆ Built-in mapping: 1%
- ◆ Ancestors
 - From concept: 49%
 - From children: 2%
 - From siblings: 1%
- ◆ Other: 11%
- ◆ No mapping 12%



References

- ◆ Bodenreider O, Nelson SJ, Hole WT, Chang HF. *Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies*. Proceedings of AMIA Annual Symposium 1998:815-9.
<http://mor.nlm.nih.gov/pubs/pdf/1998-amia-ob.pdf>
- ◆ Fung KW, Bodenreider O. *Utilizing the UMLS for semantic mapping between terminologies*. Proceedings of AMIA Annual Symposium 2005:266-270.
<http://mor.nlm.nih.gov/pubs/pdf/2005-amia-kwf.pdf>



Advanced Library Services
Towards a *Biomedical Knowledge Repository*

Disclaimer

- ◆ The project presented in this talk is being proposed as a new research initiative at the Lister Hill National Center for Biomedical Communications
- ◆ It has not been approved or reviewed by NLM yet
- ◆ The ideas presented here may not reflect NLM's views
- ◆ In collaboration with Tom Rindflesch, NLM



Delivering Health Information

- ◆ Provide biomedical text to health care professionals and consumers
- ◆ Maintain NLM's cutting edge
 - Support public health and healthy behavior
 - Assist clinical practice
 - Enable biomedical research and discovery
- ◆ Exploit current Library resources and advanced technology



Why additional services?

- ◆ Biomedical literature is growing at an increasingly faster pace
 - High-throughput approach to literature processing
- ◆ Integration between literature and other resources is insufficient
 - Adequate for navigating purposes
 - Insufficient for knowledge processing
- ◆ Information retrieval is the starting point, not the end of the journey for the researcher



Integration for navigation purposes

<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>

NCBI

Entrez, The Life Sciences Search Engine

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

Search across databases GO CLEAR Help

692  PubMed: biomedical literature citations and abstracts	73  Books: online books
166  PubMed Central: free, full text journal articles	27  OMIM: online Mendelian Inheritance in Man
1  Site Search: NCBI web and FTP sites	none  OMIA: Online Mendelian Inheritance in Animals

278  Nucleotide: sequence database (GenBank)	17  UniGene: gene-oriented clusters of transcript sequences
160  Protein: sequence database	none  CDD: conserved protein domain database
1  Genome: whole genome sequences	8  3D Domains: domains from Entrez Structure
1  Structure: three-dimensional macromolecular structures	45  UniSTS: markers and mapping data
none  Taxonomy: organisms in GenBank	5  PopSet: population study data sets
790  SNP: single nucleotide polymorphism	1680  GEO Profiles: expression and molecular abundance profiles
35  Gene: gene-centered information	1  GEO DataSets: experimental sets of GEO data
19  HomoloGene: eukaryotic homology groups	none  Cancer Chromosomes: cytogenetic databases

What additional services?

- ◆ Multi-document summarization
 - Extract and visualize the facts extracted from 250 recent abstracts on the treatment of Parkinson's disease
- ◆ Question answering
 - Clinical and biological questions
- ◆ Knowledge discovery
 - Connect facts from heterogeneous resources
- ◆ Refined information retrieval
 - Indexing on relations in addition to concepts or association main heading/subheading



Fact-based vs. concept-based

- ◆ (concept, relationship, concept) triples are the common denominator to the various advanced services
 - Facts
 - Relations
 - Semantic predictions
 - RDF triples



Biomedical knowledge repository

- ◆ Knowledge integration
 - Unique repository
 - Common format
 - Seamless environment
 - Phenotype and genotype information together
- ◆ Enabling resource for the various services
 - Summarization
 - Question answering
 - Knowledge discovery
 - Refined information retrieval



Sources of knowledge

◆ Biomedical literature

- Facts extracted from MEDLINE abstracts and full-text publicly available articles using text mining techniques
- Other corpora

◆ Structured databases / knowledge bases

- NCBI resources
- Model organism databases
- Terminological knowledge
- ...

◆ Contributed knowledge

- The repository is open to collaborators outside NLM



Annotated knowledge

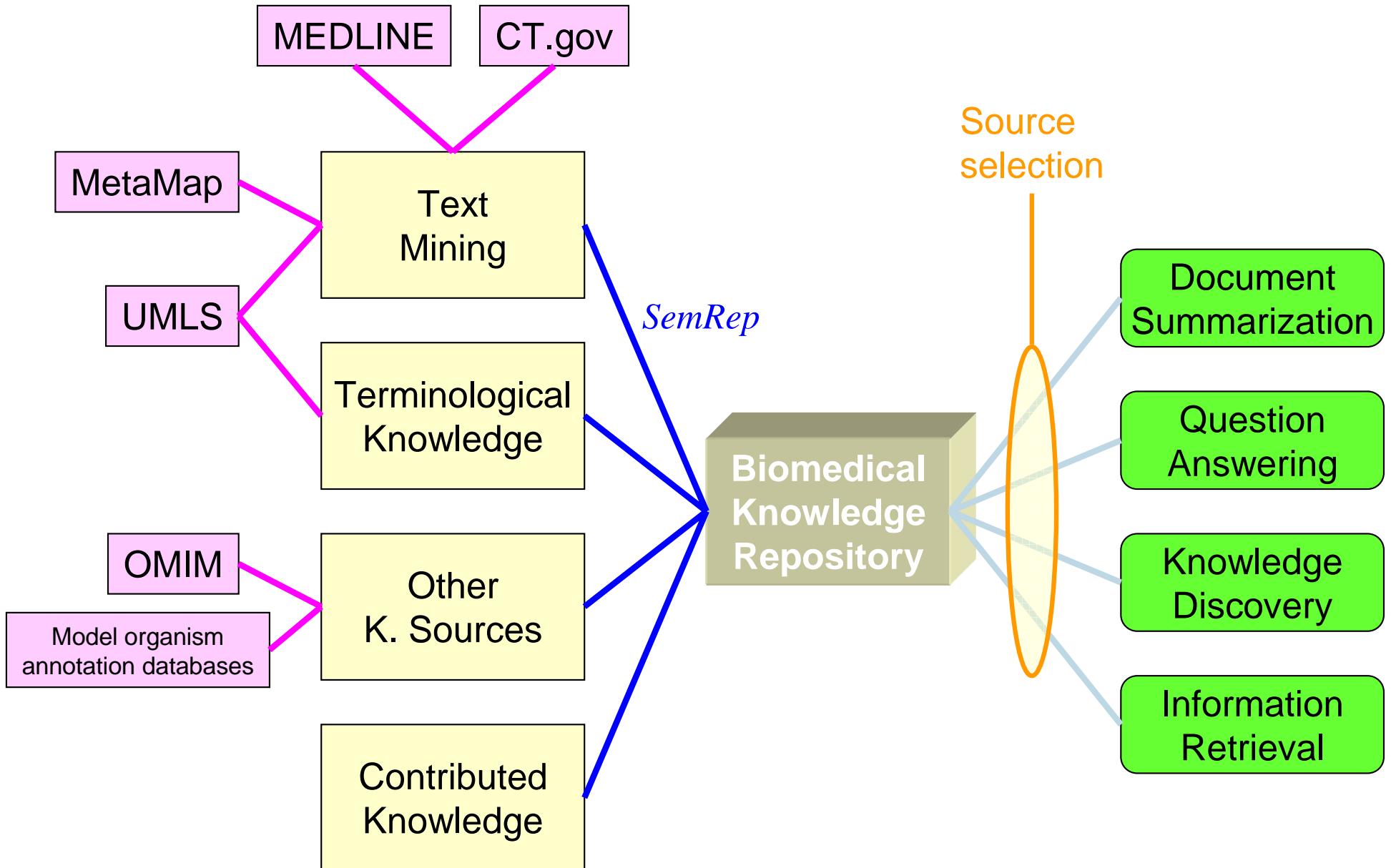
- ◆ Provenance information
 - Source (e.g., PMID)
 - Extraction mechanism
 - Timestamp
- ◆ Frequency information
 - Redundancy
- ◆ Collaborative annotation
 - “Was this information useful?”
 - Context of use/usefulness



Semantic Web perspective

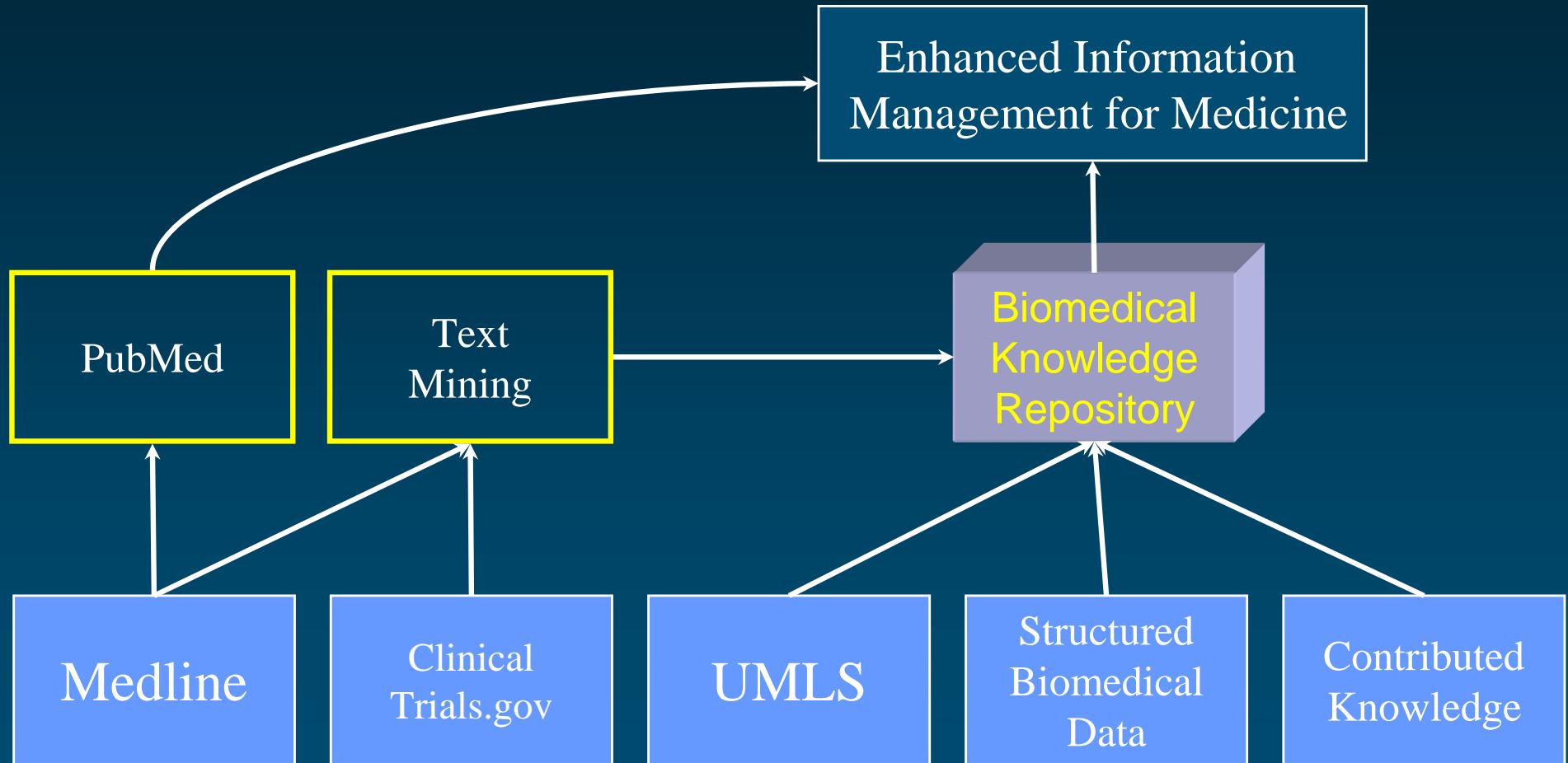
- ◆ Common format for knowledge
 - Resource Description Format (RDF)
- ◆ Common identification scheme
 - Unified Resource Identifier (URI)
- ◆ Standard tools
 - RDF browsers
 - RDF “reasoners”
- ◆ High level of interest for biomedicine in the SW community
 - Health Care and Life Sciences Interest Group



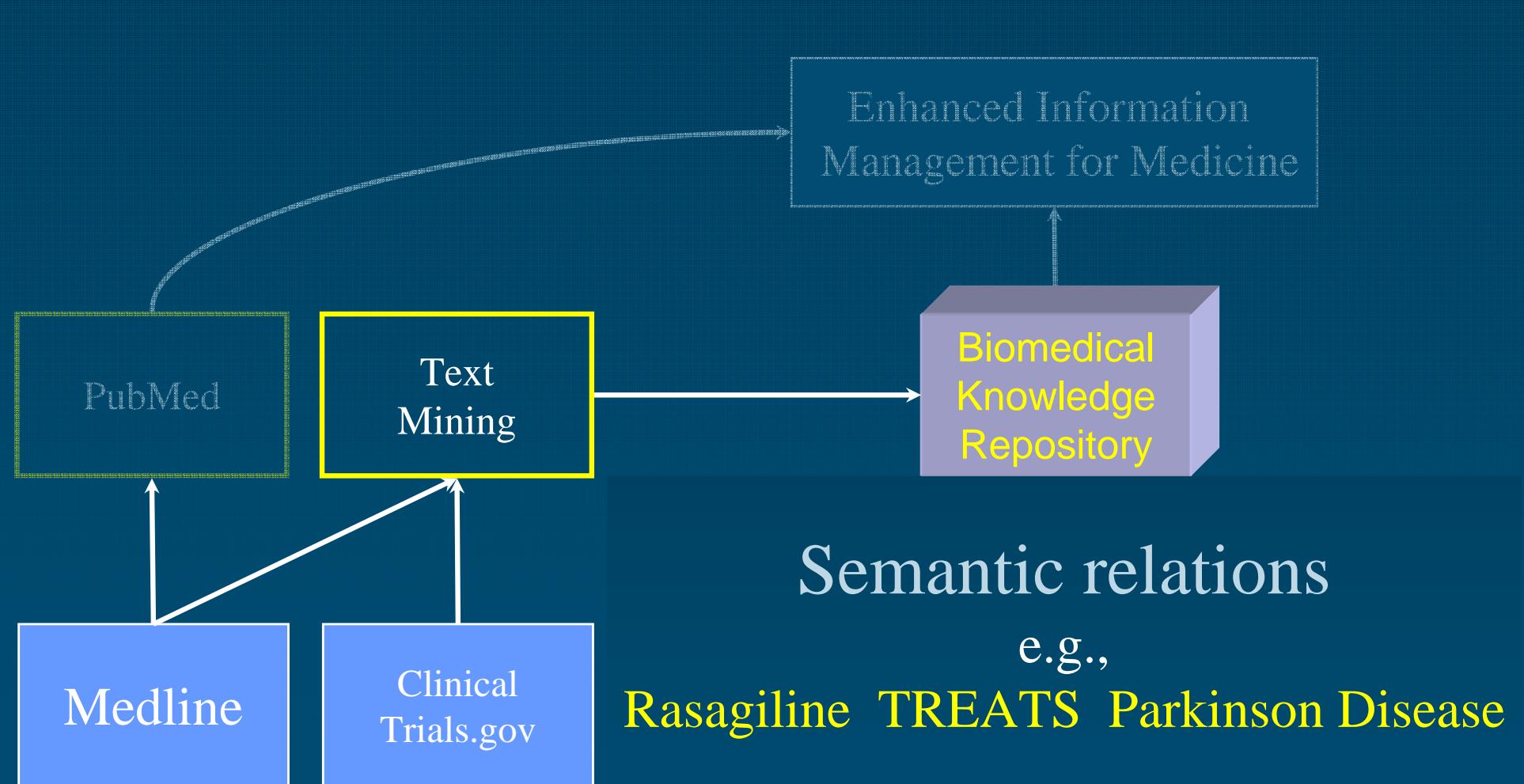


Towards a Biomedical Knowledge Repository

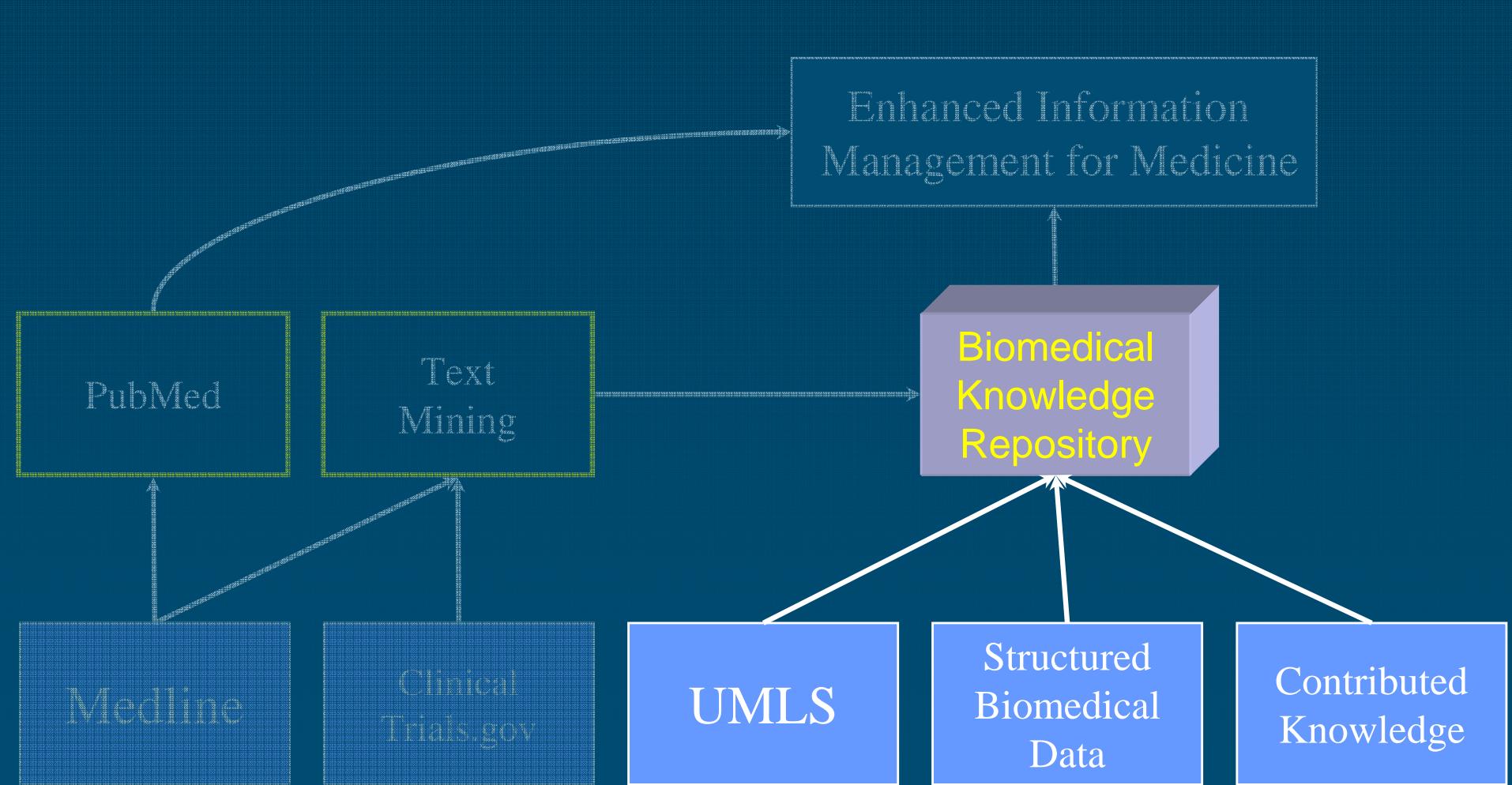
Creating the repository



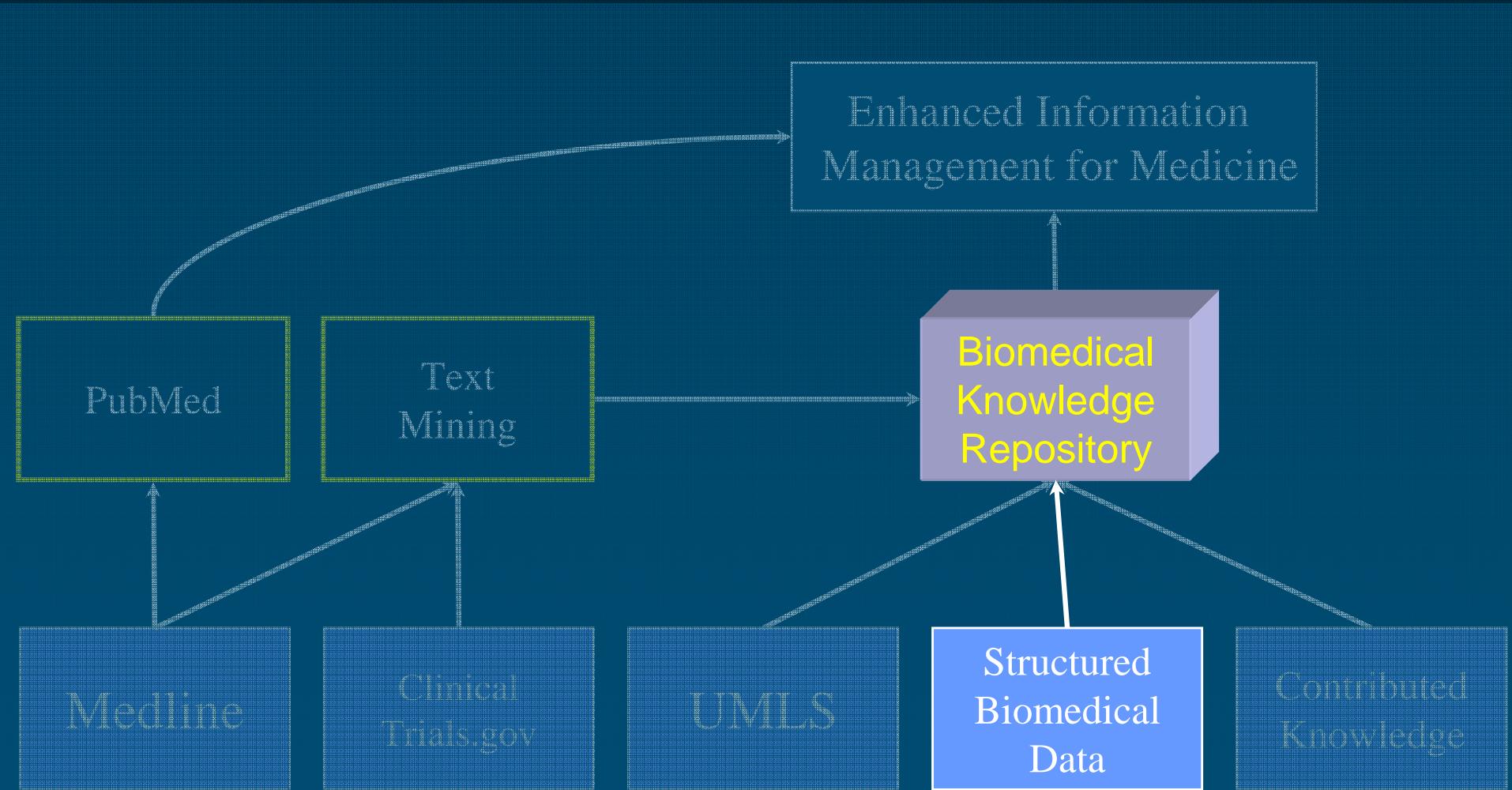
Creating the repository



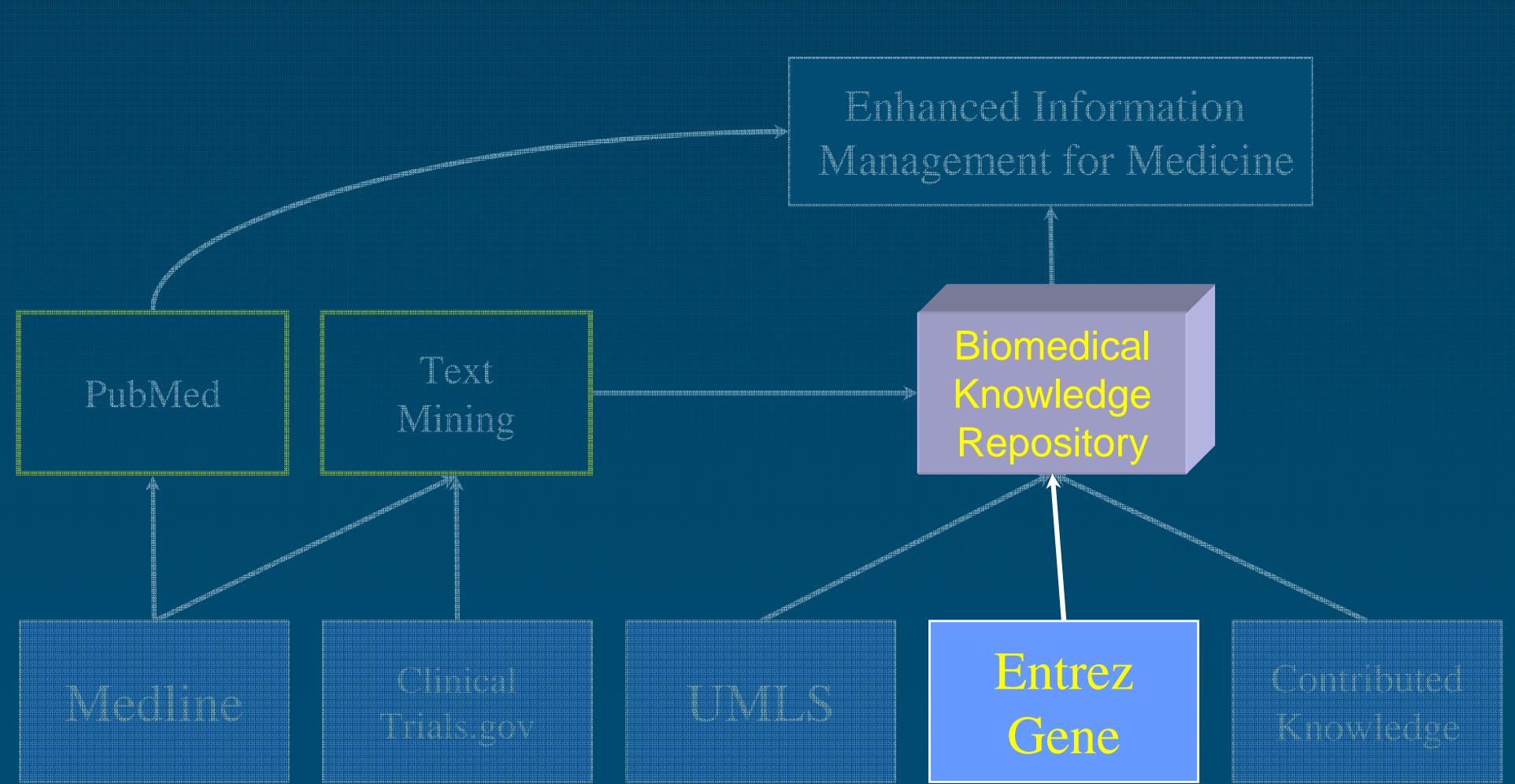
Creating the repository



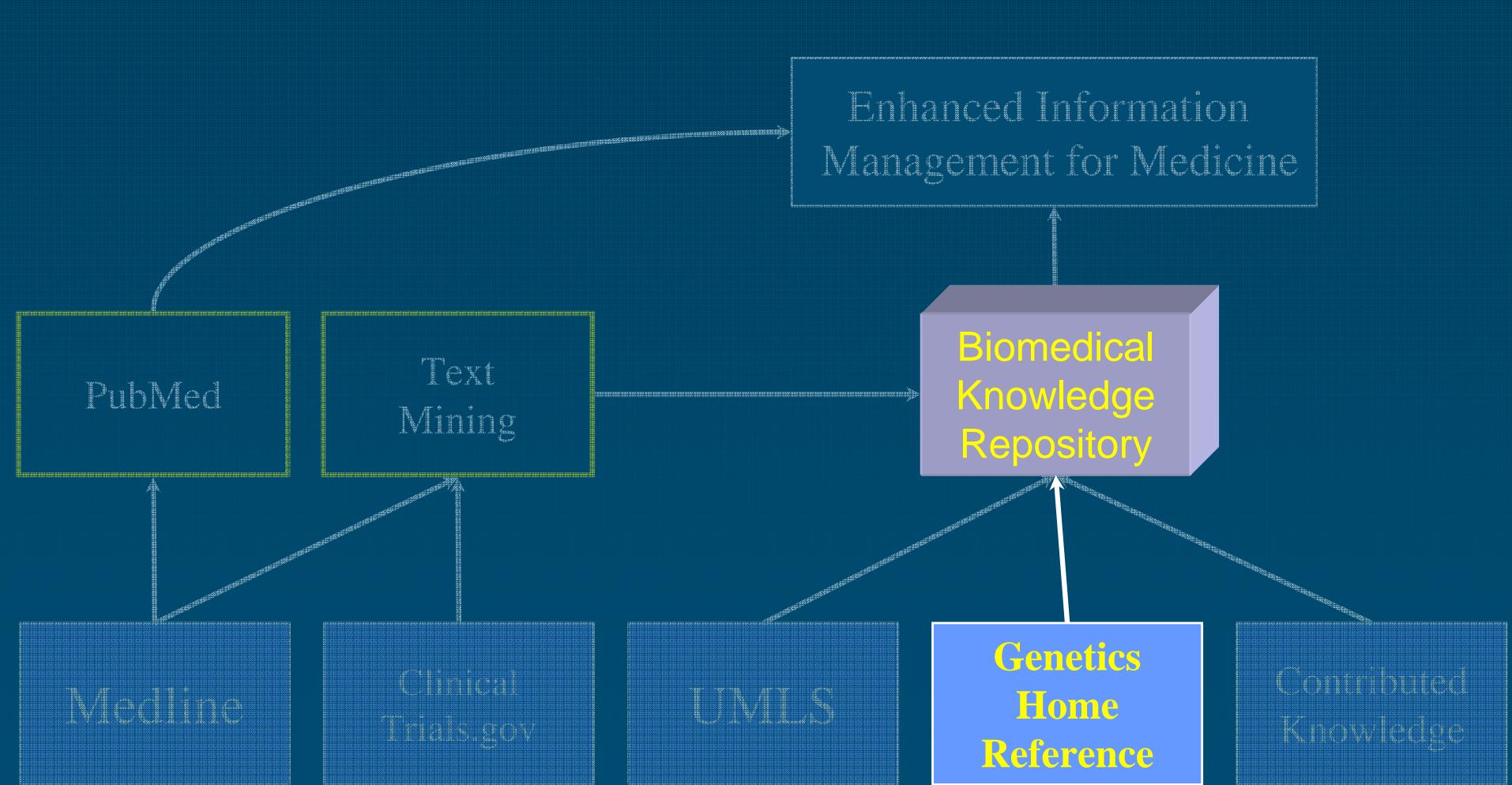
Creating the repository



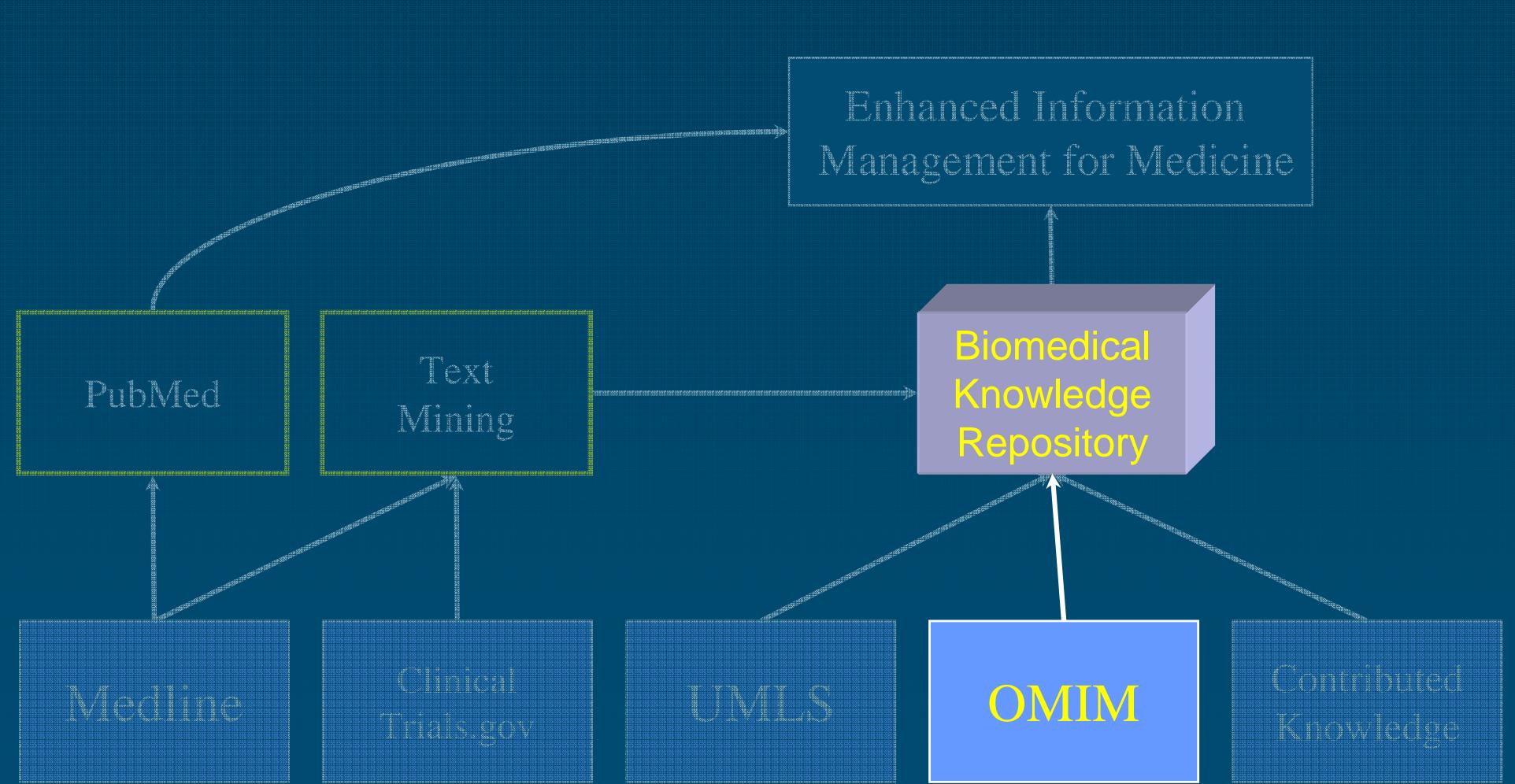
Creating the repository



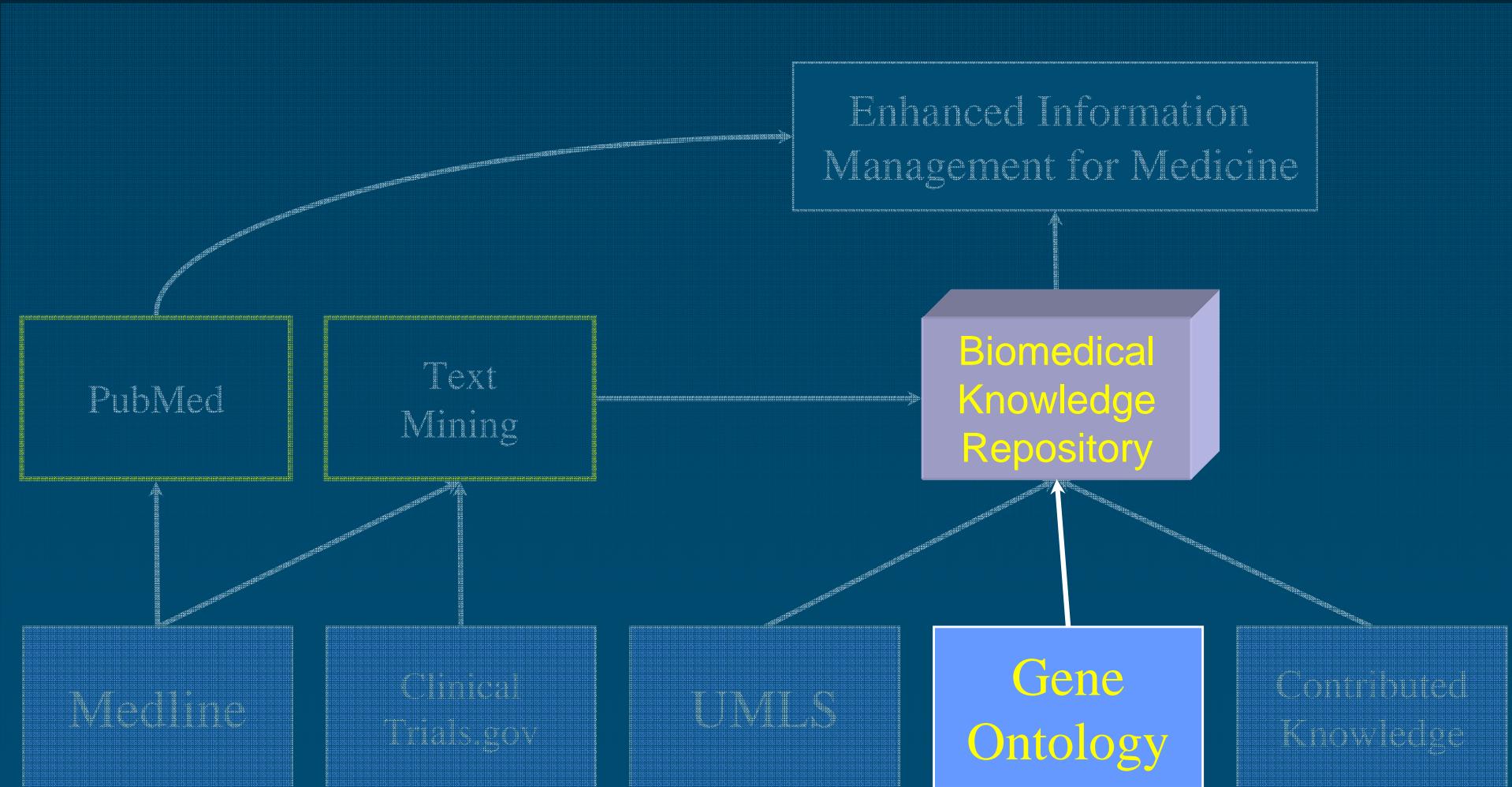
Creating the repository



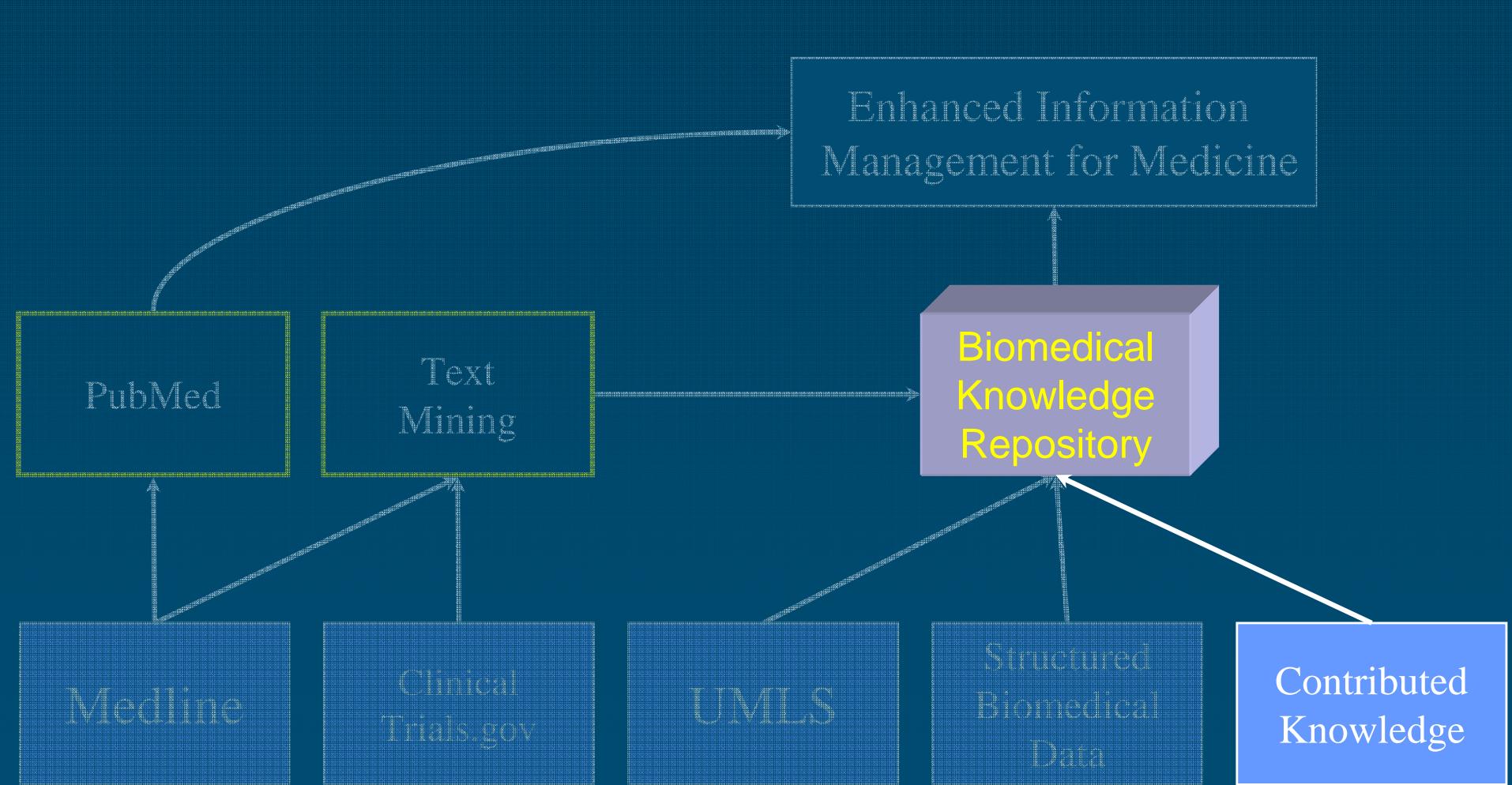
Creating the repository



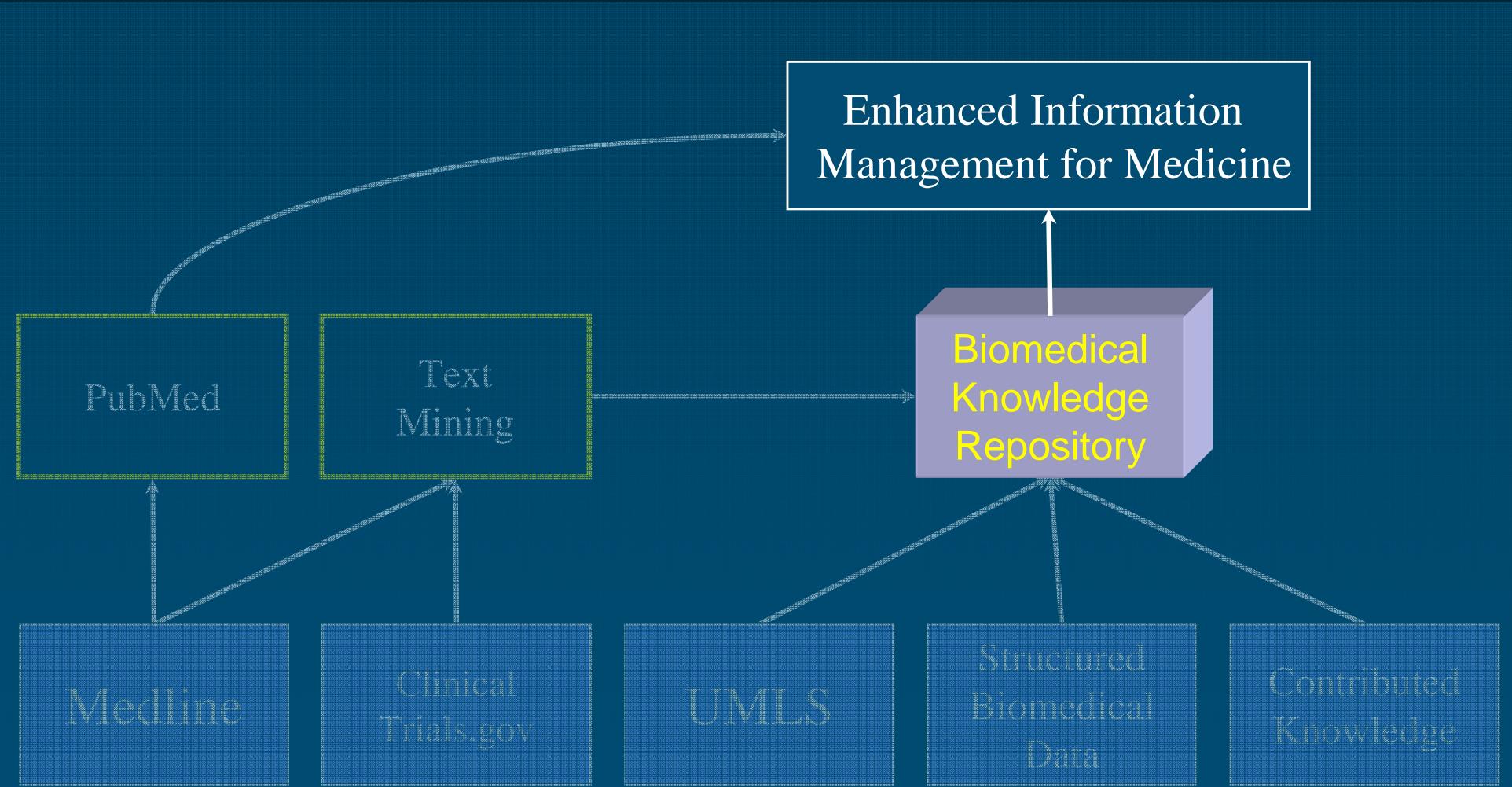
Creating the repository



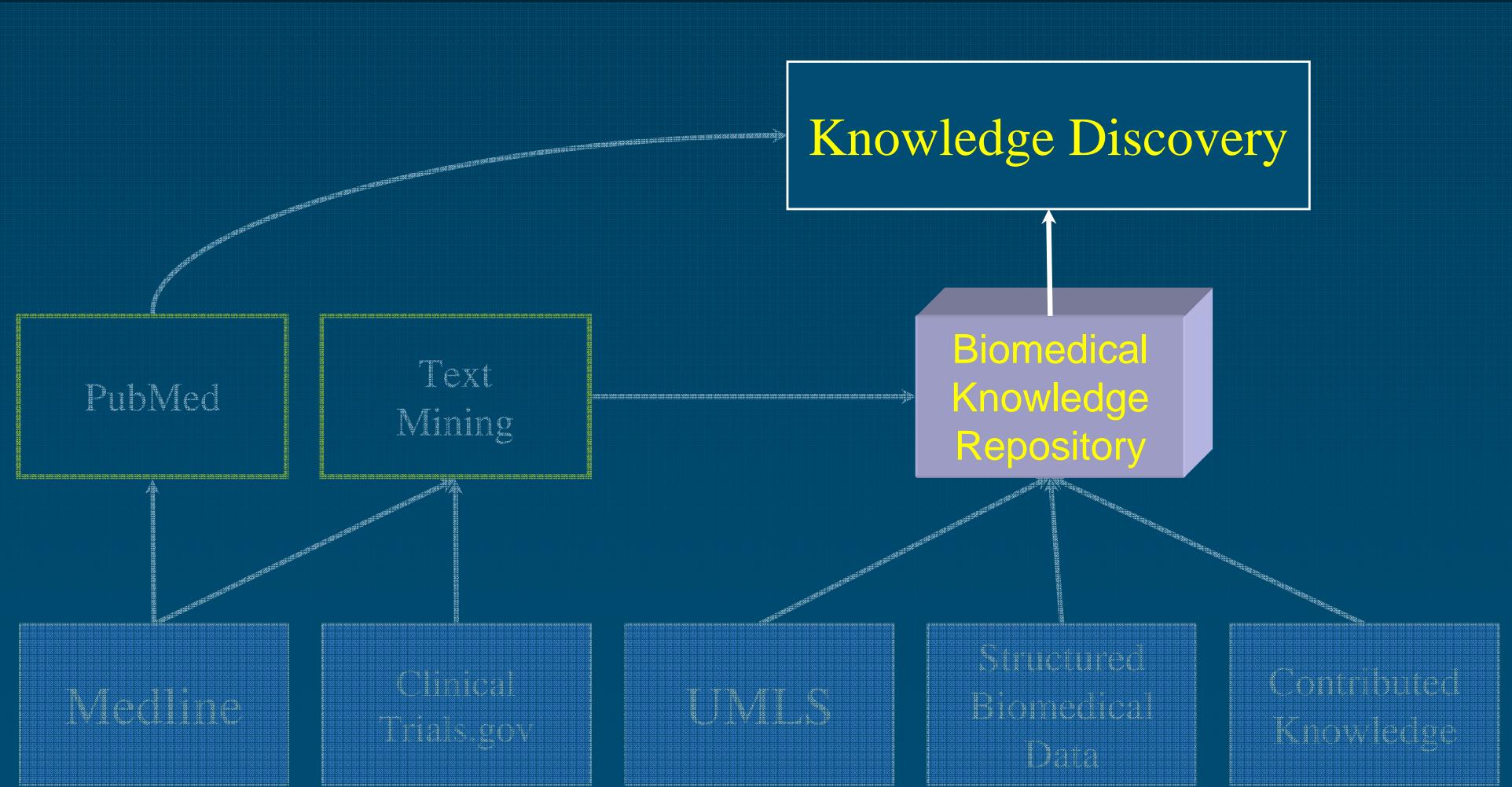
Creating the repository



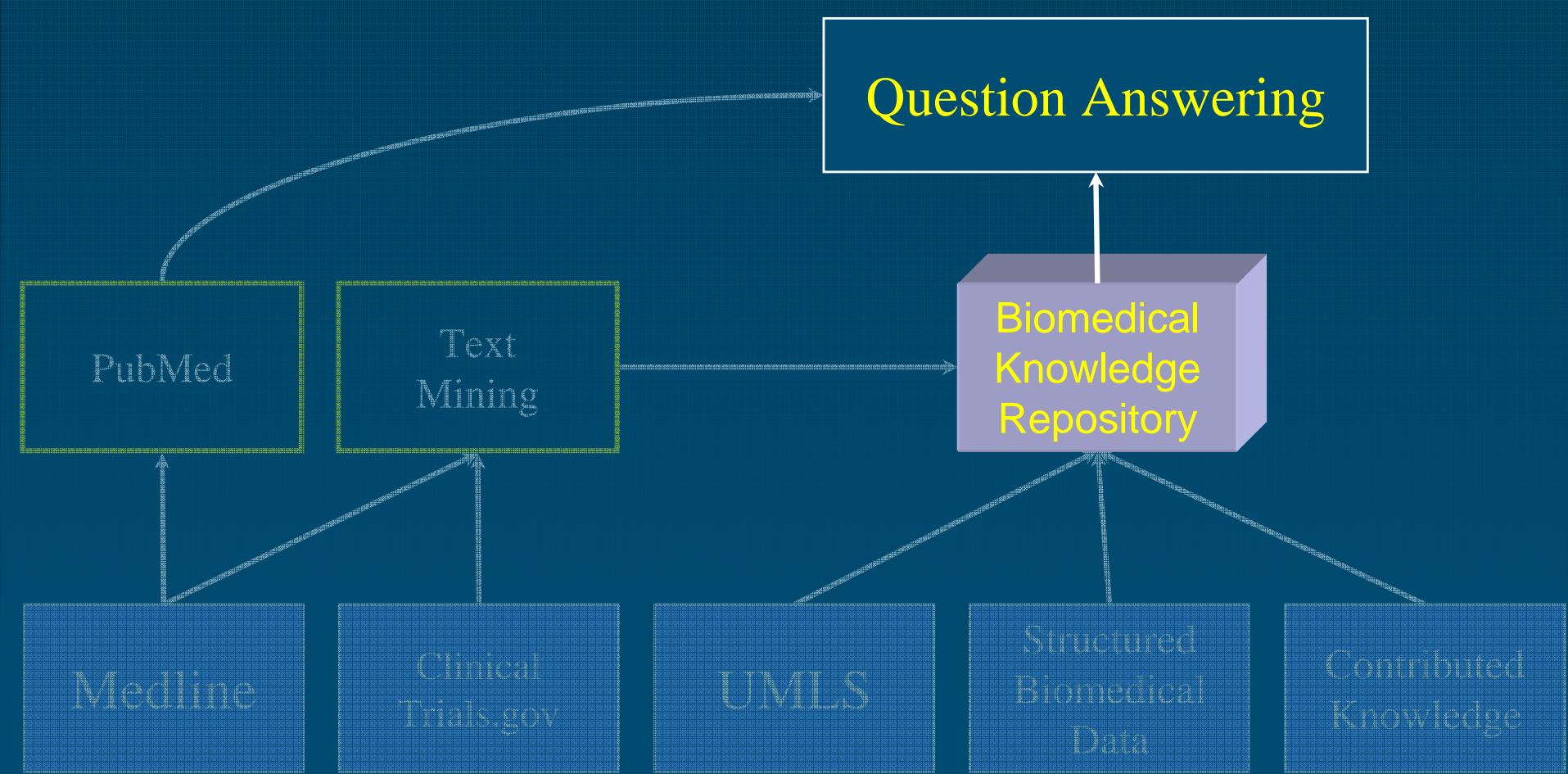
Advanced library services



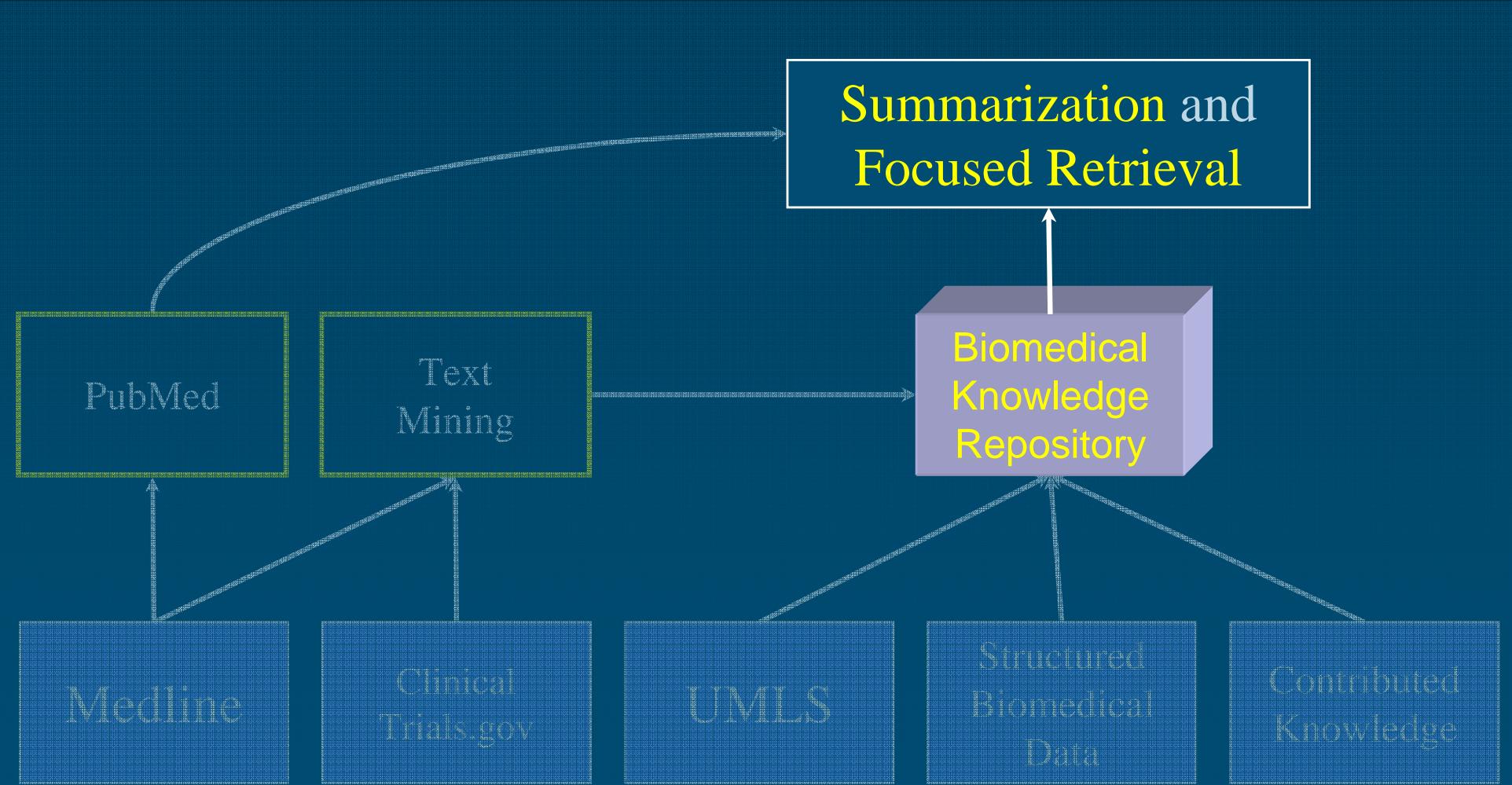
Advanced library services



Advanced library services



Advanced library services



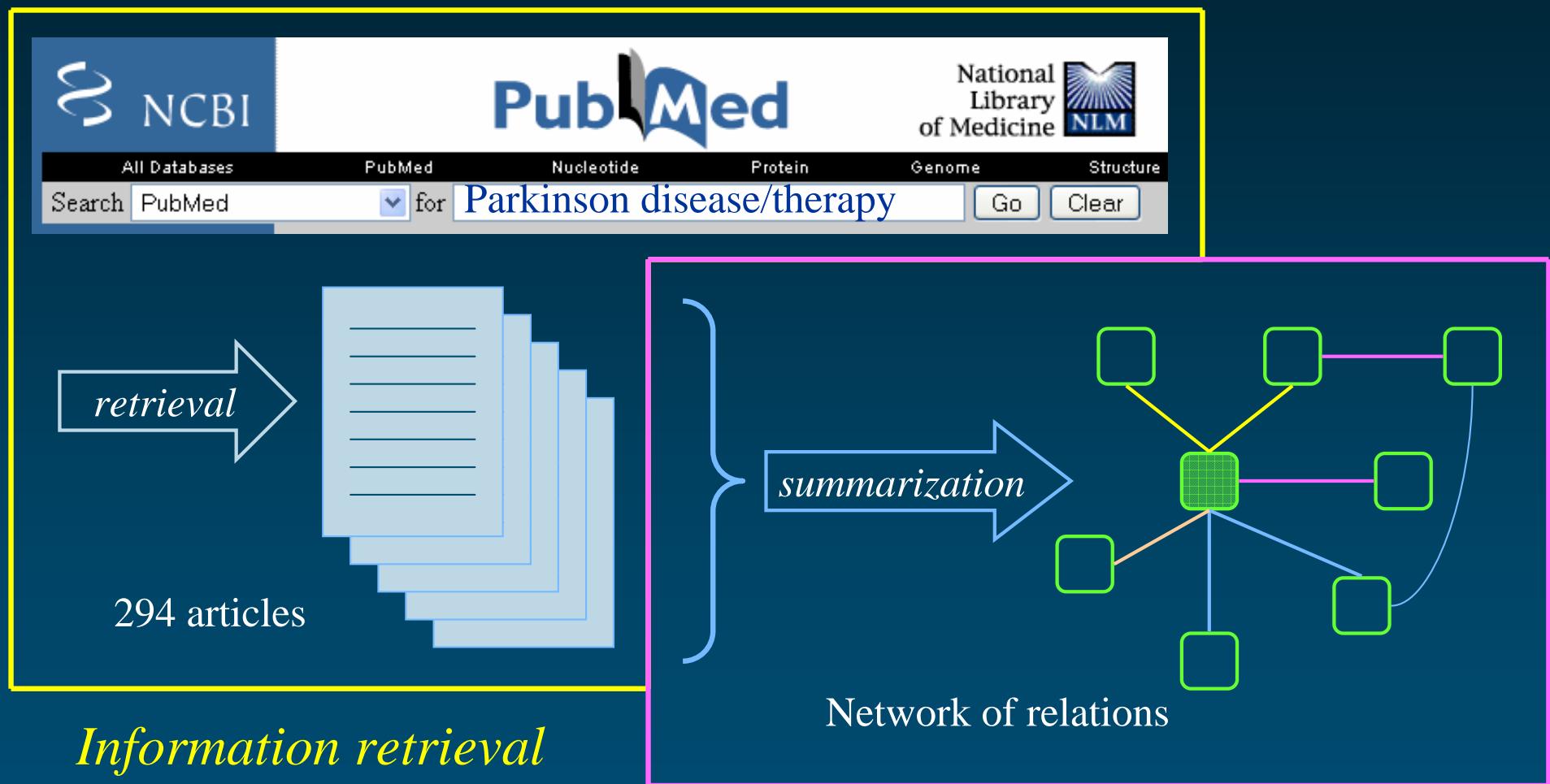
Summarizing Biomedical Text

Summarizing Biomedical Text

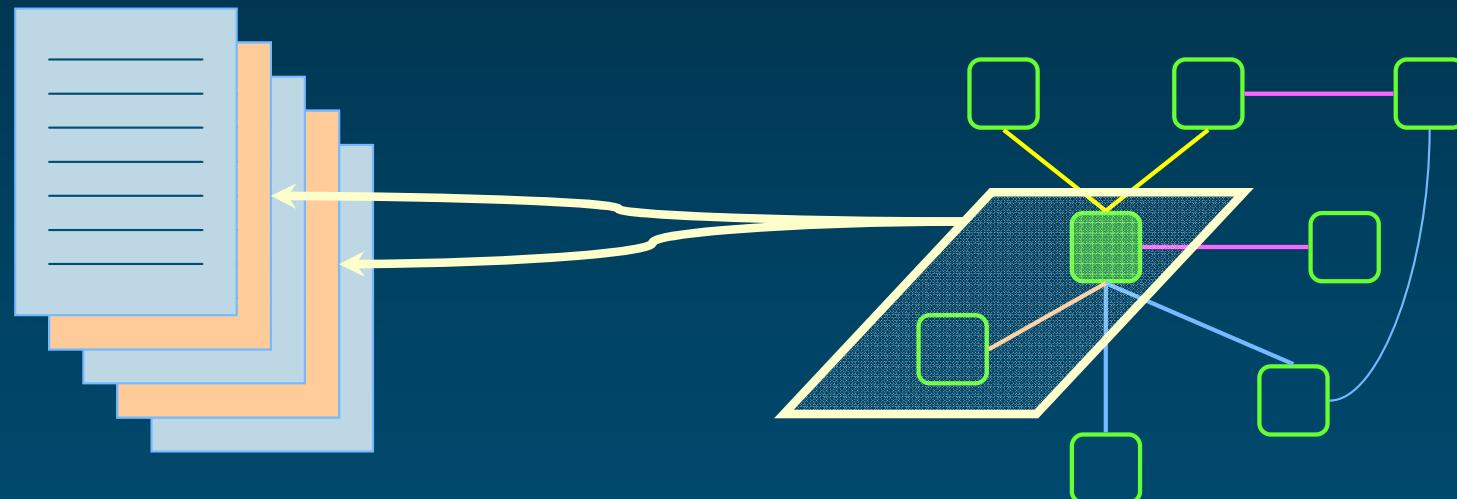
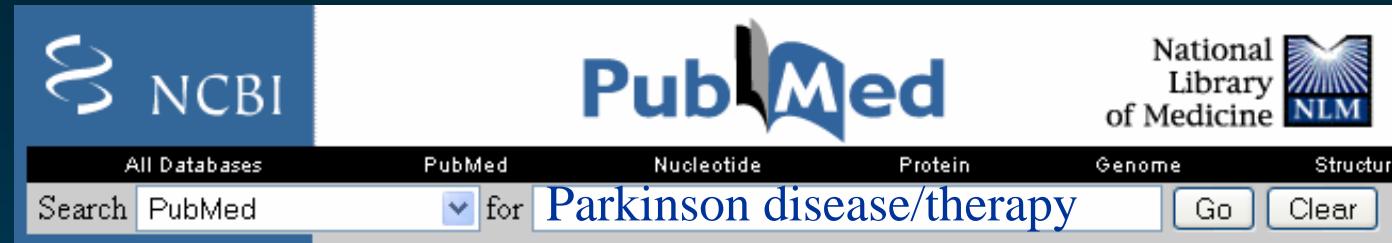
- ◆ Search
 - Medline
 - ClinicalTrials.gov
- ◆ Summarize documents
 - Most salient semantic relations
- ◆ Visualize the summary
- ◆ Link the semantic relations to
 - Original text
 - Related structured knowledge

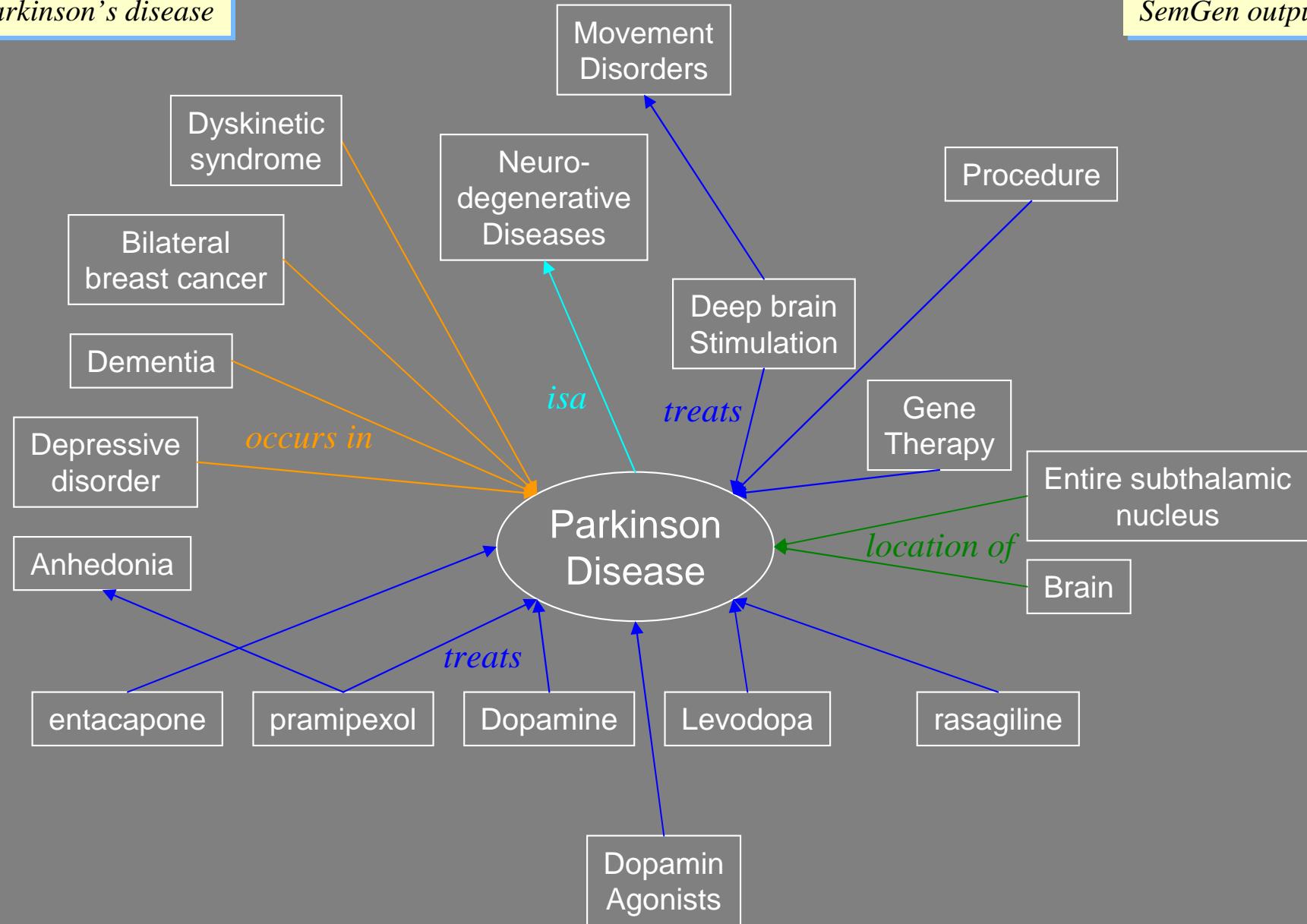


Text Mining Workflow



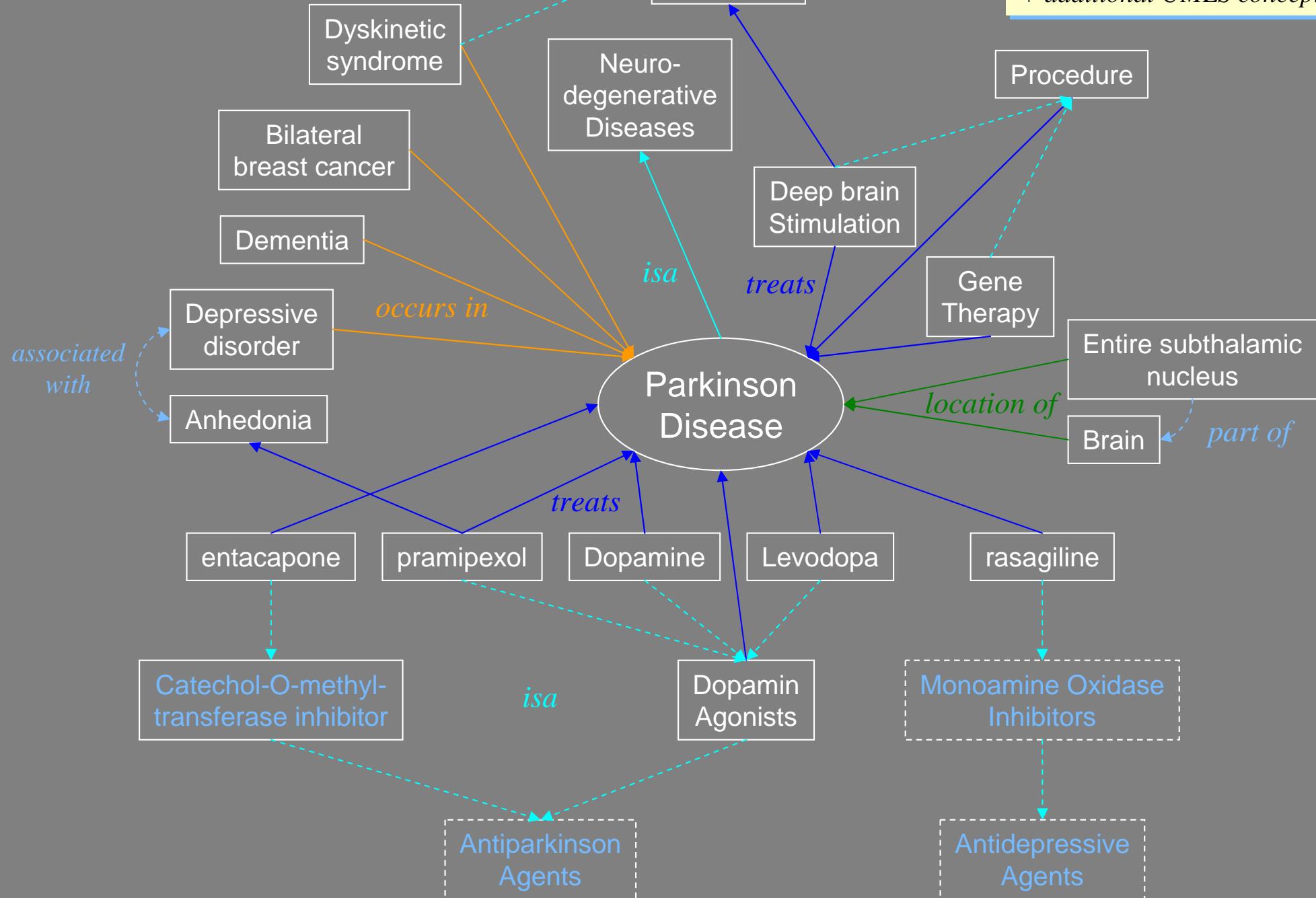
Text Mining Workflow





Treatment of Parkinson's disease

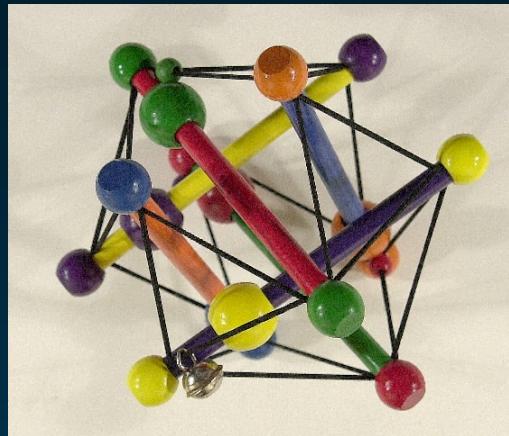
SemGen output
+ UMLS relations
+ additional UMLS concepts



Conclusions

- ◆ Need to go beyond information retrieval
- ◆ Need to integrate multiple, heterogeneous knowledge sources to support knowledge processing, not only navigation
- ◆ Synergistic with the Semantic Web
 - Emerging standard framework
 - W3C Health Care and Life Sciences Interest Group
<http://www.w3.org/2001/sw/hcls/>





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