University of Pisa, Italy
June 12, 2007

NETTAB 2007 - A Semantic Web for Bioinformatics

Tutorial T5
The Unified Medical Language System (UMLS) and the Semantic Web

Olivier Bodenreider

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

Outline

- Information integration in biomedicine
  - Some issues: naming, normalization, mapping
  - Semantic Web perspective
- Terminology integration in biomedicine
  Unified Medical Language System
- Some differences between UMLS and SW

Information integration in biomedicine

Some issues: naming, normalization, mapping

❶ Naming

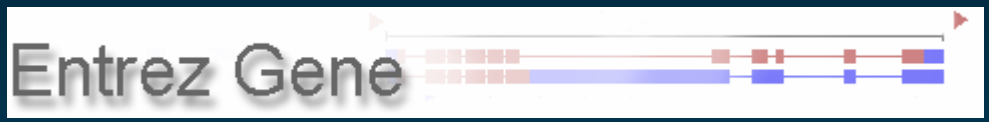
- Many biomedical entities have several names (synonymy)
  - Drug names
  - Gene names
  - Disease names
  - …
- A given name may refer to several different entities (polysemy)
  - Nail (body part)
  - Nail (medical device)

Brand names for paracetamol (acetaminophen)

http://en.wikipedia.org/wiki/List_of_paracetamol_brand_names

Brand name	Countries
Acamol	Israel
Atamel	Venezuela
Adol	Oman
Aldolor	Israel
Alvedon	Sweden
APAP	Poland
Benuron	Portugal, Germany
Biogesic	Philippines
Buscapina	Argentina
Cemol	Thailand
Crocin	India
Dafalgan	Belgium, France, Portugal, Russia, Ukraine
Daleron	Slovenia
Depon	Greece
Dexamol	Israel
Dolex	Colombia
Doliprane	France, Portugal, Russia, Ukraine
Efferalgan	France, Italy, Portugal, Russia, Spain, Ukraine
FeverAll	United States
Gelocatil	Spain
Gripin	Turkey
Lekadol	Croatia, Slovenia
Metacin	India
Pamol	Denmark, Finland, France
Panado	South Africa
Panadol	Australia, Azerbaijan, Central America, Egypt, Finland, Greece, Hong Kong, Hungary, Indonesia, Ireland, Kenya, Lebanon, Macedonia, Malaysia, Malta, Netherlands, New Zealand, Nigeria, Pakistan, Poland, Portugal, Romania, Russia, Saudi Arabia, Singapore, Sri Lanka, Switzerland, Taiwan, Ukraine, Estonia, United Kingdom
Panamax	Australia, United Kingdom
Panodil	Denmark, Iceland, Sweden
Paracet	Norway
Paralen	Czech Republic, Slovakia
Paramed	Botswana, South Africa, Zimbabwe
Paramol	Israel, Taiwan
Perdolan	Belgium
Perfalgan	Germany
Pinex	Denmark, Iceland, Norway
Plicet	Croatia
Reliv	Sweden
Rokamol	Israel
Sara	Thailand
Tachipirina	Italy
Tylenol	Brazil, Canada, Japan, South Korea, Thailand, United States
Tempra	Philippines

Names for dystrophin



http://www.ncbi.nlm.nih.gov/sites/entrez

DMD

Order cDNA clone, Links

Official Symbol DMD and **Name:** dystrophin (muscular dystrophy, Duchenne and Becker types) [*Homo sapiens*]
Other Aliases: GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272
Other Designations: Duchenne muscular dystrophy protein; dystrophin
Chromosome: X; **Location:** Xp21.2
Annotation: Chromosome X, NC_000023.9 (33267646..31047265, complement)
MIM: 300377
GeneID: 1756

Names for renal cell carcinoma

Details of 'clear cell carcinoma of kidney' | Distributed Relationships

ConceptStatus **Current**

Descriptions

- F clear cell carcinoma of kidney (disorder)
- P clear cell carcinoma of kidney
- S adenocarcinoma of kidney
- S carcinoma of kidney
- S Grawitz tumor
- S renal cell adenocarcinoma
- S renal cell carcinoma

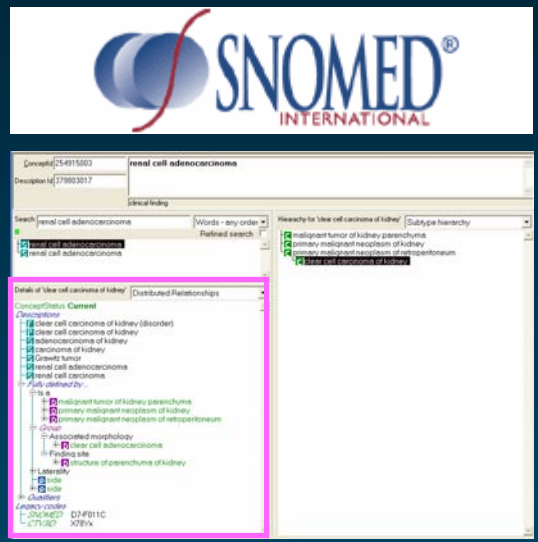
Fully defined by...

- Is a
  - D maligqnant tumor of kidney parenchyma
  - D primary maligqnant neoplasm of kidney
  - D primary maligqnant neoplasm of retroperitoneum
- *Group*
  - Associated morphology
    - D clear cell adenocarcinoma
  - Finding site
    - D structure of parenchyma of kidney
- Laterality
  - side
  - side

Qualifiers

Legacy codes

- *SNOMED:* D7-F011C
- *CTV3ID:* X78Yx



http://www.clininfo.co.uk/clue5/clue.htm

Entity recognition

- Identifying biomedical entities in text
  - Names entity recognition
  - Tagging “mentions”
  - Semantic annotation
- Supported by terminology
  - Collects the names used in the domain
  - Often incompletely
- Example: BioCreative
  - 1A – Gene name identification
  - 2GM – Gene mention tagging

❷ Normalization

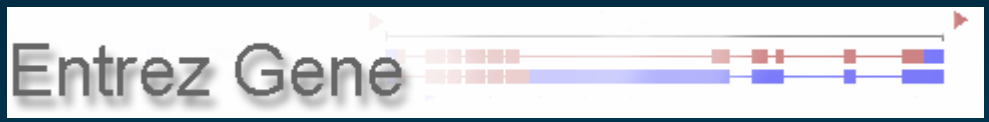
- Biomedical entities are identified by unique identifiers in various terminology systems
- Resolve names into identifiers (in a given namespace)
- Supported (in part) by terminology resources
- Example: BioCreative
  - 1B and 2GN – Gene Normalization

Identifier for paracetamol (acetaminophen)

Master Drug Data Base. Medi-Span	5005	Acetaminophen
FDA National Drug Code Directory	50612	PARACETAMOL
FDA Structured Product Labels	362O9ITL9D	ACETAMINOPHEN
First DataBank NDDF Plus	001605	Acetaminophen
SNOMED Clinical Terms	90332006	Acetaminophen (product)
SNOMED Clinical Terms	387517004	Acetaminophen (substance)
VA National Drug File	4017513	ACETAMINOPHEN

Source: RxNorm database (5/3/2007)

Identifier for dystrophin



http://www.ncbi.nlm.nih.gov/sites/entrez

DMD

Order cDNA clone, Links

Official Symbol DMD and **Name:** dystrophin (muscular dystrophy, Duchenne and Becker types) [*Homo sapiens*]
Other Aliases: GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272
Other Designations: Duchenne muscular dystrophy protein; dystrophin
Chromosome: X; **Location:** Xp21.2
Annotation: Chromosome X, NC_000023.9 (33267646..31047265, complement)
MIM: 300377
GeneID: 1756

Identifier for renal cell carcinoma

Details of 'clear cell carcinoma of kidney' | Distributed Relationships

ConceptStatus **Current**

Descriptions

- F clear cell carcinoma of kidney (disorder)
- P clear cell carcinoma of kidney
- S adenocarcinoma of kidney
- S carcinoma of kidney
- S Grawitz tumor
- S renal cell adenocarcinoma
- S renal cell carcinoma

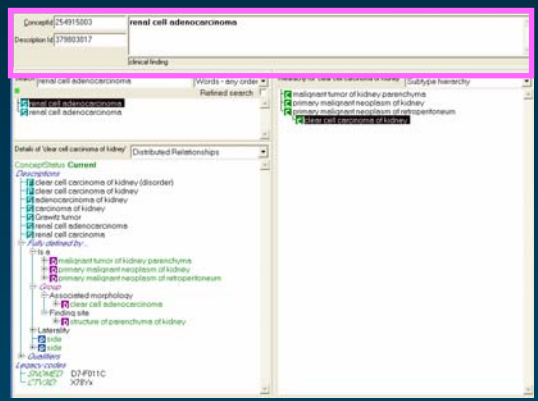
Fully defined by...

- Is a
  - D maligqnant tumor of kidney parenchyma
  - D primary maligqnant neoplasm of kidney
  - D primary maligqnant neoplasm of retroperitoneum
- *Group*
  - Associated morphology
    - D clear cell adenocarcinoma
  - Findinq site
    - D structure of parenchyma of kidney
- Laterality
  - side
  - side

Qualifiers

Legacy codes

- *SNOMED:* D7-F011C
- *CTV3ID:* X78Yx



ConceptId 254915003 | **renal cell adenocarcinoma**

Description Id 379803017 | clinical finding

http://www.clininfo.co.uk/clue5/clue.htm

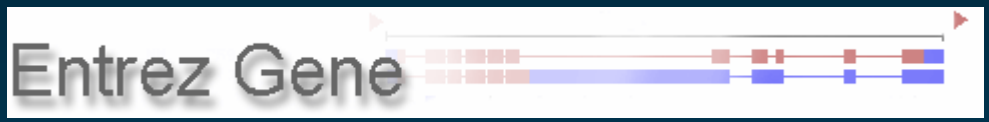
❸ Mapping / Integration

- Identify equivalent entities across systems (across namespaces)
  - Shared identifiers
  - Existing mappings (e.g., SNOMED CT to ICD-9-CM)
  - Ontology alignment techniques (lexical + structural)
- Align equivalent entities
  - Pairwise: mapping
  - More broadly: integration
- Forms the basis for information integration in the Semantic Web (mashups)

Identifier for paracetamol (acetaminophen)

Master Drug Data Base. Medi-Span	5005	Acetaminophen
FDA National Drug Code Directory	50612	PARACETAMOL
FDA Structured Product Labels	362O9ITL9D	ACETAMINOPHEN
First DataBank NDDF Plus	001605	Acetaminophen
SNOMED Clinical Terms	90332006	Acetaminophen (product)
SNOMED Clinical Terms	387517004	Acetaminophen (substance)
VA National Drug File	4017513	ACETAMINOPHEN
RxNorm	161	Acetaminophen

Identifier for dystrophin



http://www.ncbi.nlm.nih.gov/sites/entrez

DMD — Order cDNA clone, Links

Official Symbol DMD and **Name:** dystrophin (muscular dystrophy, Duchenne and Becker types) [*Homo sapiens*]
Other Aliases: GS1-19O24.1, BMD, CMD3B, DXS142, DXS164, DXS206, DXS230, DXS239, DXS268, DXS269, DXS270, DXS272
Other Designations: Duchenne muscular dystrophy protein; dystrophin
Chromosome: X; **Location:** Xp21.2
Annotation: Chromosome X, NC_000023.9 (33267646..31047265, complement)
MIM: 300377
GeneID: 1756

Identifier for renal cell carcinoma

Details of 'clear cell carcinoma of kidney' | Distributed Relationships

ConceptStatus **Current**

Descriptions

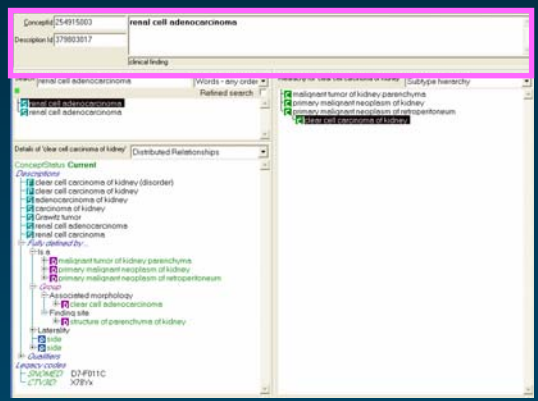
F clear cell carcinoma of kidney (disorder)	645875019
P clear cell carcinoma of kidney	379798014
S adenocarcinoma of kidney	379801015
S carcinoma of kidney	379800019
S Grawitz tumor	379797016
S renal cell adenocarcinoma	379803017
S renal cell carcinoma	379802010

Fully defined by...

- Is a
  - D maligqnant tumor of kidney parenchyma
  - D primary maligqnant neoplasm of kidney
  - D primary maligqnant neoplasm of retroperitoneum
- *Group*
  - Associated morphology
    - D clear cell adenocarcinoma
  - Findinq site
    - D structure of parenchyma of kidney
- Laterality
  - R side
  - R side
- *Qualifiers*

Legacy codes

- *SNOMED:* D7-F011C
- *CTV3ID:* X78Yx



Concept Id 254915003

Description Id 379803017

renal cell adenocarcinoma

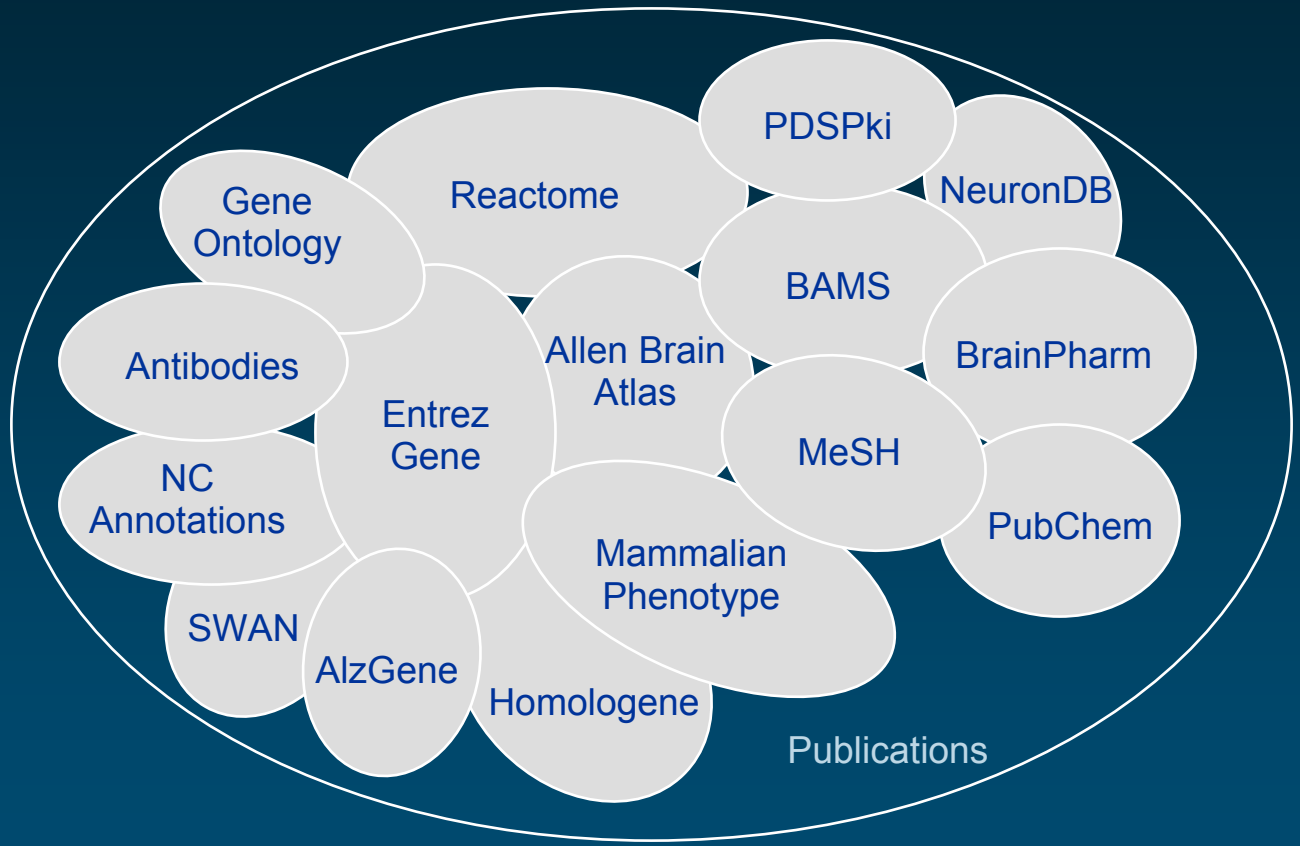
clinical finding

http://www.clininfo.co.uk/clue5/clue.htm

Information integration in biomedicine

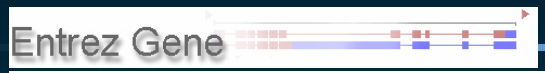
Semantic Web perspective

HCLS mashup



http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo

Shared identifiers Example



CH25H — Order cDNA clone, Links

Official Symbol CH25H and Name: cholesterol 25-hydroxylase [*Homo sapiens*]
Other Aliases: C25H
Chromosome: 10; **Location:** 10q23
Annotation: Chromosome 10, NC_000010.9 (90957050..90955509, complement)
MIM: 604551
GeneID: 9023

Pathways

Reactome Event: 73923 Lipid and lipoprotein metabolism

Homology

Mouse, Rat
Map Viewer

GeneOntology

Function
- iron ion binding
- metal ion binding
- steroid hydroxylase activity

Process
- cholesterol metabolic process
- lipid metabolic process
- metabolic process
- sterol biosynthetic process

Component
- endoplasmic reticulum
- integral to membrane
- membrane
- membrane fraction

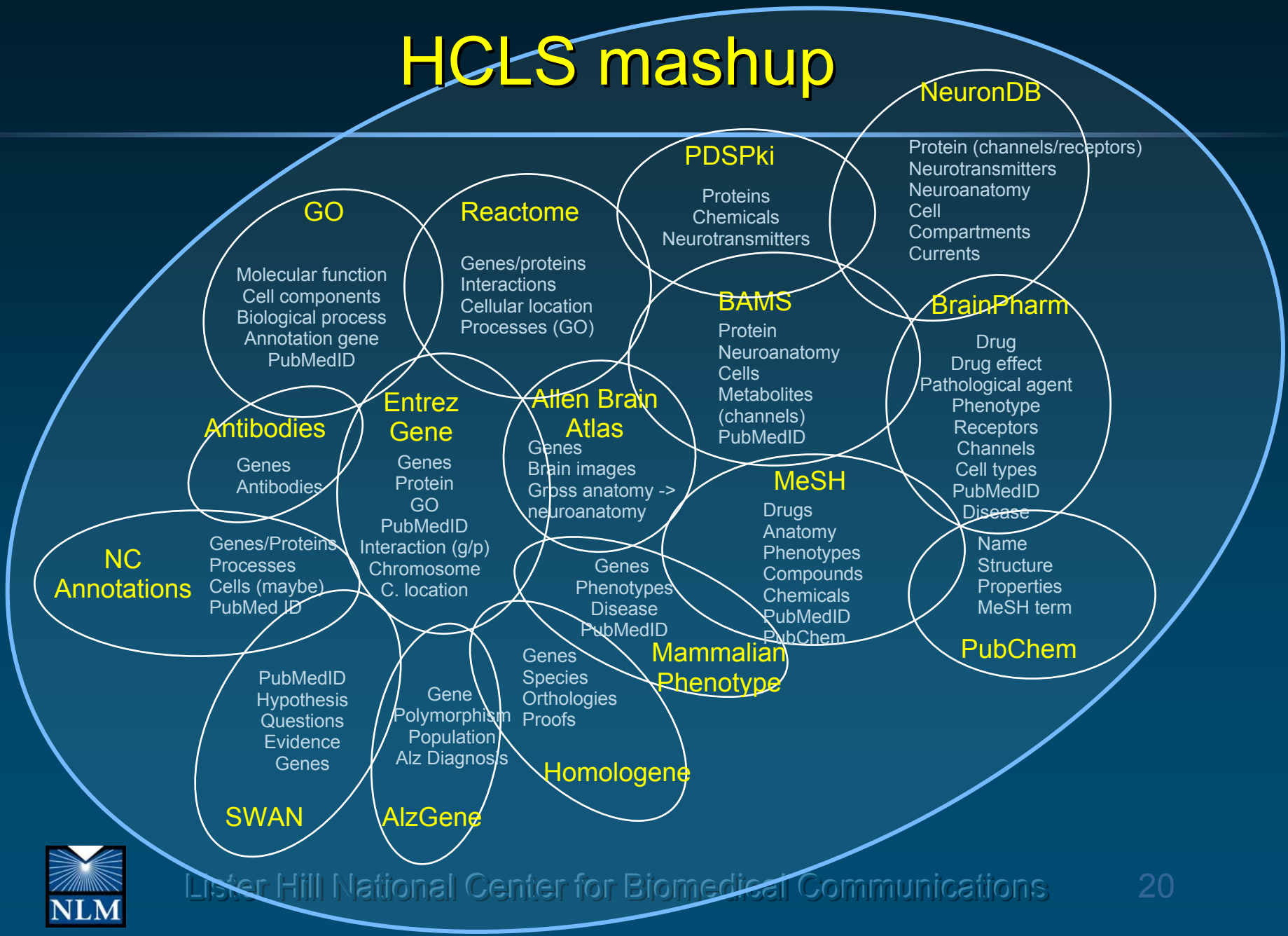
Cholesterol 25-hydroxylase [cytosol]

Details

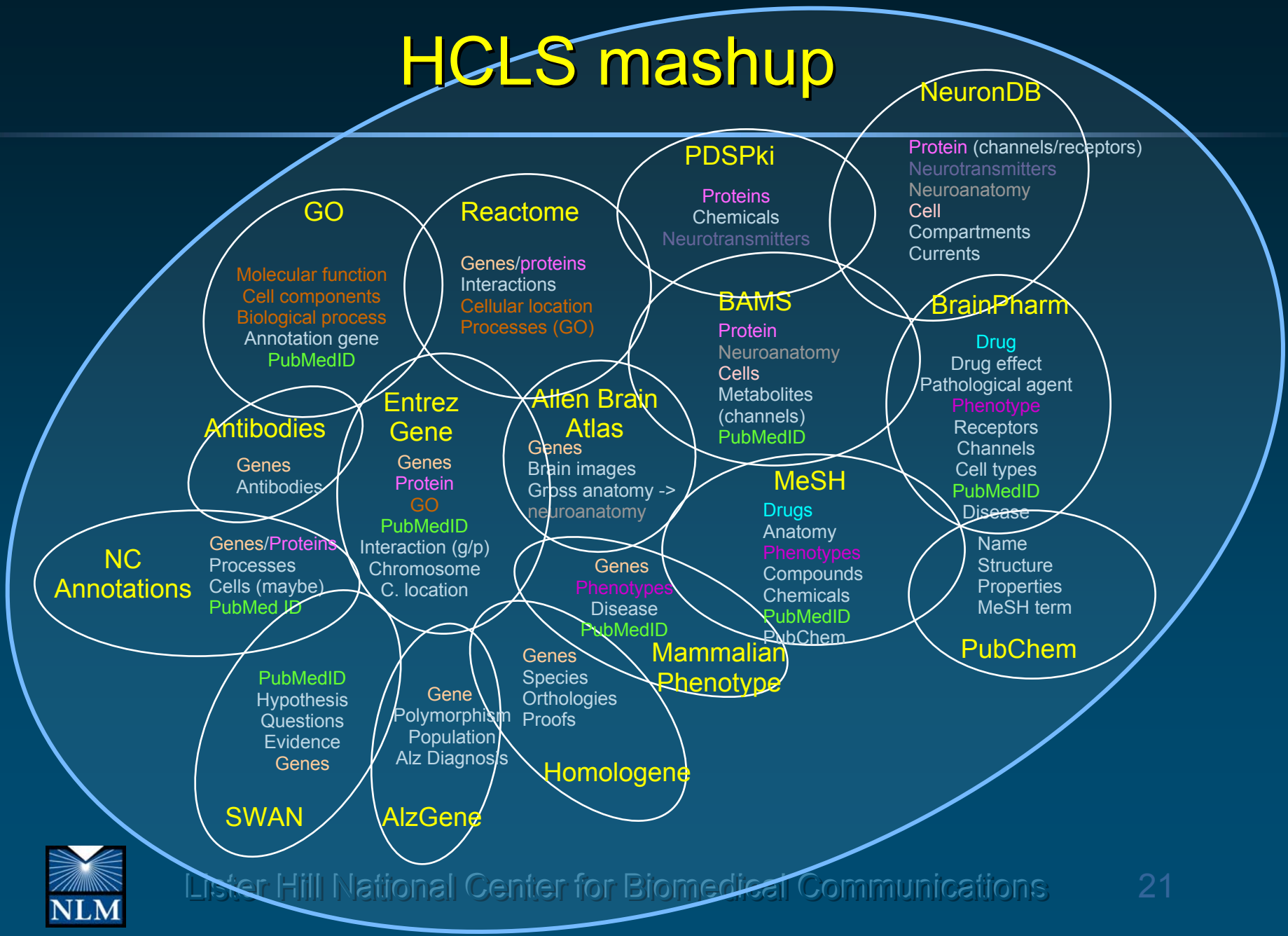
Name	Cholesterol 25-hydroxylase CH25H_HUMAN CH25H
Stable identifier	REACT_10656.1
Links to corresponding entries in other databases	ENSEMBL:ENSG00000138135 Entrez Gene:9023 HapMap:NM_003956 KEGG Gene:9023 MIM:604551 RefSeq:NM_003956 RefSeq:NP_003947 UCSC:O95992 UniProt:O95992
Other identifiers related to this sequence	CH25H_HUMAN, ENSG00000138135, ENST00000371852, ENSP00000360918, ENST00000260706, ENSP00000260706, 206932_at, 323[illegible]_at, 45019_at, g4502498_3p_at, A_14_P139081, A_23_P86470, CCDS7400, GE62104, AF059212, AF059214, AL513533, BC017843, BC072430, EntrezGene:9023, GI_31542304-S, LMN_8057, IPI00022560, MIM:604551, OTTHUMT00000049291, AAC97481, AAC97483, CAI13519, AAH17843, AAH72430, NM_003956, NP_003947, Hs.47357, Hs.597033, O95992, CH25H_HUMAN, IPR006088
Reference entity	UniProt:O95992 Cholesterol 25-hydroxylase
Coordinates in the reference sequence	..
Cellular compartment	cytosol GO
Organism	Homo sapiens
Component of	CH25H (Fe2+ cofactor) [endoplasmic reticulum membrane]
Participates in processes	

Lipid and lipoprotein metabolism
- Steroid metabolism
  - Metabolism of bile acids and bile salts
    - Synthesis of bile acids and bile salts
      - **Cholesterol is hydroxylated to 25-hydroxycholesterol [Homo sapiens]**

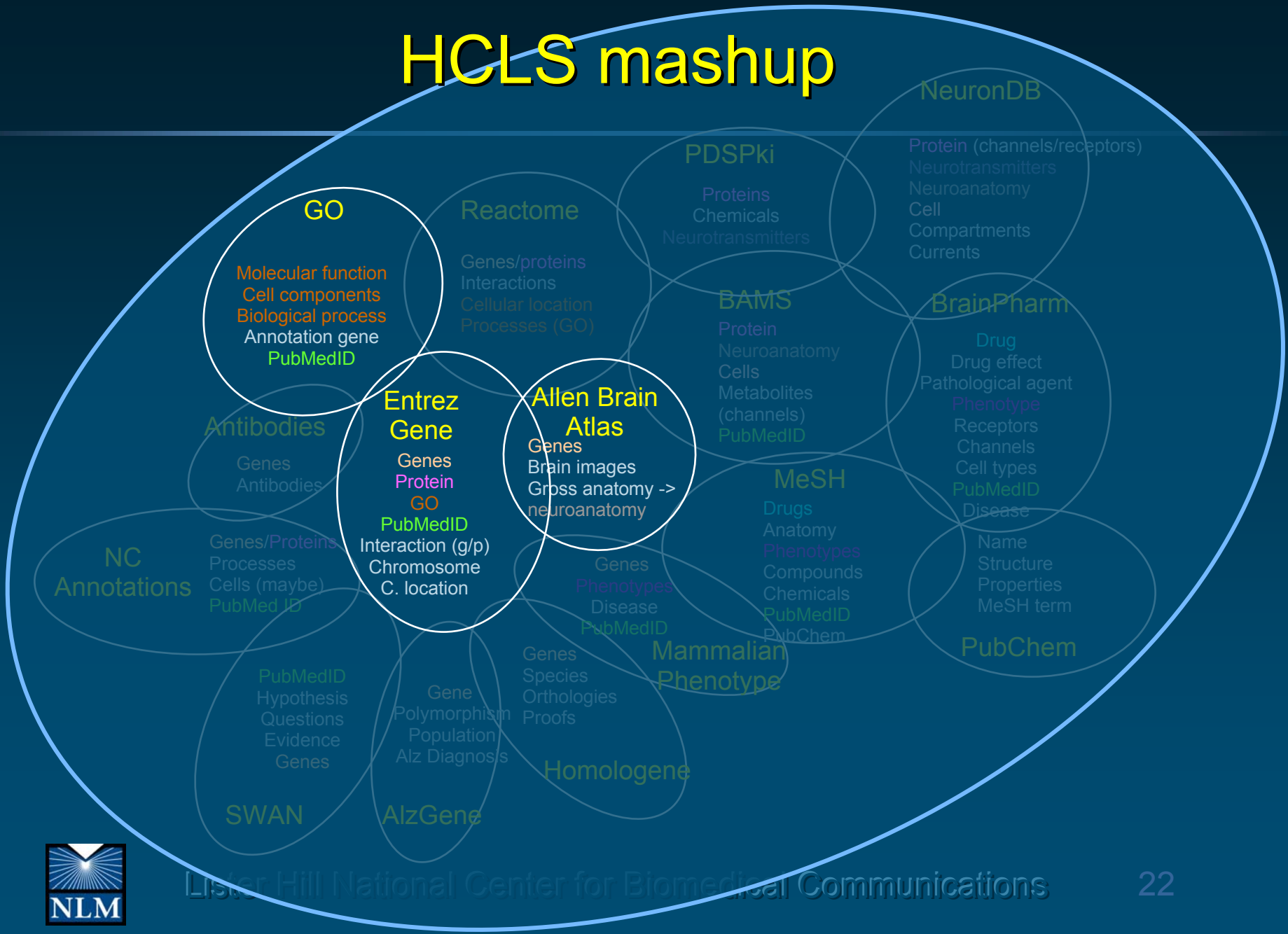
HCLS mashup


NeuronDB
Protein (channels/receptors)
Neurotransmitters
Neuroanatomy
Cell
Compartments
Currents
PDSPki
Proteins
Chemicals
Neurotransmitters
GO
Molecular function
Cell components
Biological process
Annotation gene
PubMedID
Reactome
Genes/proteins
Interactions
Cellular location
Processes (GO)
BAMS
Protein
Neuroanatomy
Cells
Metabolites
(channels)
PubMedID
BrainPharm
Drug
Drug effect
Pathological agent
Phenotype
Receptors
Channels
Cell types
PubMedID
Disease
Antibodies
Genes
Antibodies
Entrez Gene
Genes
Protein
GO
PubMedID
Interaction (g/p)
Chromosome
C. location
Allen Brain Atlas
Genes
Brain images
Gross anatomy -> neuroanatomy
MeSH
Drugs
Anatomy
Phenotypes
Compounds
Chemicals
PubMedID
PubChem
NC Annotations
Genes/Proteins
Processes
Cells (maybe)
PubMed ID
Genes
Phenotypes
Disease
PubMedID
Mammalian Phenotype
Name
Structure
Properties
MeSH term
PubChem
PubMedID
Hypothesis
Questions
Evidence
Genes
SWAN
Gene
Polymorphism
Population
Alz Diagnosis
AlzGene
Genes
Species
Orthologies
Proofs
Homologene

HCLS mashup

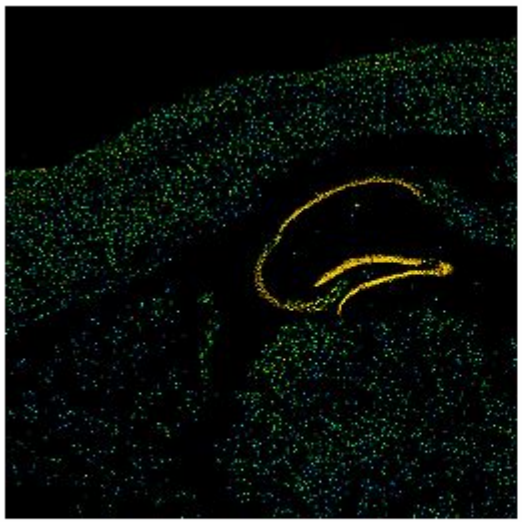

NeuronDB
Protein (channels/receptors)
Neurotransmitters
Neuroanatomy
Cell
Compartments
Currents
PDSPki
Proteins
Chemicals
Neurotransmitters
GO
Molecular function
Cell components
Biological process
Annotation gene
PubMedID
Reactome
Genes/proteins
Interactions
Cellular location
Processes (GO)
BAMS
Protein
Neuroanatomy
Cells
Metabolites
(channels)
PubMedID
BrainPharm
Drug
Drug effect
Pathological agent
Phenotype
Receptors
Channels
Cell types
PubMedID
Disease
Antibodies
Genes
Antibodies
Entrez Gene
Genes
Protein
GO
PubMedID
Interaction (g/p)
Chromosome
C. location
Allen Brain Atlas
Genes
Brain images
Gross anatomy -> neuroanatomy
MeSH
Drugs
Anatomy
Phenotypes
Compounds
Chemicals
PubMedID
PubChem
Name
Structure
Properties
MeSH term
PubChem
NC Annotations
Genes/Proteins
Processes
Cells (maybe)
PubMed ID
Genes
Phenotypes
Disease
PubMedID
Mammalian Phenotype
PubMedID
Hypothesis
Questions
Evidence
Genes
SWAN
Gene
Polymorphism
Population
Alz Diagnosis
AlzGene
Genes
Species
Orthologies
Proofs
Homologene

HCLS mashup


NeuronDB
Protein (channels/receptors)
Neurotransmitters
Neuroanatomy
Cell
Compartments
Currents
PDSPki
Proteins
Chemicals
Neurotransmitters
GO
Molecular function
Cell components
Biological process
Annotation gene
PubMedID
Reactome
Genes/proteins
Interactions
Cellular location
Processes (GO)
BAMS
Protein
Neuroanatomy
Cells
Metabolites
(channels)
PubMedID
BrainPharm
Drug
Drug effect
Pathological agent
Phenotype
Receptors
Channels
Cell types
PubMedID
Disease
Entrez Gene
Genes
Protein
GO
PubMedID
Interaction (g/p)
Chromosome
C. location
Allen Brain Atlas
Genes
Brain images
Gross anatomy -> neuroanatomy
Antibodies
Genes
Antibodies
MeSH
Drugs
Anatomy
Phenotypes
Compounds
Chemicals
PubMedID
PubChem
NC Annotations
Genes/Proteins
Processes
Cells (maybe)
PubMed ID
Genes
Phenotypes
Disease
PubMedID
Mammalian Phenotype
Name
Structure
Properties
MeSH term
PubChem
PubMedID
Hypothesis
Questions
Evidence
Genes
Gene
Polymorphism
Population
Alz Diagnosis
Genes
Species
Orthologies
Proofs
Homologene
SWAN
AlzGene

HCLS mashup

6.



432508

Entrez-Gene 432508

cleavage and polyadenylation specific factor 6

location: nucleus

Transcript Region

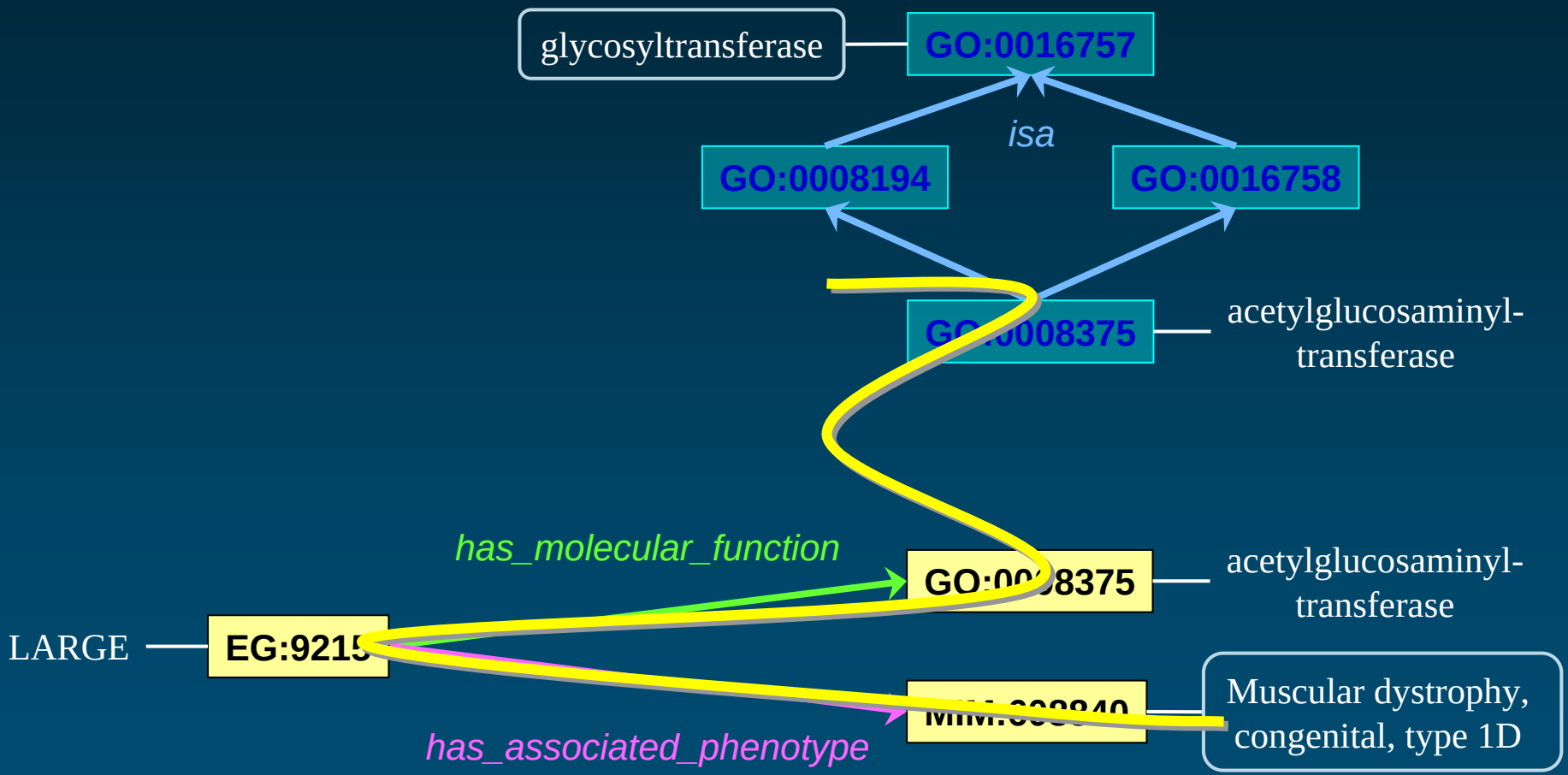
NC_000076.4

Genomic Context

Open Map View

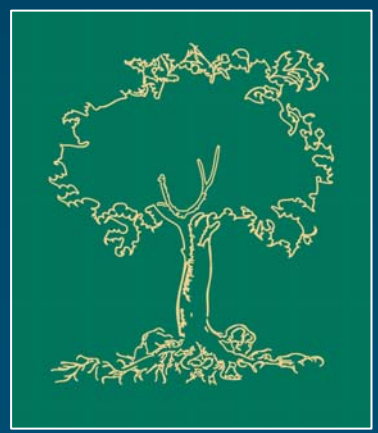
http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo

From *glycosyltransferase* to *congenital muscular dystrophy*



Terminology integration in biomedicine

Unified Medical Language System



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.»

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

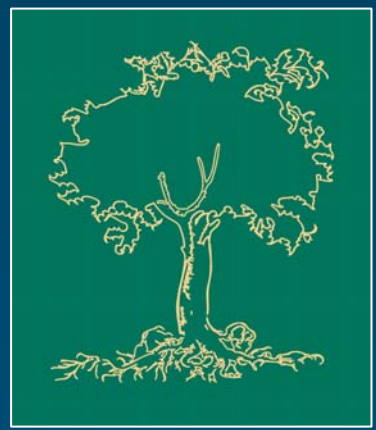
Lexical resources

Terminological resources

Ontological resources

Addison’s disease

Example



Addison's disease in medical vocabularies

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

Organize terms

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

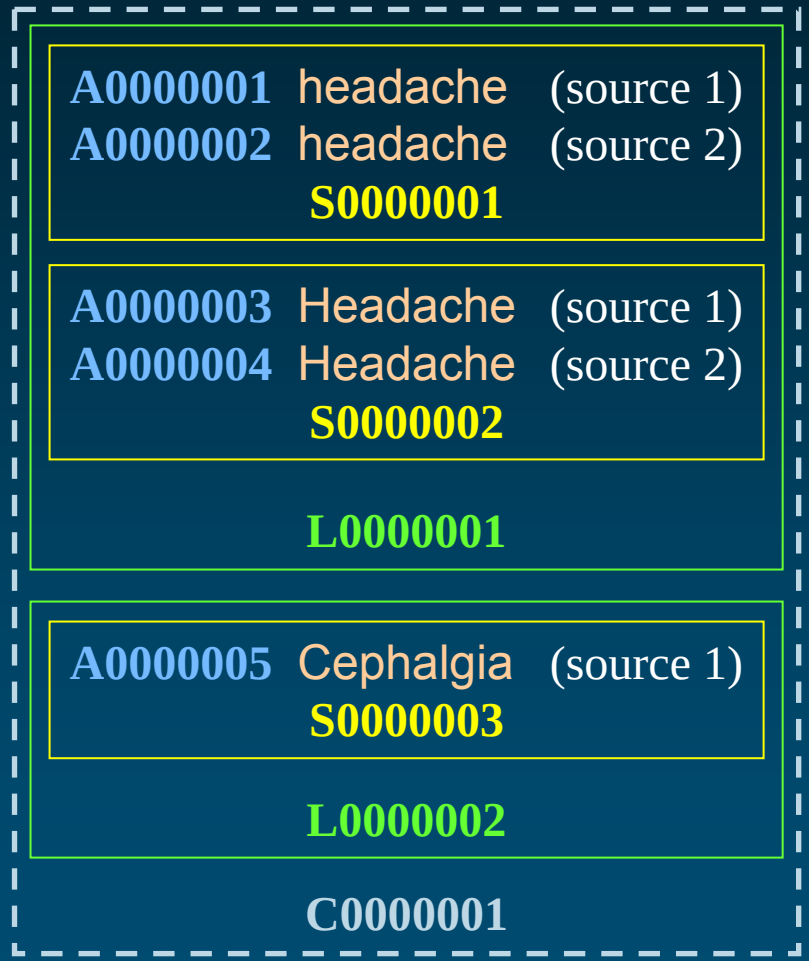
Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

C0001403

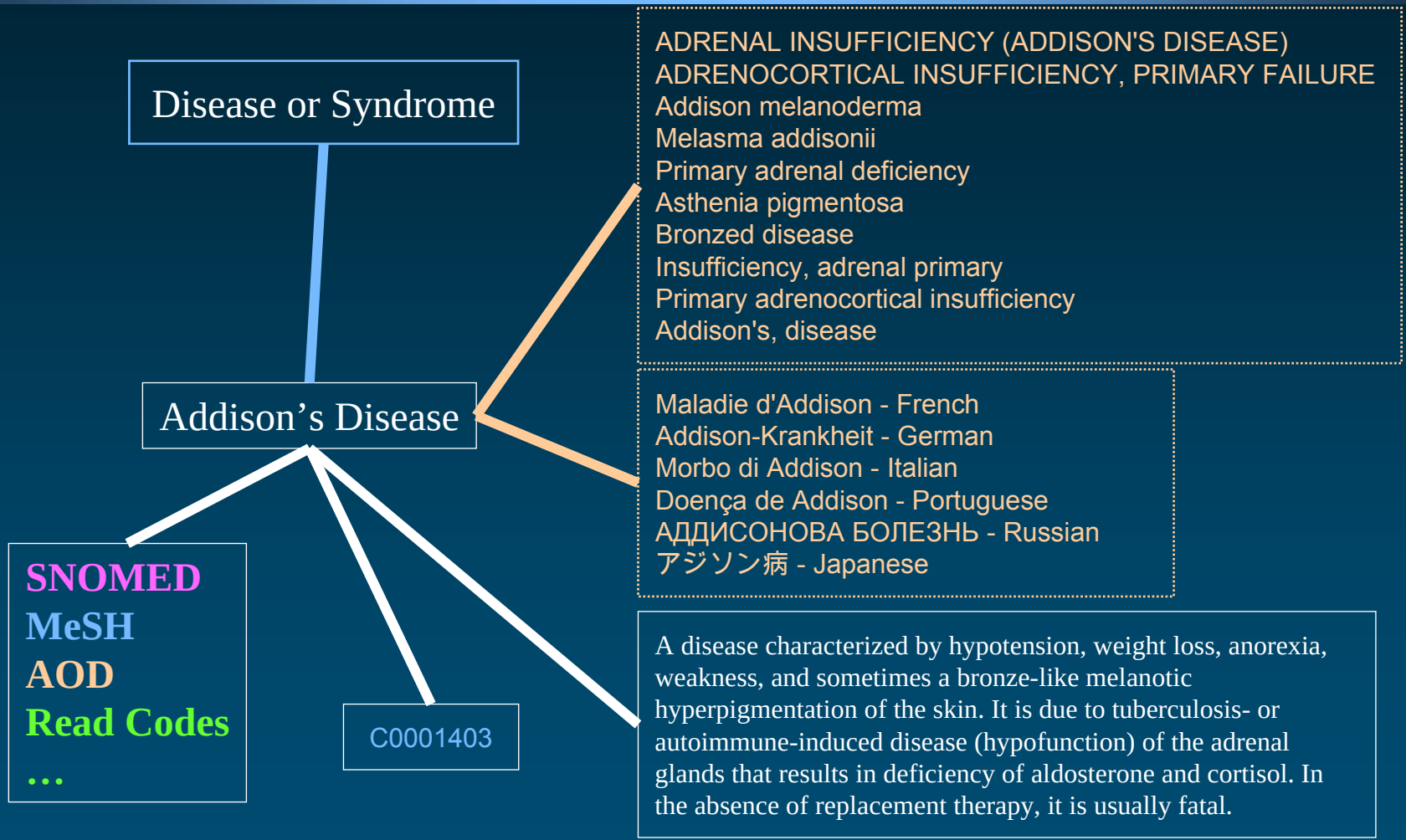
Addison's disease

Metathesaurus Concepts (2007AA)

- Concept (~ 1.4 M) CUI
  - Set of synonymous concept names
- Term (~ 4.9 M) LUI
  - Set of normalized names
- String (~ 5.5 M) SUI
  - Distinct concept name
- Atom (~ 6.8 M) AUI
  - Concept name in a given source

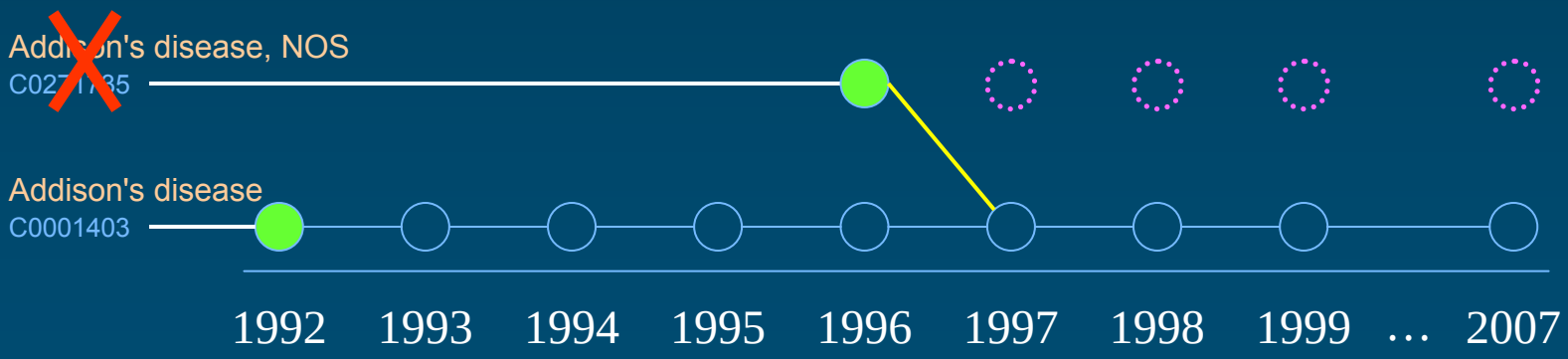


Addison's Disease: Concept

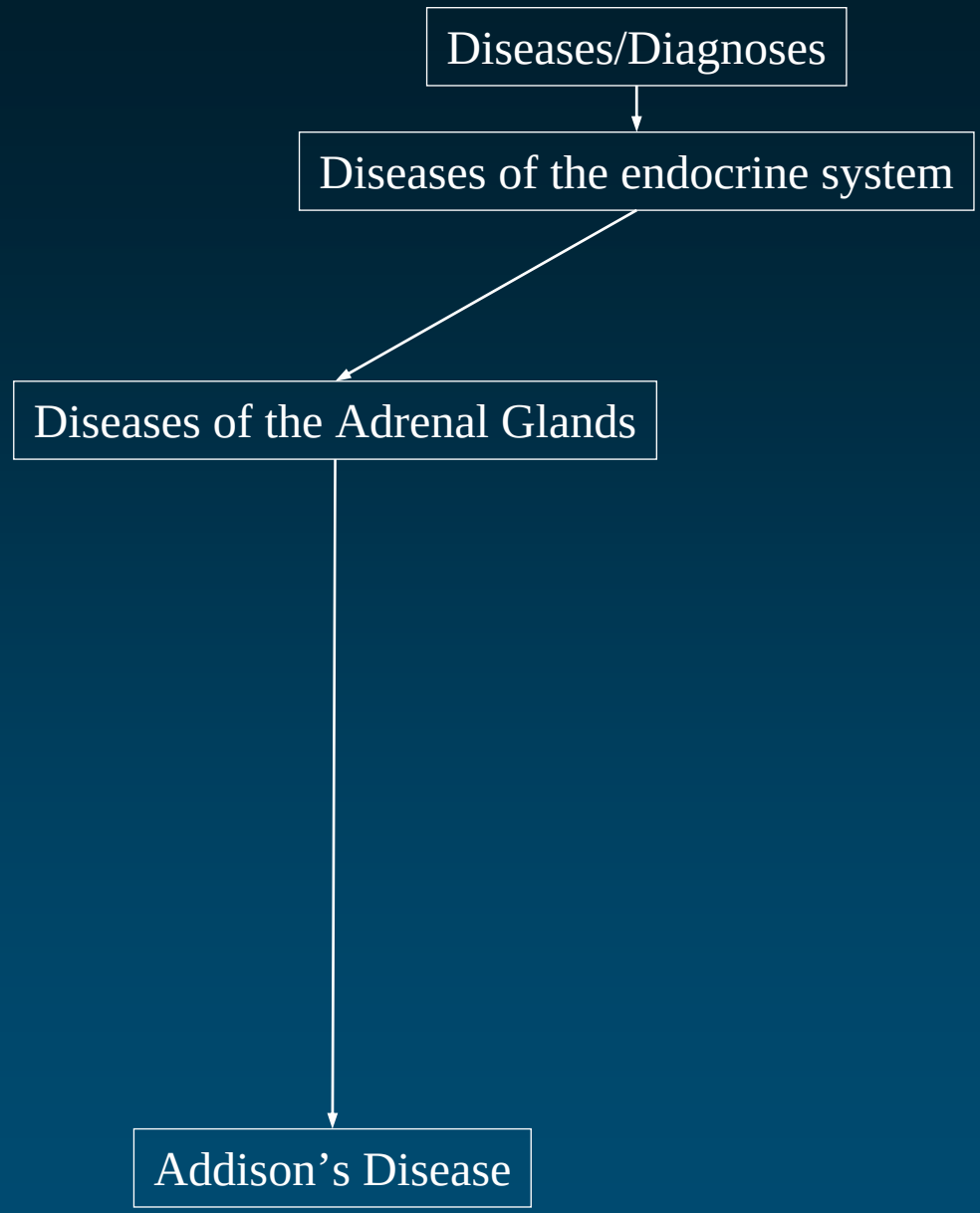


Metathesaurus Evolution over time

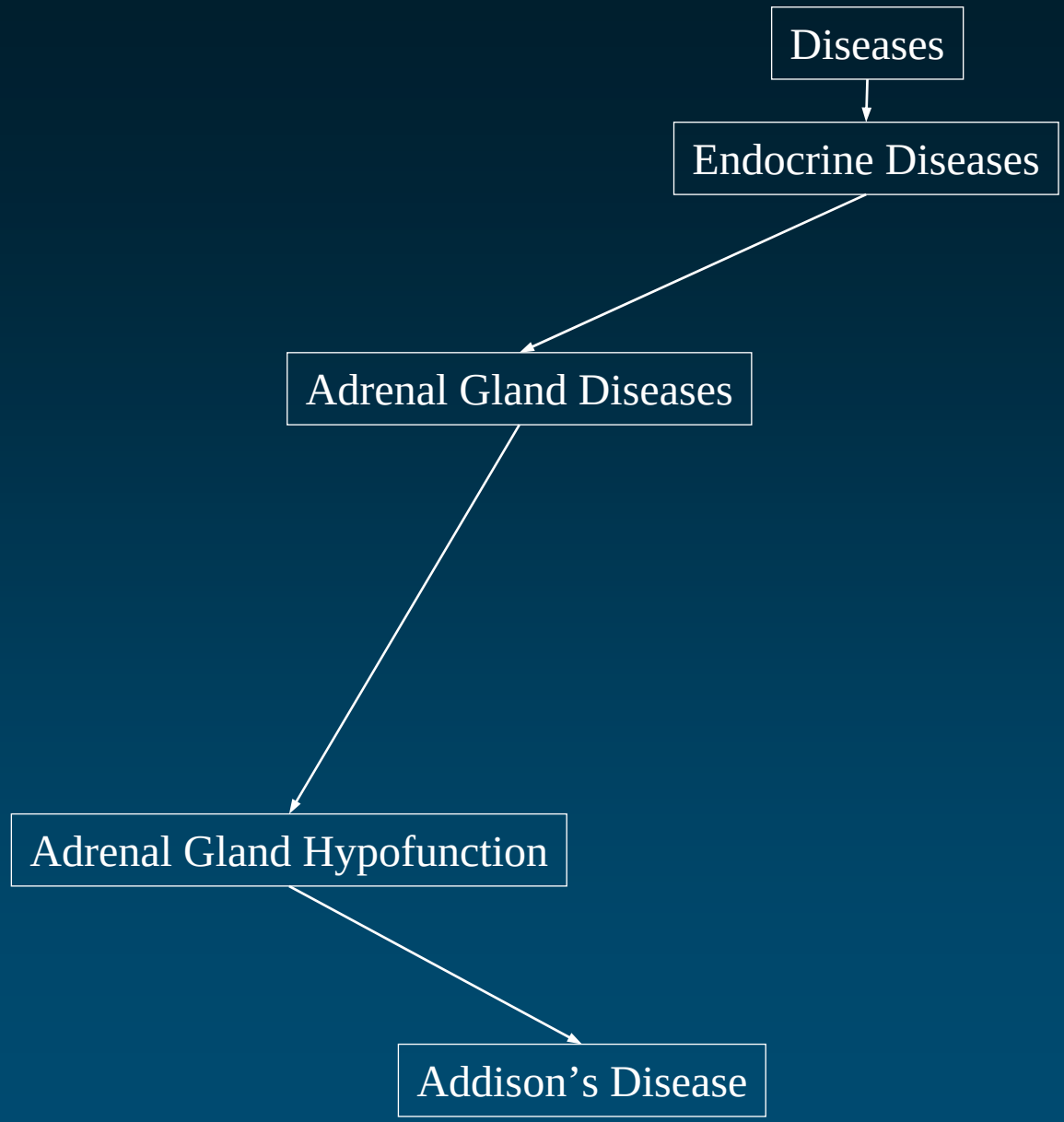
- Concepts never die (in principle)
  - CUIs are permanent identifiers
- What happens when they do die (in reality)?
  - Concepts can merge or split
  - Resulting in new concepts and deletions



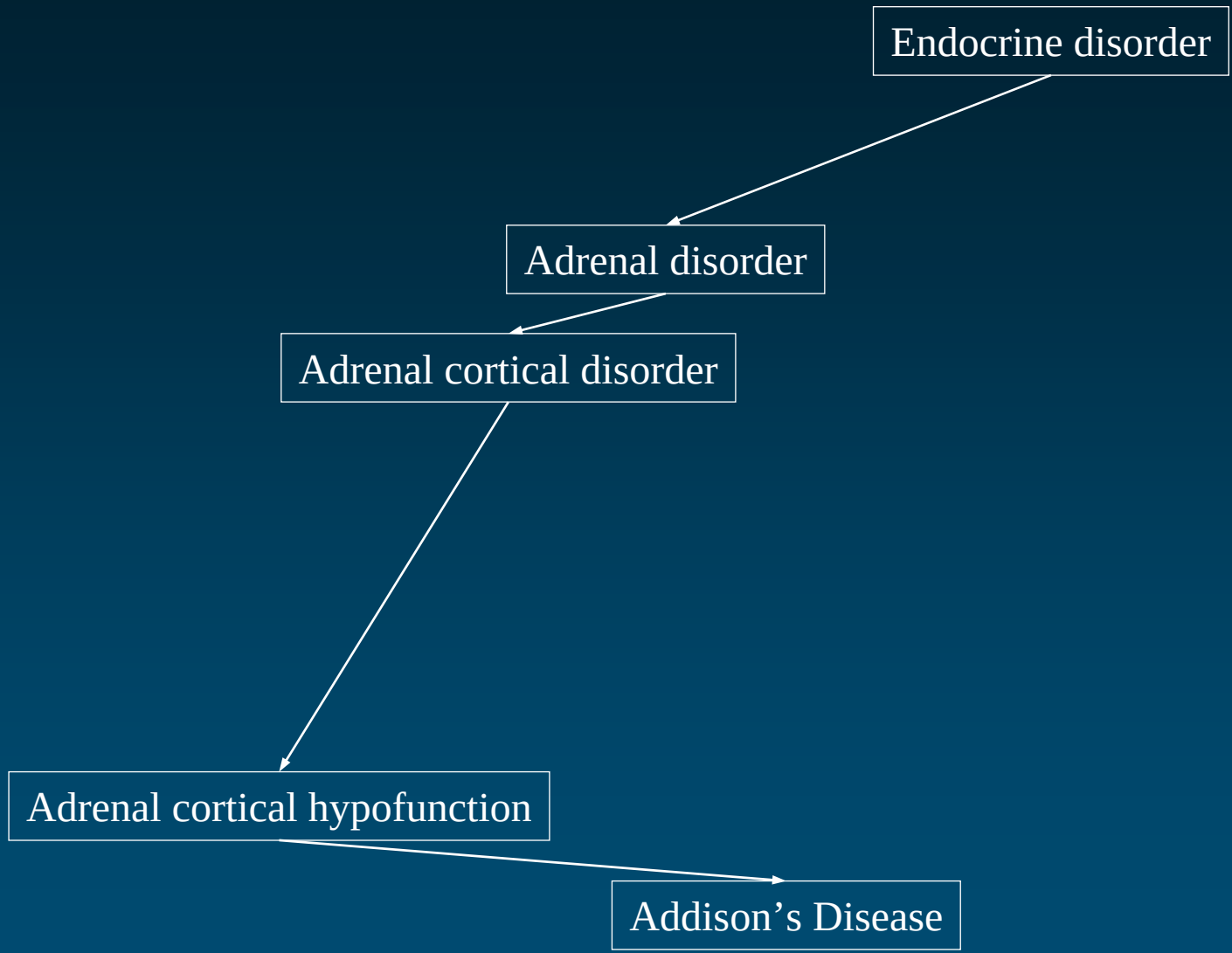
SNOMED International



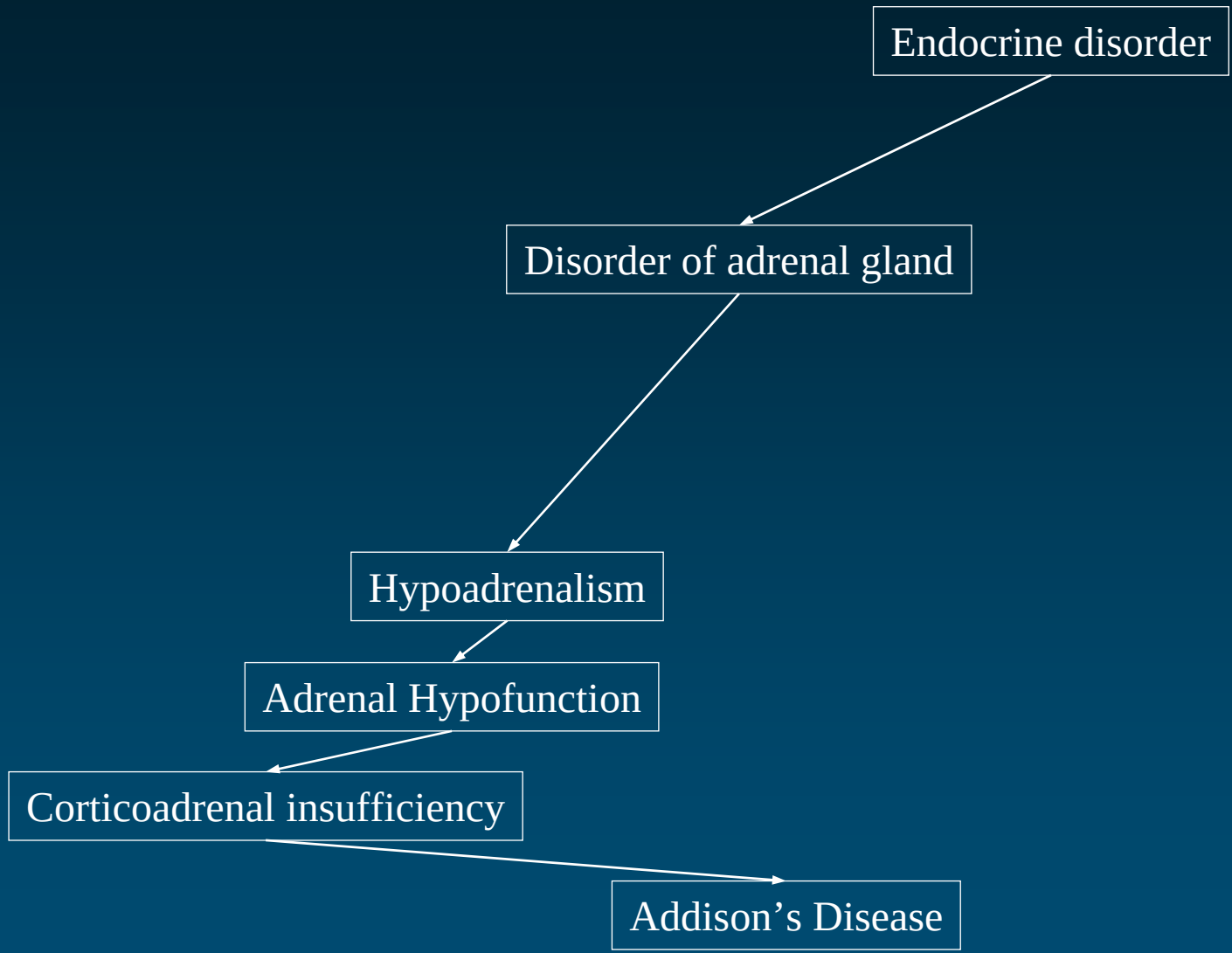
MeSH


Diseases
Endocrine Diseases
Adrenal Gland Diseases
Adrenal Gland Hypofunction
Addison’s Disease

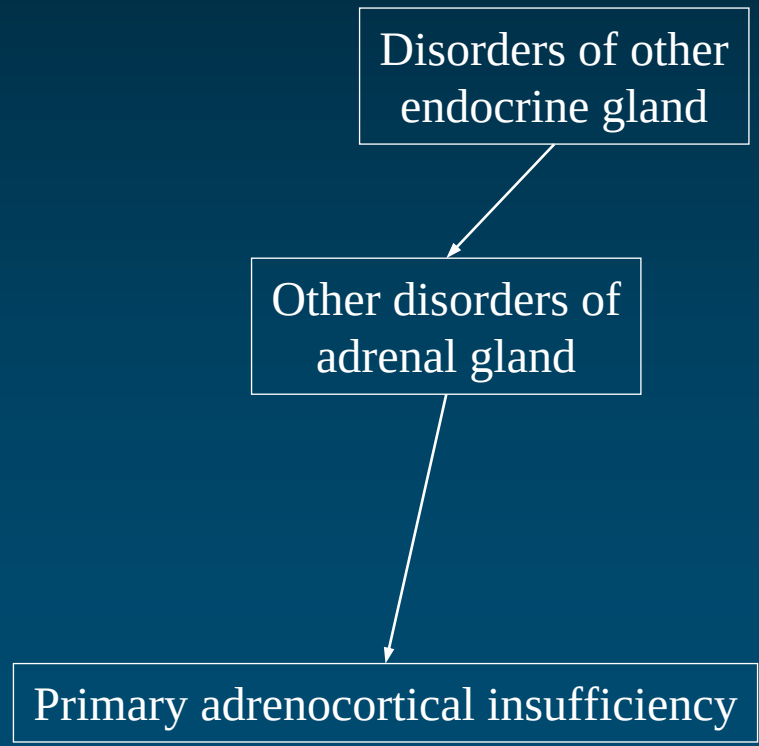
AOD


Endocrine disorder
Adrenal disorder
Adrenal cortical disorder
Adrenal cortical hypofunction
Addison’s Disease

Read Codes

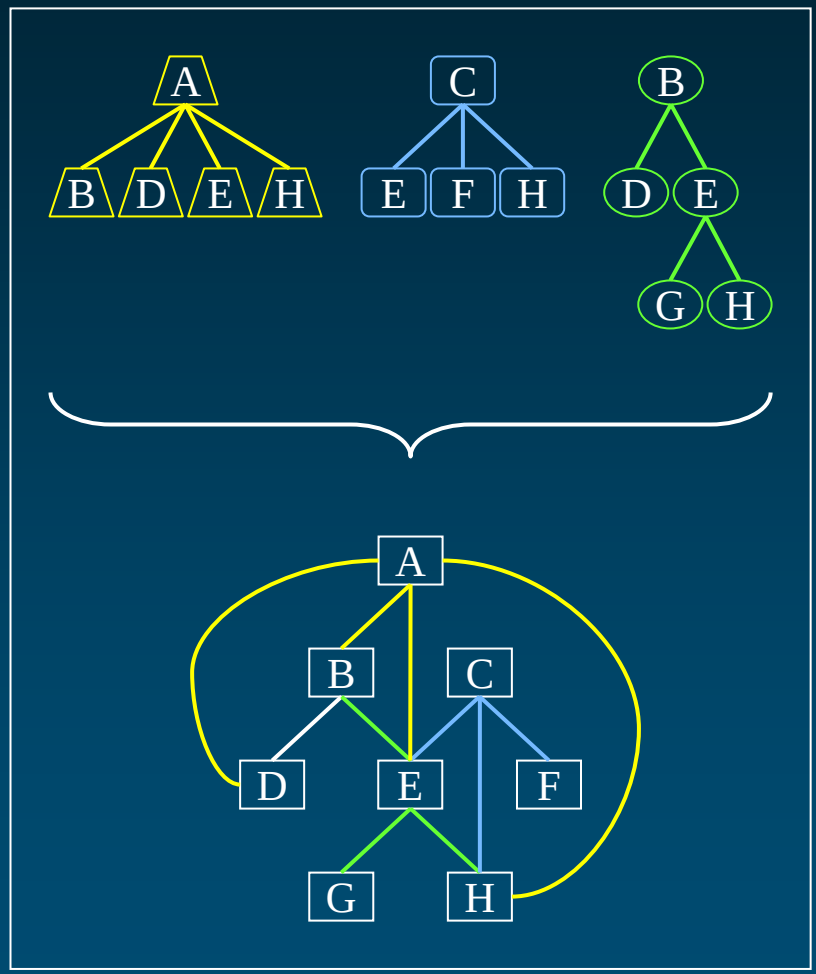

Endocrine disorder
Disorder of adrenal gland
Hypoadrenalism
Adrenal Hypofunction
Corticoadrenal insufficiency
Addison’s Disease

ICD-10


Disorders of other endocrine gland
Other disorders of adrenal gland
Primary adrenocortical insufficiency

Organize concepts

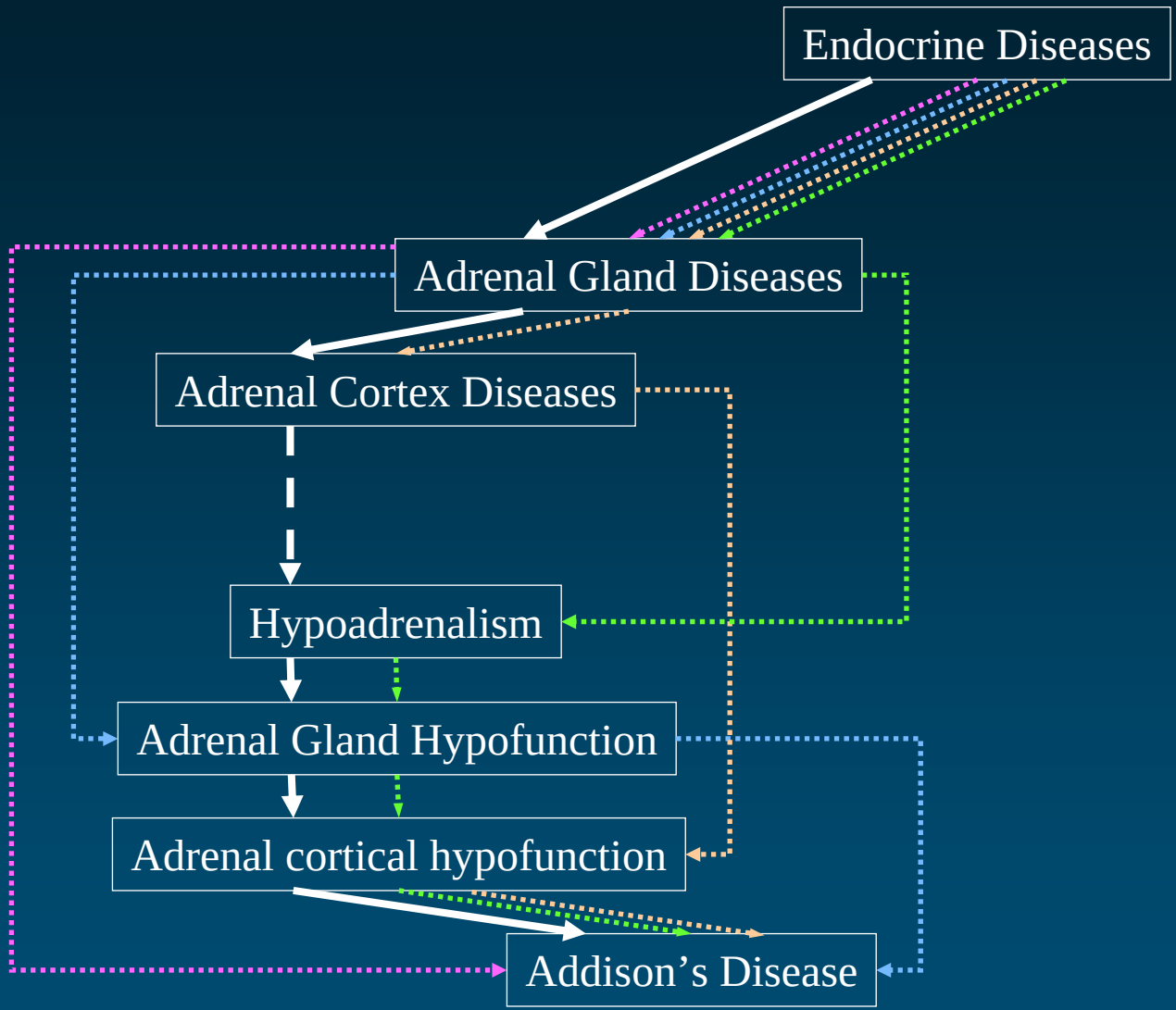
- Inter-concept relationships: hierarchies from the source vocabularies
- Redundancy: multiple paths
- One graph instead of multiple trees (multiple inheritance)

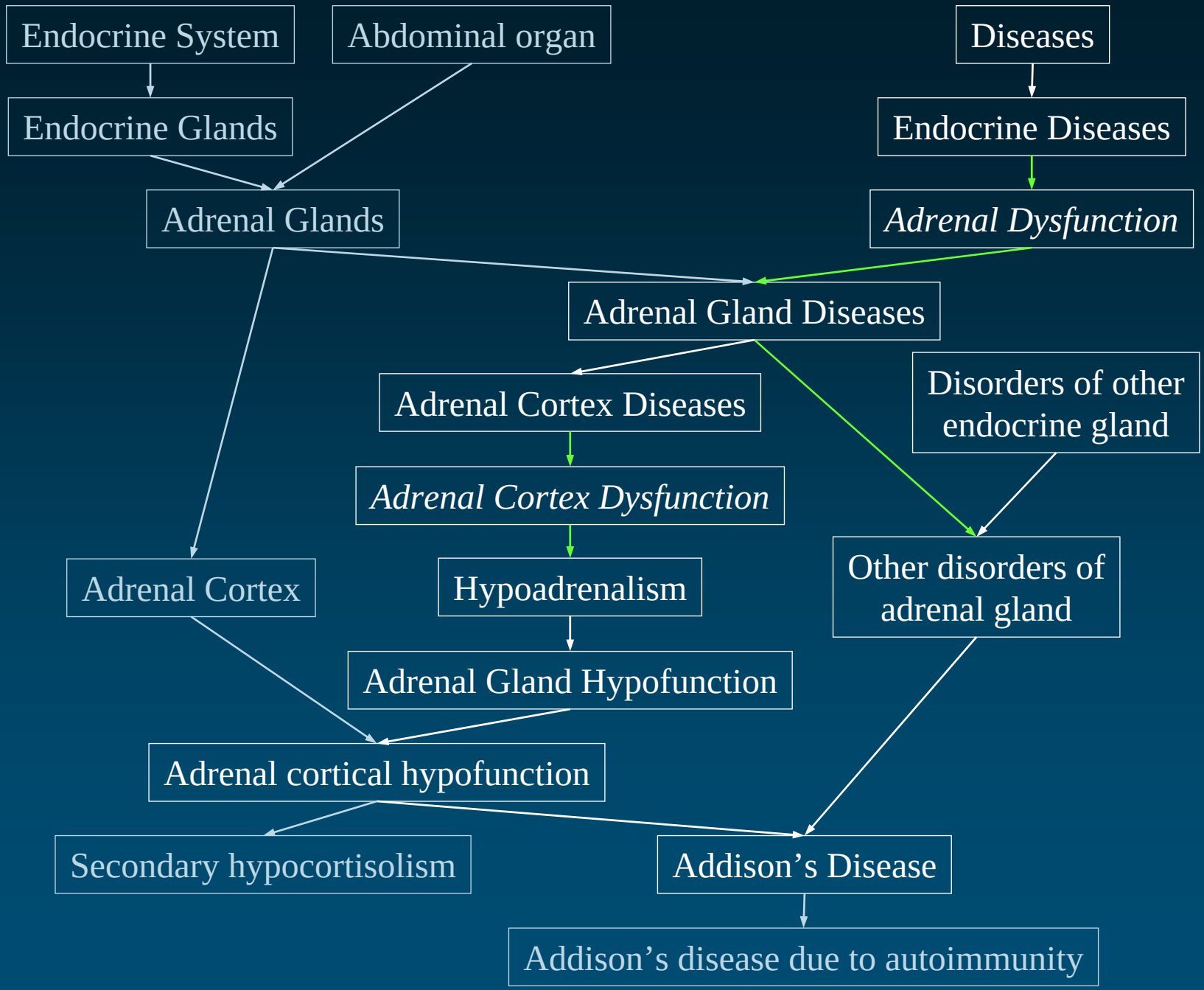


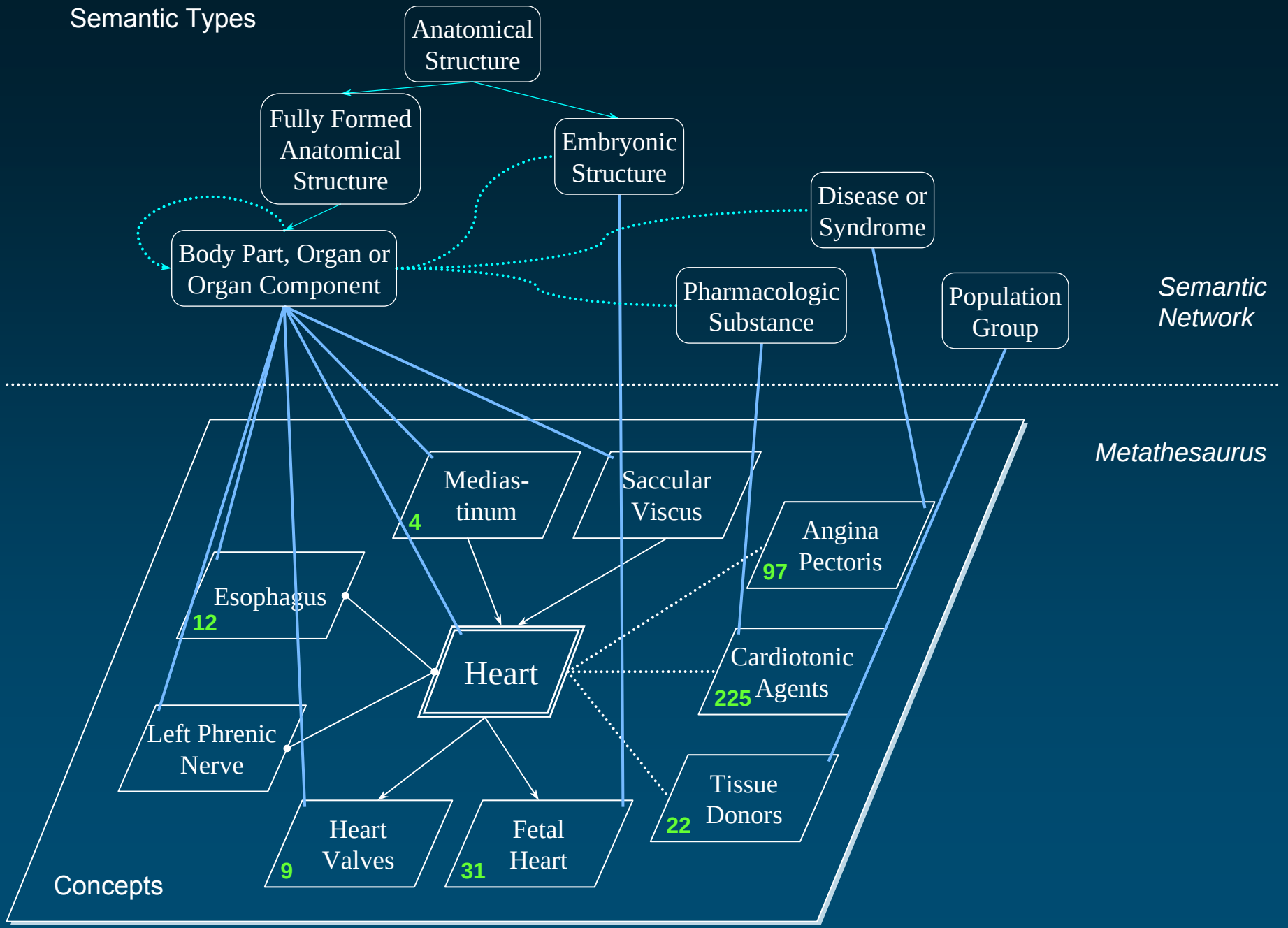
organize concepts

SNOMED
MeSH
AOD
Read Codes

UMLS




Endocrine System
Abdominal organ
Diseases
Endocrine Glands
Endocrine Diseases
Adrenal Glands
Adrenal Dysfunction
Adrenal Gland Diseases
Adrenal Cortex Diseases
Disorders of other endocrine gland
Adrenal Cortex Dysfunction
Adrenal Cortex
Hypoadrenalism
Other disorders of adrenal gland
Adrenal Gland Hypofunction
Adrenal cortical hypofunction
Secondary hypocortisolism
Addison's Disease
Addison's disease due to autoimmunity


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
Medias-tinum
4
Saccular Viscus
Angina Pectoris
97
Esophagus
12
Heart
Cardiotonic Agents
225
Left Phrenic Nerve
Tissue Donors
22
Heart Valves
9
Fetal Heart
31
Concepts

Source Vocabularies (2007AA)

- 139 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 5.5M names
  - 1.4M concepts
  - 16M relations
- Common presentation

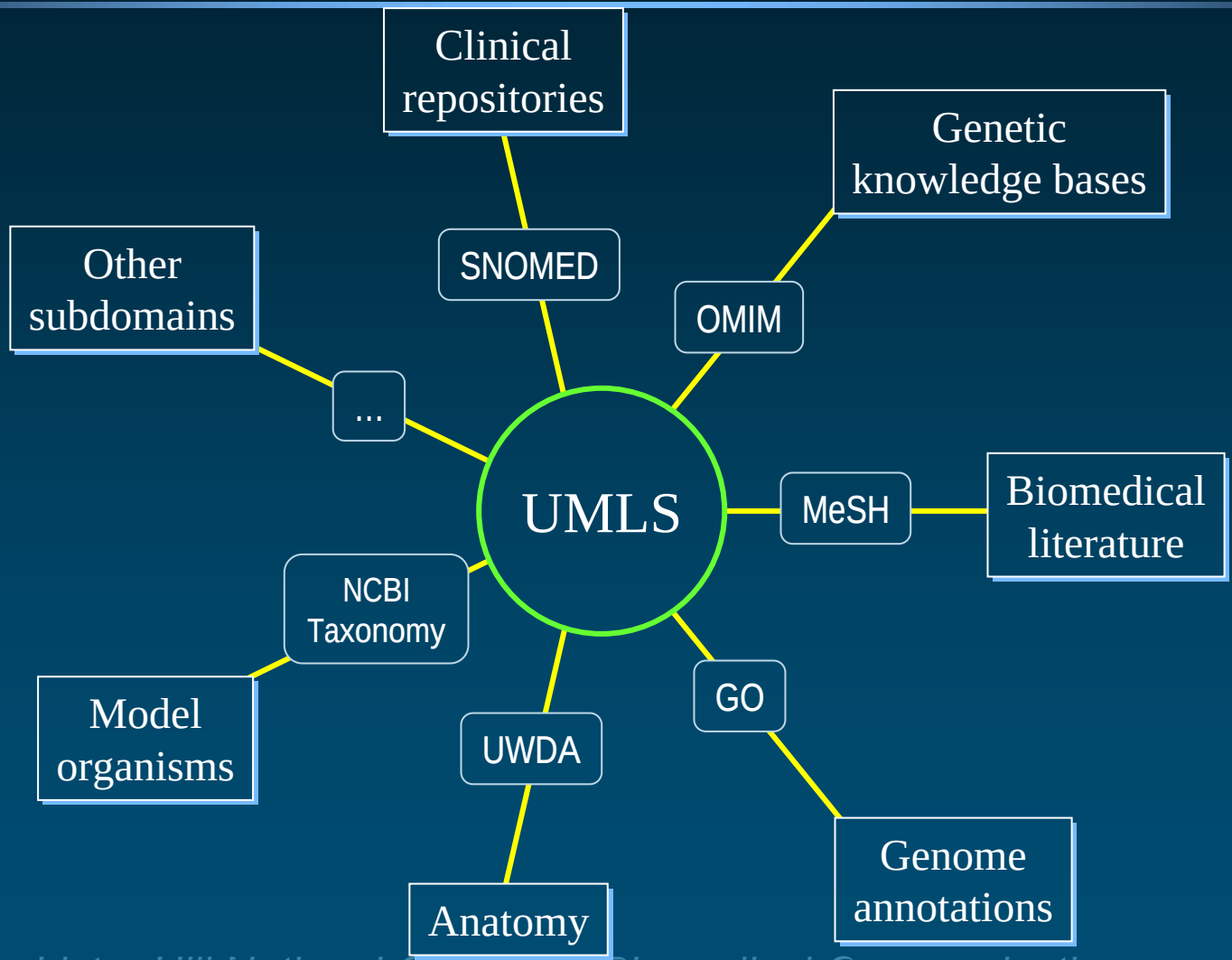
Biomedical terminologies

- 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)

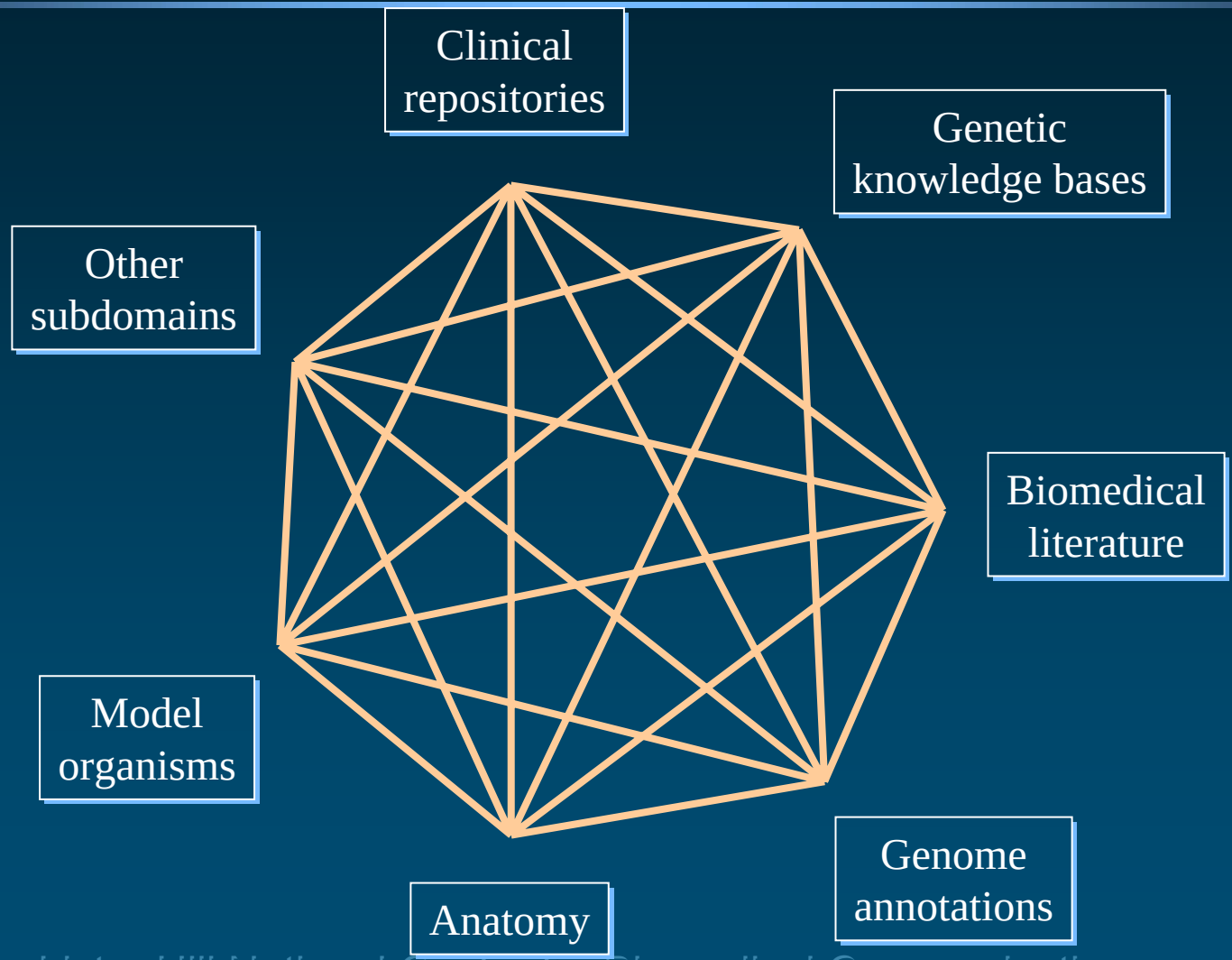
Biomedical terminologies (cont'd)

- Specialized vocabularies
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - oncology (NCI Thesaurus, PDQ)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART, MedDRA)
  - primary care (ICPC)
  - genomics (Gene Ontology, HUGO, OMIM)
- Terminology of knowledge bases (AI/Rheum, DXplain, QMR)

Integrating subdomains



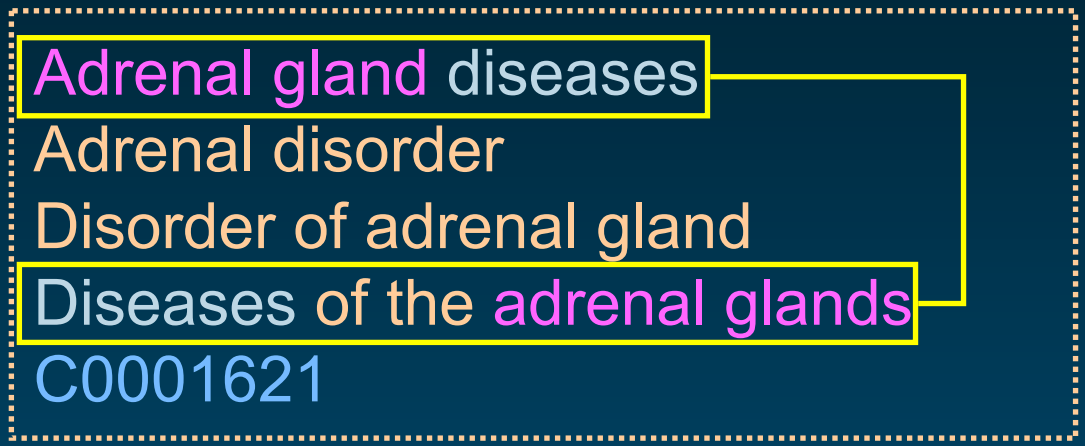
Integrating subdomains



How do they do that?

- Lexical knowledge
- Semantic pre-processing
- UMLS editors

Lexical knowledge

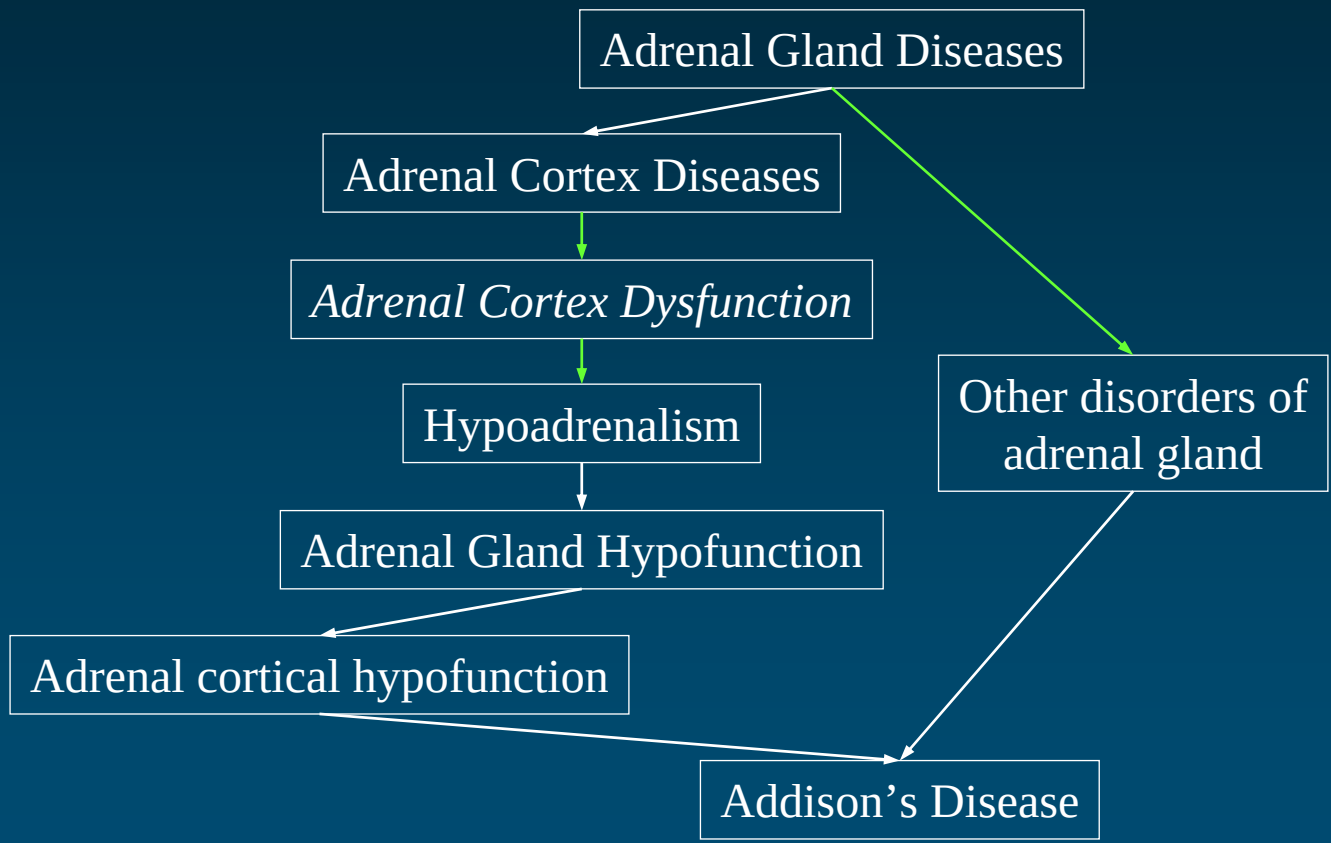


Semantic pre-processing

- Metadata in the source vocabularies

- Tentative categorization
- Positive (or negative) evidence for tentative synonymy relations based on lexical features

Additional knowledge: UMLS editors



UMLS vs. Semantic Web

Similarities, differences and unresolved issues

- Identifying biomedical entities
  - Trans-namespace integration
  - No UMLS-based URIs
- Availability
  - Intellectual property restrictions
  - Application Programming Interface
- Formats
  - RRF vs. SW languages
- UMLS as an ontology?
  - Underspecified semantics

❶ Identifying biomedical entities

- Syntax vs. semantics
  - URI, LSID,… vs. reference ontologies
- Integrative resources vs. individual namespaces
  - Unified Medical Language System (UMLS) vs. GO, MeSH, SNOMED, …

No UMLS-based URIs Syntax

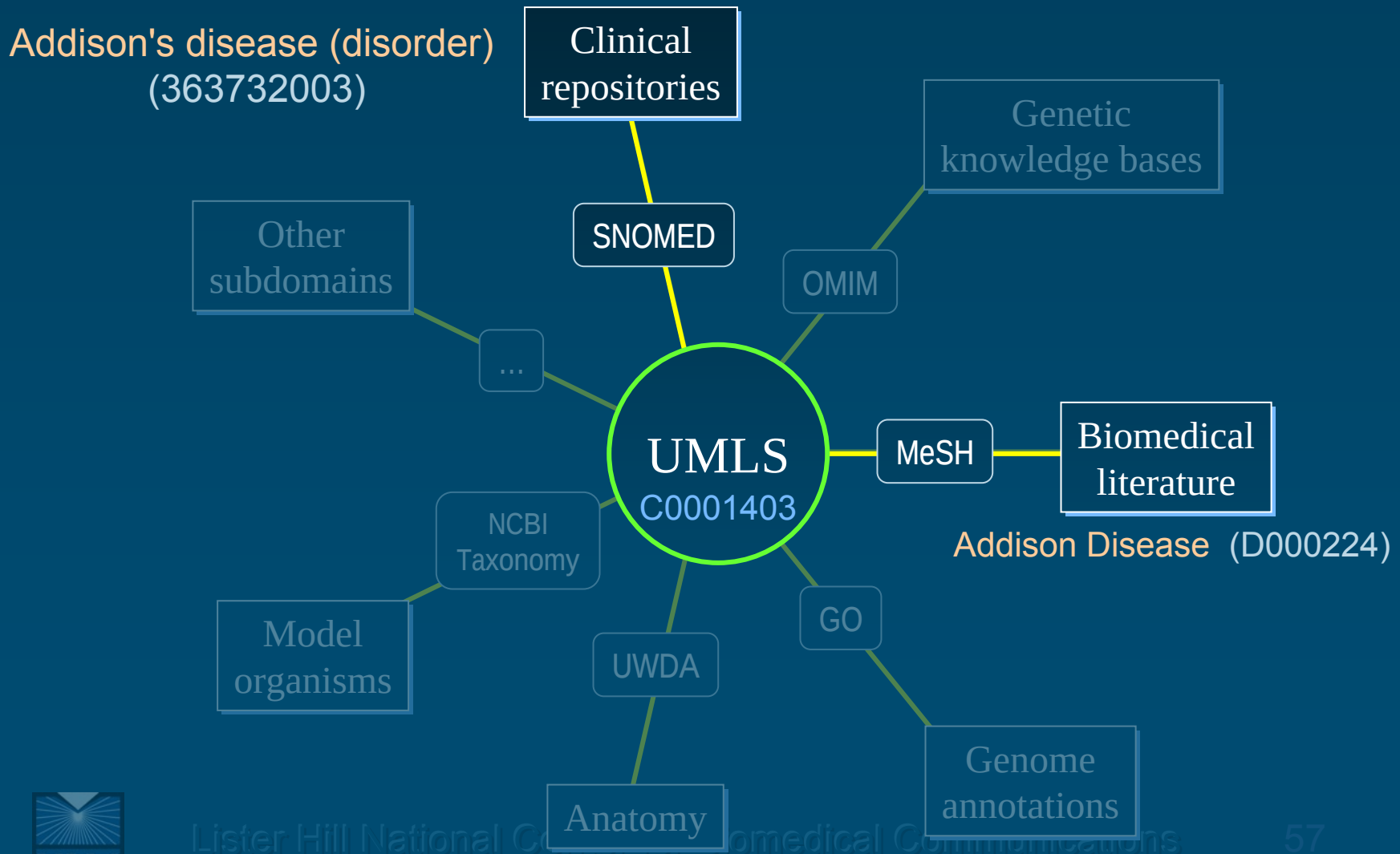
- No officially supported UMLS-based URIs for biomedical entities

  e.g., http://umls.org/C0001403
- Possible alternatives
  - Redirection service (e.g., PURL) http://purl.org/
- Resolution issues: what is expected to be returned?
  - Acknowledgment of existence
  - Preferred term
  - Set of names, relations,… in RDF

No UMLS-based URIs Semantics

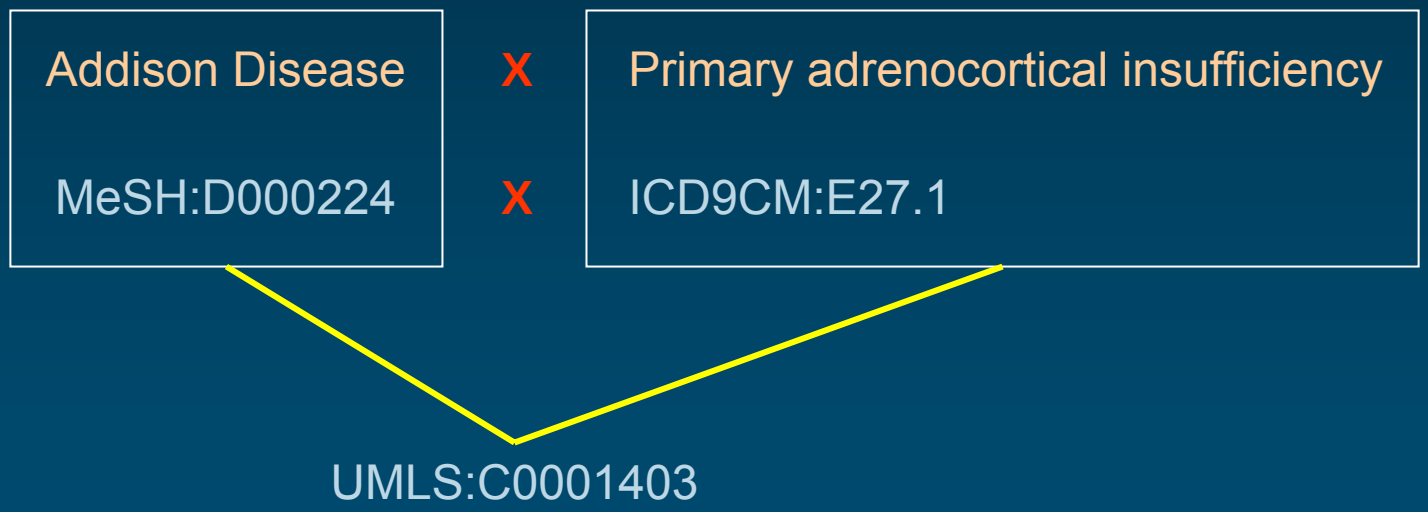
- Potential resources for trans-namespace identification of biomedical entities
  - Clinical medicine: UMLS CUIs
  - [Genomics: Entrez Gene]
- Ontology of biomedical relationships
  - No comprehensive integrative resource available
    - OBO relations
    - UMLS Semantic Network relations
    - GALEN relations

Trans-namespace integration

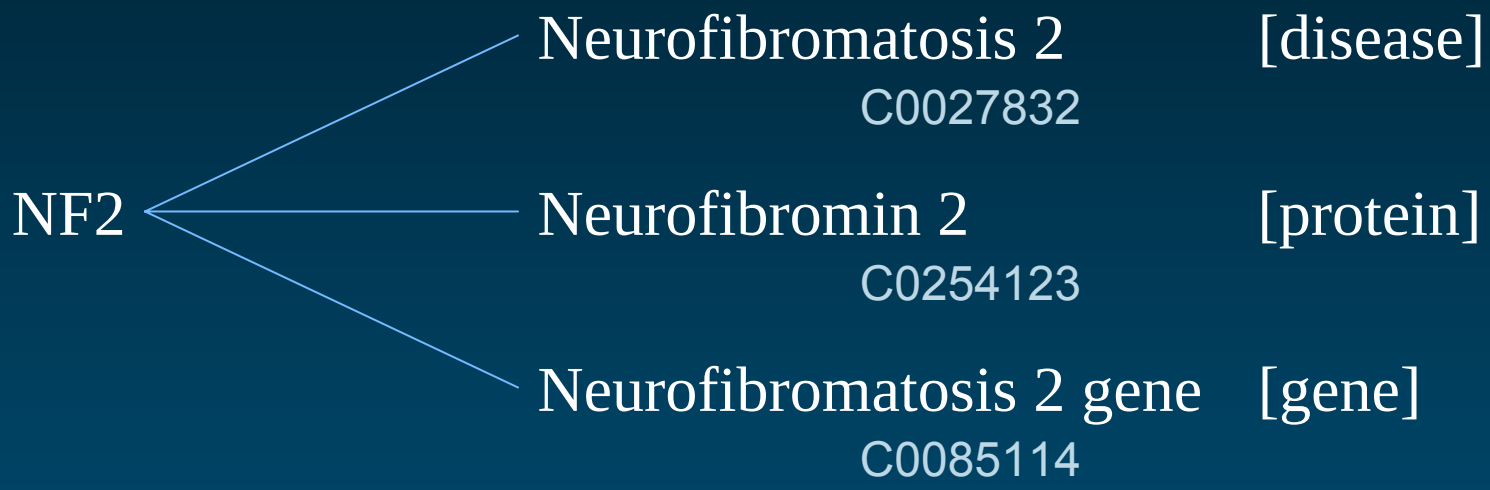


Trans-namespace integration

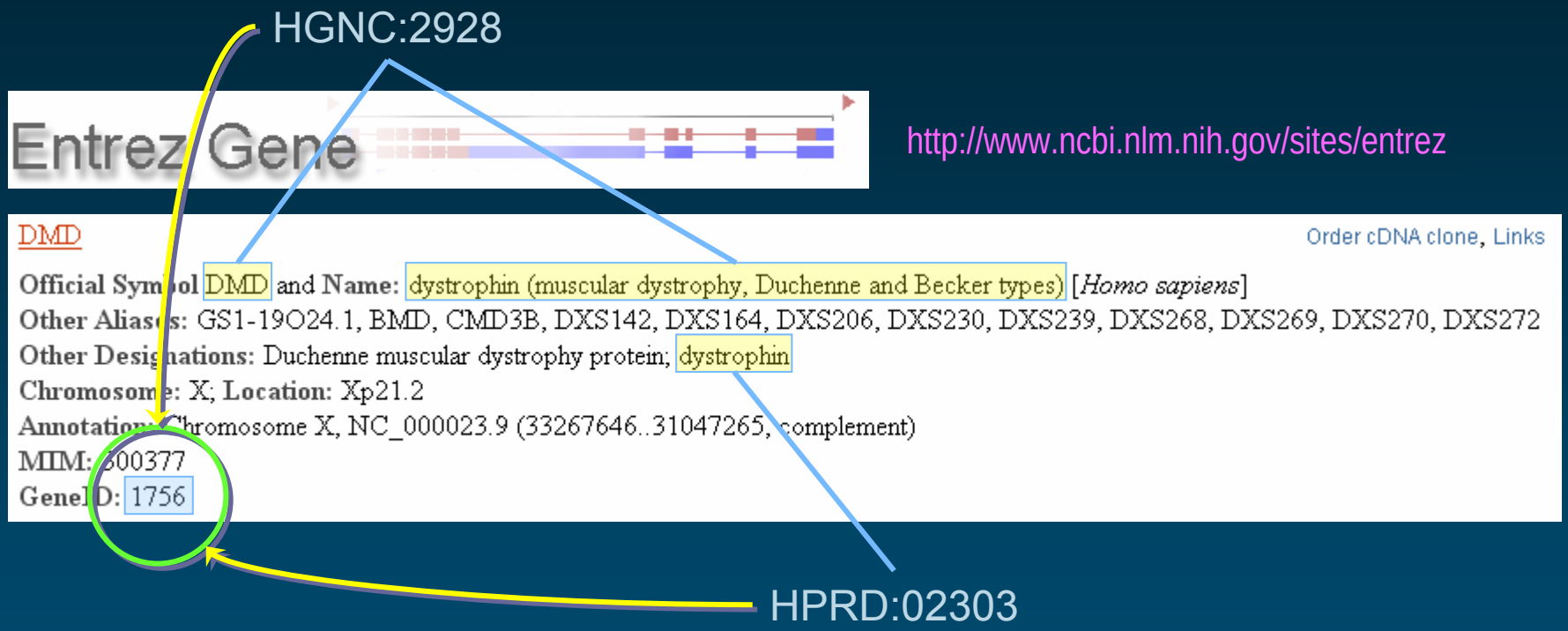
- Advantages
  - Over shared identifiers (increased recall)
  - Over lexical mapping (increased recall + precision)



Ambiguity resolution



Other integrative resources



❷ Availability Intellectual property restrictions

- UMLS: free license required
  http://www.nlm.nih.gov/research/umls/license.html
- Some intellectual property restrictions
  - 2/3 of the names freely available (in the US)

Name Count by Source Restriction Level (SRL):

SRL	Source Count	% of Sources
0	2181959	32.22%
1	94059	1.39%
2	22156	0.33%
3	2111546	31.18%
4	2362949	34.89%
0+4	4544908	67.11%

http://www.nlm.nih.gov/research/umls/

- Web browser: username/password required

Availability Application Programming Interfaces

- Remote server at NLM
- Local application connected through

Java RMI

- Java-based applications
- Developer's Guide: Chapter 3
- Set of Java classes (part of the UMLSKS API download)
- Detailed *Javadoc* documentation online and with API download

TCP/IP socket

- XML-based queries
- Developer's Guide: Chapter 5
- XML schema
- Socket server
  - Host: umlsks.nlm.nih.gov
  - Port: 8042

Availability Web Services-based API

- Part of the Knowledge Source Server version 3
  - Portlet-based, customizable
  - WS architecture
- Coming soon
  - Alpha release in July 2007
  - Beta release in November 2007

❸ Representation formalism

- UMLS
  - Rich Release Format (RRF)
  - [Original Release Format (ORF)]
  - Support for source transparency
- Semantic Web
  - RDF – Resource Description Framework
  - OWL – Web Ontology Language
  - SKOS – Simple Knowledge Organization Systems

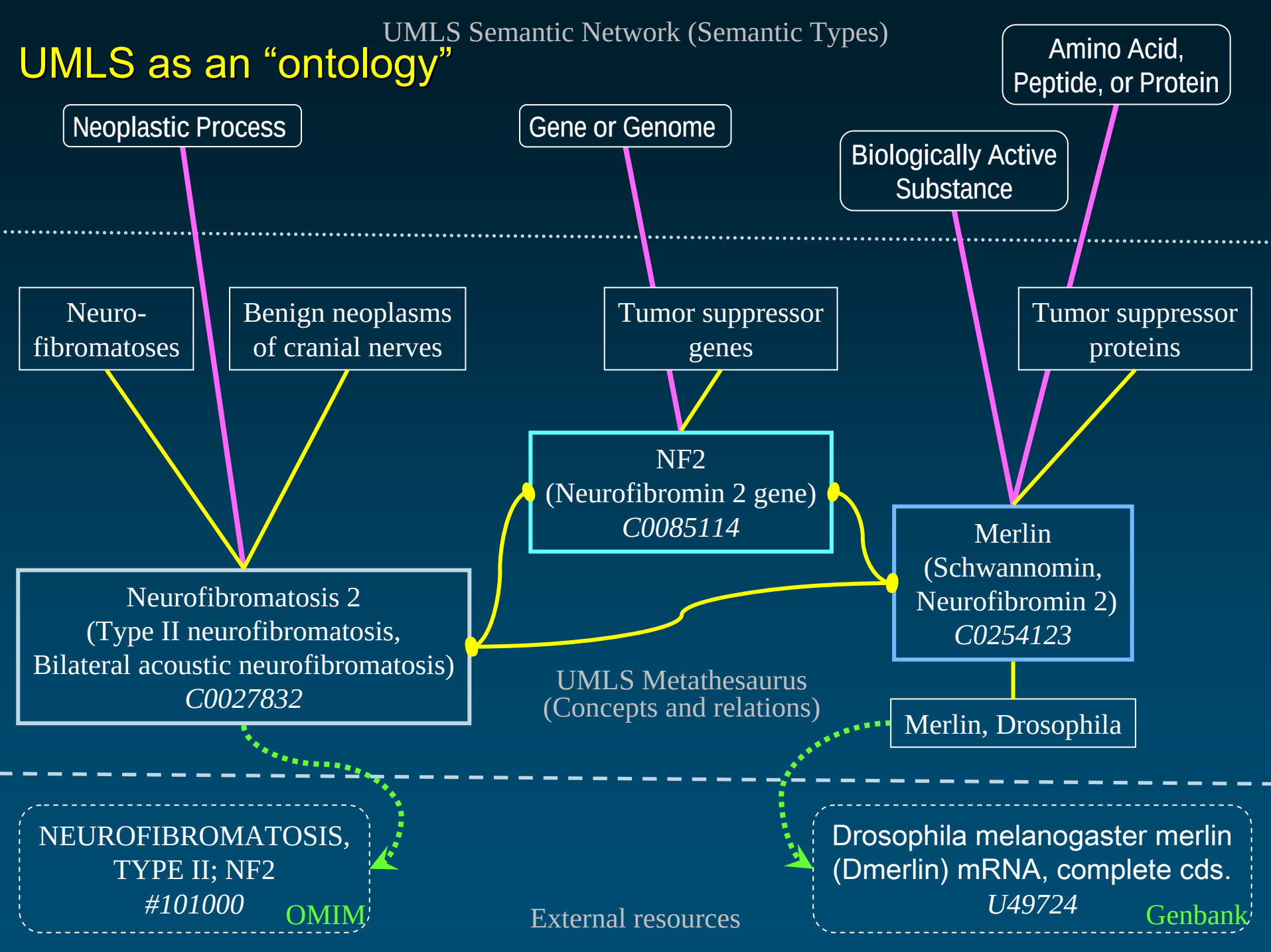
- Other formats
  - OBO – Open Biological Ontologies http://obo.sourceforge.net/browse.html
  - LexGrid http://informatics.mayo.edu/LexGrid/
- Converters
  - OBO – OWL http://www.bioontology.org/tools/oboinowl/obo_converter.html

UMLS vocabularies available in RDF/OWL

- NCI Thesaurus (OWL)
  - http://ncicb.nci.nih.gov/core/EVS
- Gene Ontology
  - http://www.geneontology.org/
- Repository of biomedical ontologies (OBO, OWL)
  - http://www.bioontology.org/ncbo/faces/index.xhtml

Porting vocabularies to OWL Experiments

- MeSH
  - Soualmia et al., KR-MED 2004
- Foundational Model of Anatomy (FMA)
  - Golbreich et al., JWS 2006 (OWL DL)
  - Noy and Rubin, SMI Tech Report 2007 (OWL Full)
- UMLS Semantic Network
  - Kashyap and Borgida, ISWC 2003
- UMLS Metathesaurus
  - Cornet and Abu-Hanna, AMIA 2002


UMLS Semantic Network (Semantic Types)
UMLS as an “ontology”
Neoplastic Process
Gene or Genome
Biologically Active Substance
Amino Acid, Peptide, or Protein
Neuro-fibromatoses
Benign neoplasms of cranial nerves
Tumor suppressor genes
Tumor suppressor proteins
NF2 (Neurofibromin 2 gene) C0085114
Merlin (Schwannomin, Neurofibromin 2) C0254123
Neurofibromatosis 2 (Type II neurofibromatosis, Bilateral acoustic neurofibromatosis) C0027832
UMLS Metathesaurus (Concepts and relations)
Merlin, Drosophila
NEUROFIBROMATOSIS, TYPE II; NF2 #101000
OMIM
External resources
Drosophila melanogaster merlin (Dmerlin) mRNA, complete cds. U49724
Genbank

❹ UMLS as an ontology 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)

Underspecified semantics

- Relationship “attribute” not always present
- Relations used to create hierarchies vs. hierachical relations

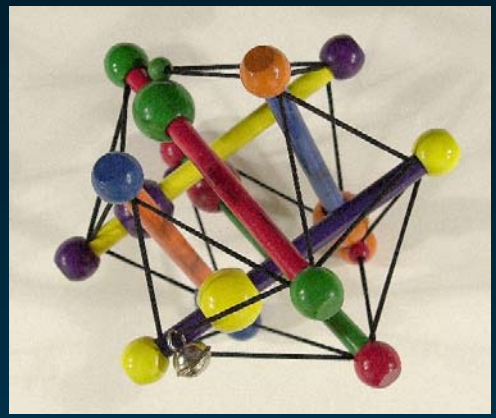
Summary

Biomedicine and Semantic Web

- Semantic Web technologies have not been widely adopted yet in biomedicine
  - OBO vs. OWL
  - caBIG vs. Taverna
- Use cases
  - Information/Data integration
- Recent efforts
  - W3C Health Care and Life Sciences Interest Group

UMLS and Semantic Web

- Terminology integration
- Based on existing terminologies
- Trans-namespace, permanent identifiers
- APIs available
  - Web Services-based API coming soon
- Can support information integration
- "Proprietary" representation (RRF)
- Some intellectual property restrictions
- Underspecified semantics
- No UMLS-based URIs



Medical Ontology Research

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

NATIONAL LIBRARY OF MEDICINE

Olivier Bodenreider

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

UMLS 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/semnav.pl
  - RRF browser
    (standalone application distributed with the UMLS)

UMLS References

- Gentle introduction
  - Bodenreider O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*; D267-D270. http://mor.nlm.nih.gov/pubs/pdf/2004-nar-ob.pdf
- Seminal paper
  - Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods Inf Med*, *32*(4), 281-91.

Semantic Web for Health Care and Life Sciences

- W3C Health Care and Life Sciences Interest Group
  - http://www.w3.org/2001/sw/hcls/

- Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung K-H. Advancing translational research with the Semantic Web. *BMC Bioinformatics* 2007;8(Suppl 3):S2. http://mor.nlm.nih.gov/pubs/pdf/2007-bmc_bioinformatics-ar.pdf

- Demo presented at the WWW2007 conference (May 2007) http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo

Biomedical information integration through RDF

- Biomedical perspective
  - Sahoo S, Zeng K, Bodenreider O, Sheth AP. (2007). From "glycosyltransferase" to "congenital muscular dystrophy": Integrating knowledge from NCBI Entrez Gene and the Gene Ontology. *Proceedings of Medinfo (in press).* http://mor.nlm.nih.gov/pubs/pdf/2007-medinfo-ss.pdf
- Semantic Web perspective
  - Sahoo S, Zeng K, Bodenreider O, Sheth AP. (2007). An experiment in integrating large biomedical knowledge resources with RDF: Application to associating genotype and phenotype information. *Proceedings of the workshop on Health Care and Life Sciences Data Integration for the Semantic Web at the 16th International World Wide Web Conference (WWW2007) (in press).* http://mor.nlm.nih.gov/pubs/pdf/2007-www_hcls-ss.pdf