IEEE International Conference on
Bioinformatics and Biomedicine
Philadelphia, Pennsylvania
November 5, 2008

Ontologies for Mining Biomedical Data

Olivier Bodenreider

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

Outline

- Biomedical ontologies
  - What they are
  - What they are for
  - Examples
- Ontologies for mining biomedical data
  - Normalization
  - Integration
  - Aggregation
- Applications of ontologies to data mining in biomedicine
  - Text mining – Information extraction
  - Biological – Mining gene expression data and functional annotations
  - Clinical – Mining adverse drug reactions

Biomedical ontologies

What is an ontology?

- The *What* question
  - Objects in the world
    - With their properties
    - With their relations to other objects
  - Also: events, processes, and states
- Explicit specification of a conceptualization
  - Support software applications [Gruber, 1993]
- Domain ontology reflects
  - Underlying reality
  - Theory of the domain

Ontology vs. other artifacts

[Smith, KR-MED 2006]
[Chute, JAMIA 2000]

- Ontology
  - Defining types of things and their relations
- Terminology
  - Naming things in a domain
- Thesaurus
  - Organizing things for a given purpose
- Classification
  - Placing things into (arbitrary) classes
- Knowledge bases
  - Assertional vs. definitional knowledge

Examples of biomedical ontologies

[J. Cimino, YBMI 2006]

- Structural perspective
  - What are they (vs. what are they for)?
- “High-impact” biomedical ontologies
  - International Classification of Diseases (ICD)
  - Logical Observation Identifiers, Names and Codes (LOINC)
  - SNOMED Clinical Terms
  - Foundational Model of Anatomy
  - Gene Ontology
  - RxNorm
  - Medical Subject Headings (MeSH)
  - NCI Thesaurus
  - Unified Medical Language System (UMLS)

Characteristics

Name	Scope	# concepts	Median	Subs. Hier	Version
SNOMED CT	Clinical medicine (patient records)	310,314	2	yes	July 31, 2007
LOINC	Clinical observations and laboratory tests	46,406	3	no	Version 2.21 (no “natural language” names)
FMA	Human anatomical structures	~72,000	?	yes	(not yet in the UMLS)
Gene Ontology	Functional annotation of gene products	22,546	1	yes	Jan. 2, 2007
RxNorm	Standard names for prescription drugs	93,426	1	no	Aug. 31, 2007
NCI Thesaurus	Cancer research, clinical care, public information	58,868	2	yes	2007_05E
ICD-10	Diseases and conditions (health statistics)	12,318	1	no	1998 (tabular)
MeSH	Biomedicine (descriptors for indexing the literature)	24,767	5	no	Aug. 27, 2007
UMLS .	Terminology integration in the life sciences	1,4 M	2	n/a	2007AC (English only)

NLM

[Bodenreider, YBMI 2008]

NCI Thesaurus

NATIONAL CANCER INSTITUTE

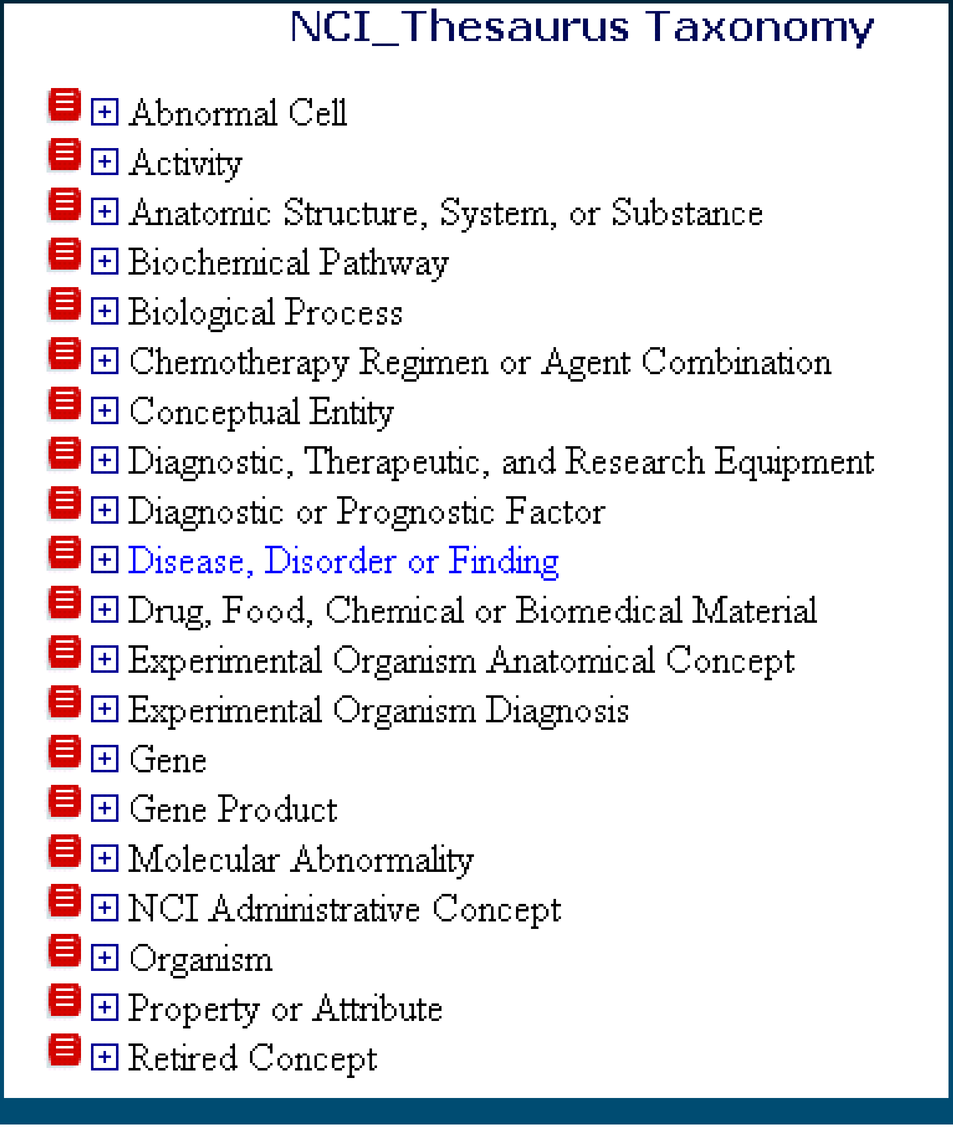
NCI thesaurus Characteristics (1)

- Current version: 08.08d (~monthly releases)
- Type: Controlled terminology / ontology
- Domain: Cancer
- Developer: NCI Center for Bioinformatics
- Funding: NCI
- Availability
  - Publicly available: Yes
  - Repositories: UMLS / OBO / NCBO BioPortal
- URL: http://nciterms.nci.nih.gov/

NCI thesaurus Characteristics (2)

- Number of
  - Concepts: 58,868 (2007_05E)
  - Terms: 2.68 per concept
- Major organizing principles:
  - Subsumption hierarchy
  - Rich set of associative relationships
  - Small proportion of defined concepts (many primitives)
  - Links to many external resources
- Formalism: OWL Lite

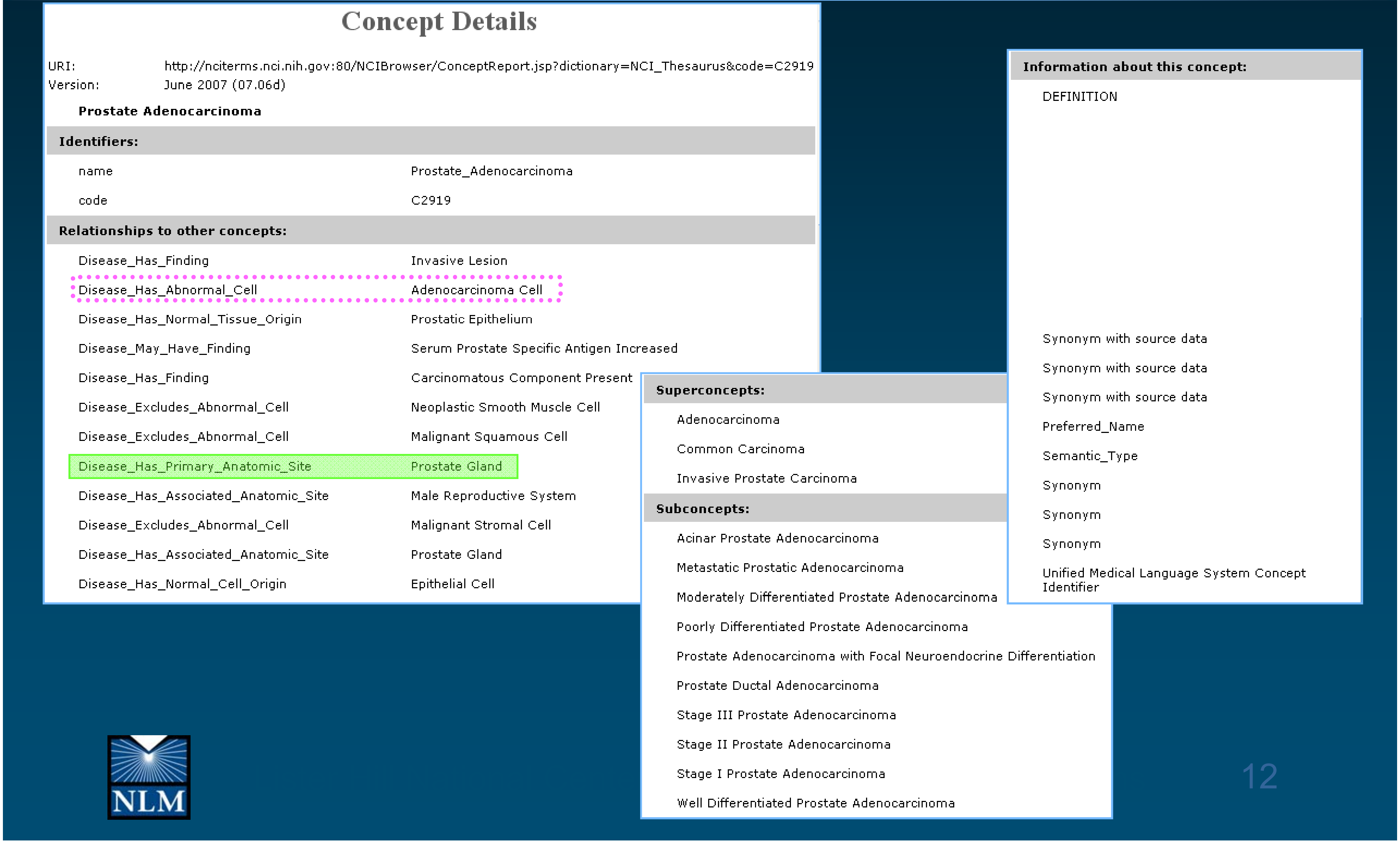
NCI thesaurus Top level



NCI_Thesaurus Taxonomy

- ⊞ Abnormal Cell
- ⊞ Activity
- ⊞ Anatomic Structure, System, or Substance
- ⊞ Biochemical Pathway
- ⊞ Biological Process
- ⊞ Chemotherapy Regimen or Agent Combination
- ⊞ Conceptual Entity
- ⊞ Diagnostic, Therapeutic, and Research Equipment
- ⊞ Diagnostic or Prognostic Factor
- ⊞ Disease, Disorder or Finding
- ⊞ Drug, Food, Chemical or Biomedical Material
- ⊞ Experimental Organism Anatomical Concept
- ⊞ Experimental Organism Diagnosis
- ⊞ Gene
- ⊞ Gene Product
- ⊞ Molecular Abnormality
- ⊞ NCI Administrative Concept
- ⊞ Organism
- ⊞ Property or Attribute
- ⊞ Retired Concept

NCI thesaurus Example



Concept Details

URI: http://nciterms.nci.nih.gov:80/NCIBrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&code=C2919
Version: June 2007 (07.06d)

Prostate Adenocarcinoma

Identifiers:

name	Prostate_Adenocarcinoma
code	C2919

Relationships to other concepts:

Disease_Has_Finding	Invasive Lesion
Disease_Has_Abnormal_Cell	Adenocarcinoma Cell
Disease_Has_Normal_Tissue_Origin	Prostatic Epithelium
Disease_May_Have_Finding	Serum Prostate Specific Antigen Increased
Disease_Has_Finding	Carcinomatous Component Present
Disease_Excludes_Abnormal_Cell	Neoplastic Smooth Muscle Cell
Disease_Excludes_Abnormal_Cell	Malignant Squamous Cell
Disease_Has_Primary_Anatomic_Site	Prostate Gland
Disease_Has_Associated_Anatomic_Site	Male Reproductive System
Disease_Excludes_Abnormal_Cell	Malignant Stromal Cell
Disease_Has_Associated_Anatomic_Site	Prostate Gland
Disease_Has_Normal_Cell_Origin	Epithelial Cell

Superconcepts:

Adenocarcinoma
Common Carcinoma
Invasive Prostate Carcinoma

Subconcepts:

Acinar Prostate Adenocarcinoma
Metastatic Prostatic Adenocarcinoma
Moderately Differentiated Prostate Adenocarcinoma
Poorly Differentiated Prostate Adenocarcinoma
Prostate Adenocarcinoma with Focal Neuroendocrine Differentiation
Prostate Ductal Adenocarcinoma
Stage III Prostate Adenocarcinoma
Stage II Prostate Adenocarcinoma
Stage I Prostate Adenocarcinoma
Well Differentiated Prostate Adenocarcinoma

Information about this concept:

DEFINITION

Synonym with source data
Synonym with source data
Synonym with source data
Preferred_Name
Semantic_Type
Synonym
Synonym
Synonym
Unified Medical Language System Concept Identifier

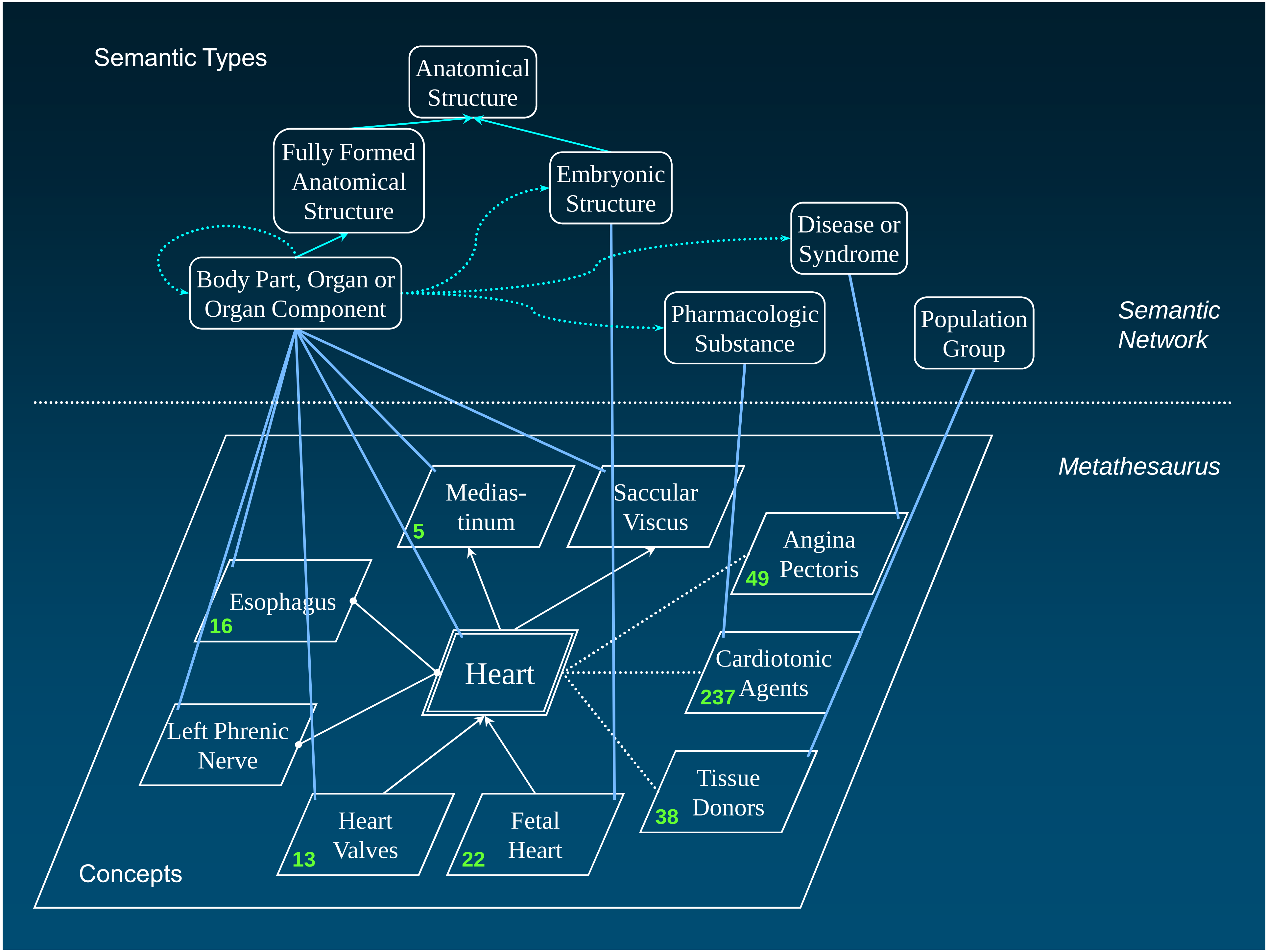
Unified Medical Language System (UMLS)

UMLS Characteristics (1)

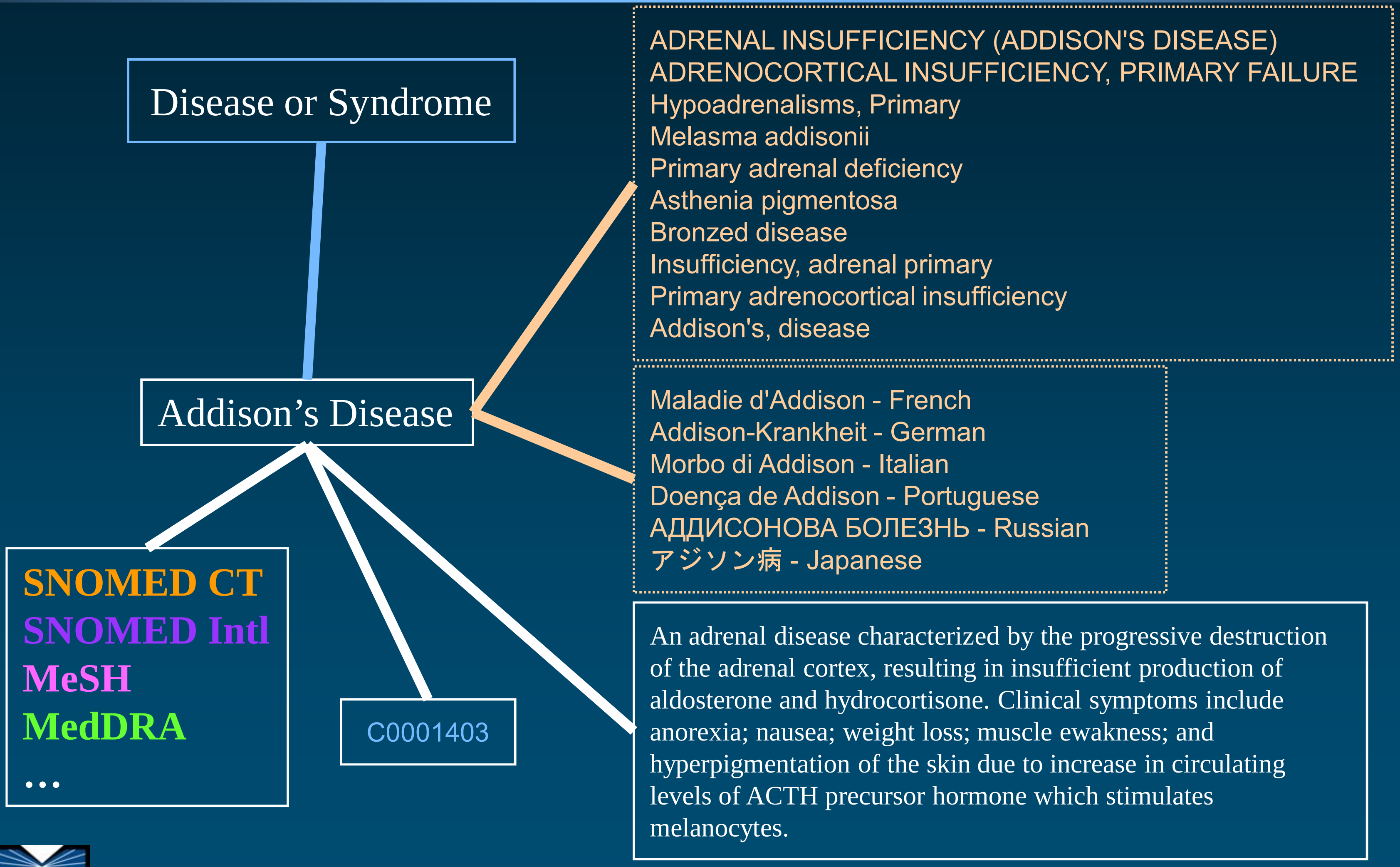
- Current version: 2008AA (2-3 annual releases)
- Type: Terminology integration system
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (intramural)
- Availability
  - Publicly available: Yes* (cost-free license required)
  - Repositories: UMLS
- URL: http://umlsks.nlm.nih.gov/

UMLS Characteristics (2)

- Number of
  - Concepts: 1.5M (2008AA)
  - Terms: ~6M
- Major organizing principles (Metathesaurus):
  - Concept orientation
  - Source transparency
  - Multi-lingual through translation
- Formalism: Proprietary format (RRF)


Semantic Types
Anatomical Structure
Fully Formed Anatomical Structure
Embryonic Structure
Body Part, Organ or Organ Component
Disease or Syndrome
Pharmacologic Substance
Population Group
Semantic Network
Metathesaurus
Medias- tinum
5
Saccular Viscus
Angina Pectoris
49
Esophagus
16
Heart
Cardiotonic Agents
237
Left Phrenic Nerve
Tissue Donors
38
Heart Valves
13
Fetal Heart
22
Concepts

Addison’s Disease: Concept



Disease or Syndrome

Addison’s Disease

ADRENAL INSUFFICIENCY (ADDISON'S DISEASE)
ADRENOCORTICAL INSUFFICIENCY, PRIMARY FAILURE
Hypoadrenalisms, Primary
Melasma addisonii
Primary adrenal deficiency
Asthenia pigmentosa
Bronzed disease
Insufficiency, adrenal primary
Primary adrenocortical insufficiency
Addison's, disease

Maladie d'Addison - French
Addison-Krankheit - German
Morbo di Addison - Italian
Doença de Addison - Portuguese
АДДИСОНОВА БОЛЕЗНЬ - Russian
アジソン病 - Japanese

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

SNOMED CT
SNOMED Intl
MeSH
MedDRA
…

C0001403

NLM

Biomedical ontologies in action

[Bodenreider, YBMI 2008]

- Functional perspective
  - What are they for (vs. what are they)?
- “High-impact” biomedical ontologies
- 3 major categories of use
  - Knowledge management (indexing and retrieval of data and information, access to information, mapping among ontologies)
  - Data integration, exchange and semantic interoperability
  - Decision support and reasoning (data selection and aggregation, decision support, natural language processing applications, knowledge discovery).

NLM

Ontologies for mining biomedical data

Normalization

Integration

Aggregation

Ontologies for mining biomedical data

Normalization

Integration

Aggregation

Issues

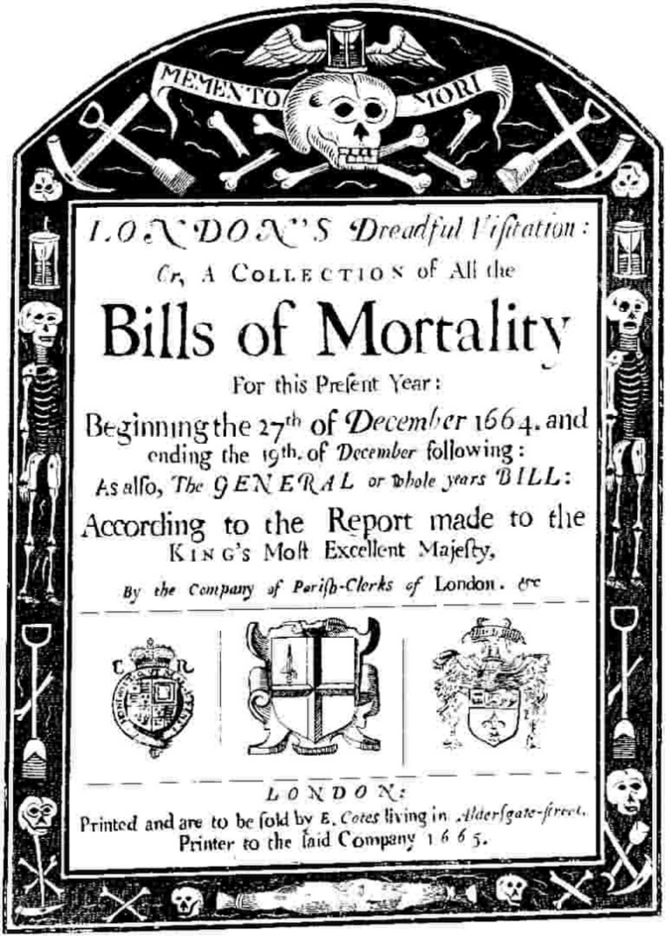
- Variability of natural language
  - Lexical variants
    - Lung cancer
    - Cancer of the lung
    - Lung cancers
  - Synonyms
    - Pulmonary cancer
    - Malignant neoplasm of lung
    - Malignant tumor of lung

Solutions with ontologies

- Controlled vocabulary
  - Standard list of terms to be used for a given purpose
    - Gene Ontology (functional annotation of gene products)
    - MeSH (indexing of biomedical articles)
    - ICD (mortality and morbidity reporting)

NLM

London Bills of Mortality


MEMENTO MORI

LONDON'S Dreadful Viſitation:

Or, A COLLECTION of All the

Bills of Mortality

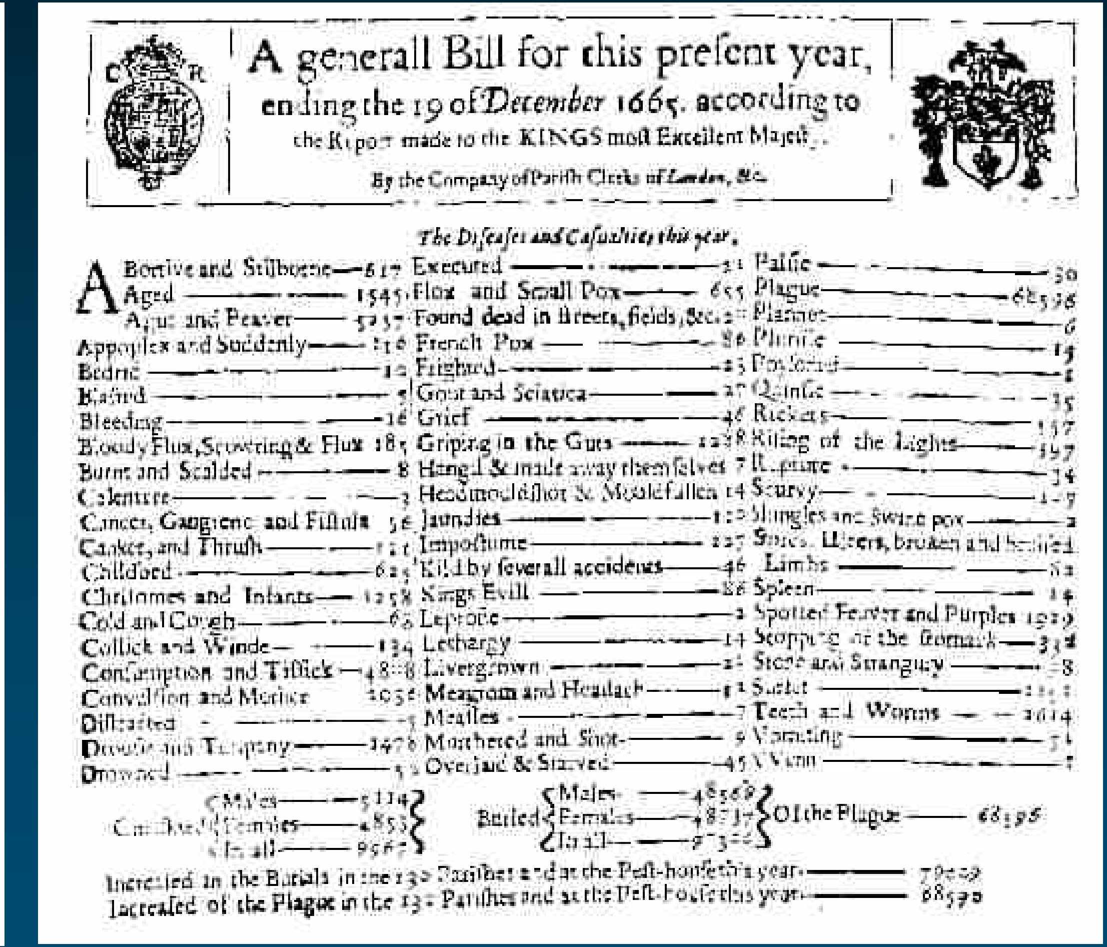
For this Preſent Year:

Beginning the 27th of *December* 1664. and
ending the 19th. of *December* following:
As alſo, *The GENERAL or whole years BILL:*

According to the Report made to the
KING's Moſt Excellent Majeſty.

By the Company of Pariſh-Clerks of London. &c

LONDON:
Printed and are to be ſold by *E. Cotes* living in *Alderſgate-ſtreet.*
Printer to the ſaid Company 1665.



A generall Bill for this preſent year,
ending the 19 of *December* 1665. according to
the Report made to the KINGS moſt Excellent Majeſty.

By the Company of Pariſh Clerks of *London*, &c.

The Diſeaſes and Caſualties this year.

Abortive and Stilborne	617	Executed	21	Palſie	30
Aged	1545	Flox and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in ſtreets, fields, &c.	20	Planet	6
Appoplex and Suddenly	116	French Pox	86	Pluriſie	15
Bedrid	10	Frighted	23	Poyſoned	1
Blaſted	5	Gout and Sciatica	27	Quinſie	35
Bleeding	16	Grief	46	Rickets	557
Bloudy Flux, Scowring & Flux	185	Griping in the Guts	1288	Riſing of the Lights	397
Burnt and Scalded	8	Hang'd & made away themſelves	7	Rupture	34
Calenture	3	Headmouldſhot & Mouldfallen	14	Scurvy	105
Cancer, Gangrene and Fiſtula	56	Jaundies	110	Shingles and Swine pox	2
Canker, and Thruſh	111	Impoſtume	227	Sores, Ulcers, broken and bruiſed	
Childbed	625	Kill'd by ſeverall accidents	46	Limbs	82
Chriſomes and Infants	1258	Kings Evill	86	Spleen	14
Cold and Cough	68	Leproſie	2	Spotted Feaver and Purples	1929
Collick and Winde	134	Lethargy	14	Stopping of the Stomack	332
Conſumption and Tiſſick	4808	Livergrown	20	Stone and Strangury	98
Convulſion and Mother	2036	Meagrom and Headach	12	Surfet	1251
Diſtracted	5	Meaſles	7	Teeth and Worms	2614
Dropſie and Timpany	1478	Murthered and Shot	9	Vomiting	51
Drowned	50	Overlaid & Starved	45	Wenn	1

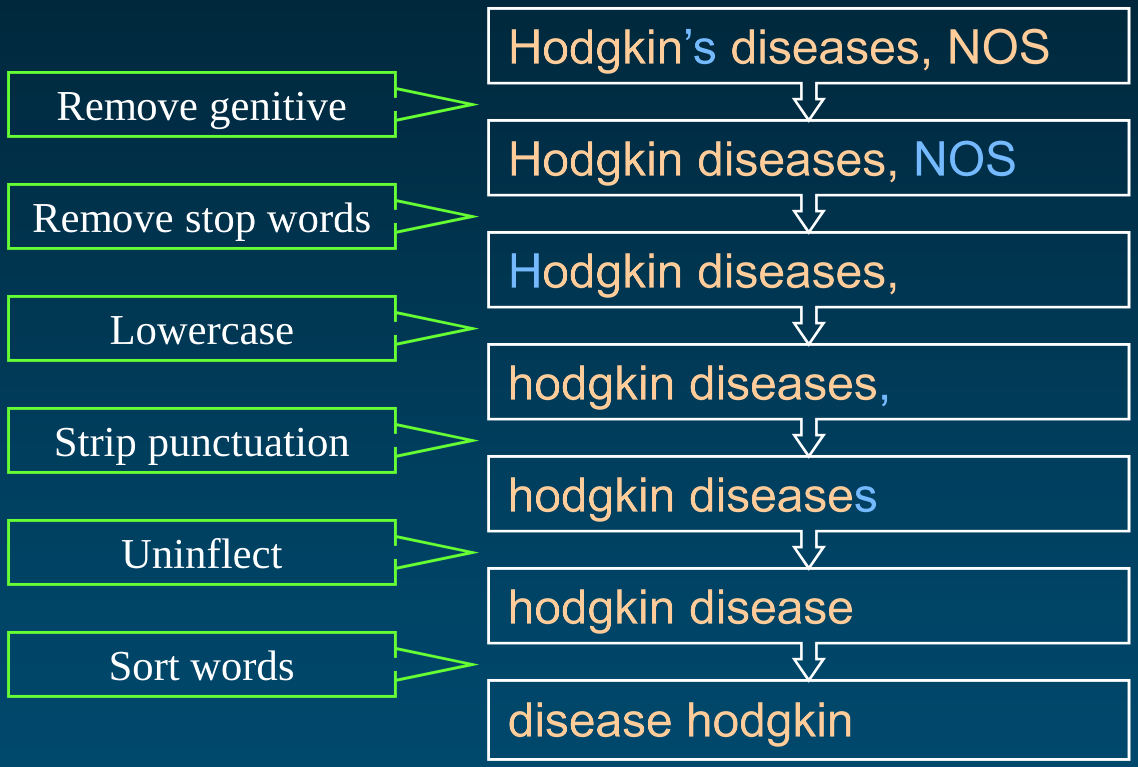
Chriſtned	Males	5114	Buried	Males	48569	Of the Plague 68596
	Females	4853		Females	48737	
	In all	9967		In all	97306	

Increaſed in the Burials in the 130 Pariſhes and at the Peſt-houſe this year — 79009
Increaſed of the Plague in the 130 Pariſhes and at the Peſt-houſe this year — 68590

Solutions

- Controlled vocabulary
  - Standard list of terms to be used for a given purpose
    - Gene Ontology (functional annotation of gene products)
    - MeSH (indexing of biomedical articles)
    - ICD (mortality and morbidity reporting)
- Lexical normalization programs [McCray, SCAMC 1994]
  - Management of lexical terminological variability
  - UMLS *Lexical Variant Generation* program
  - Used in terminology integration systems (e.g., UMLS)
  - Useful for indexing and text mining applications

Normalization



Hodgkin’s diseases, NOS

Remove genitive

Hodgkin diseases, NOS

Remove stop words

Hodgkin diseases,

Lowercase

hodgkin diseases,

Strip punctuation

hodgkin diseases

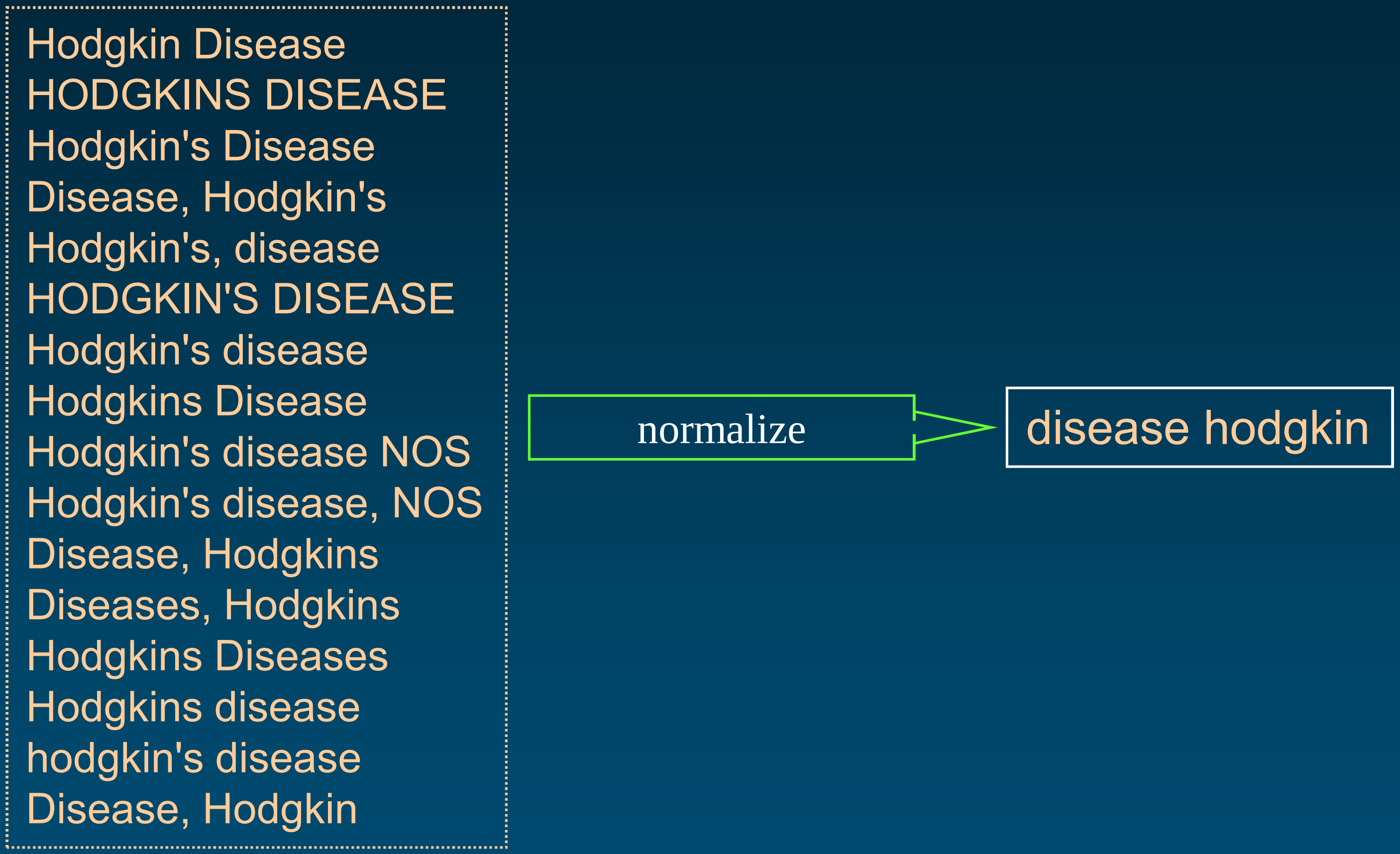
Uninflect

hodgkin disease

Sort words

disease hodgkin

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

Ontologies for mining biomedical data

Normalization

Integration

Aggregation

Issues

- Different codes for the same biomedical entity in different ontologies

SNOMED CT:	363732003	Addison's disease
MeSH:	D000224	Addison Disease
NCI Thesaurus:	C26689	Addison's Disease
ICD 9-CM:	255.41	Addison's disease NOS
ICD 10:	E27.1	Primary adrenocortical insufficiency
MedDRA:	10001130	Addison's disease
…		

- Hindrance to the integration of datasets (e.g., clinical, research and epidemiology data)

Solutions with ontologies

- Identify equivalent concepts across ontologies

- Specific mappings
  - SNOMED to ICD 9-CM (provided by SNOMED)
- Terminology integration systems
  - Manually curated
    - Unified Medical Language System (UMLS)
    - RxNorm (for drug vocabularies)
  - Automatically aligned
    - BioPortal

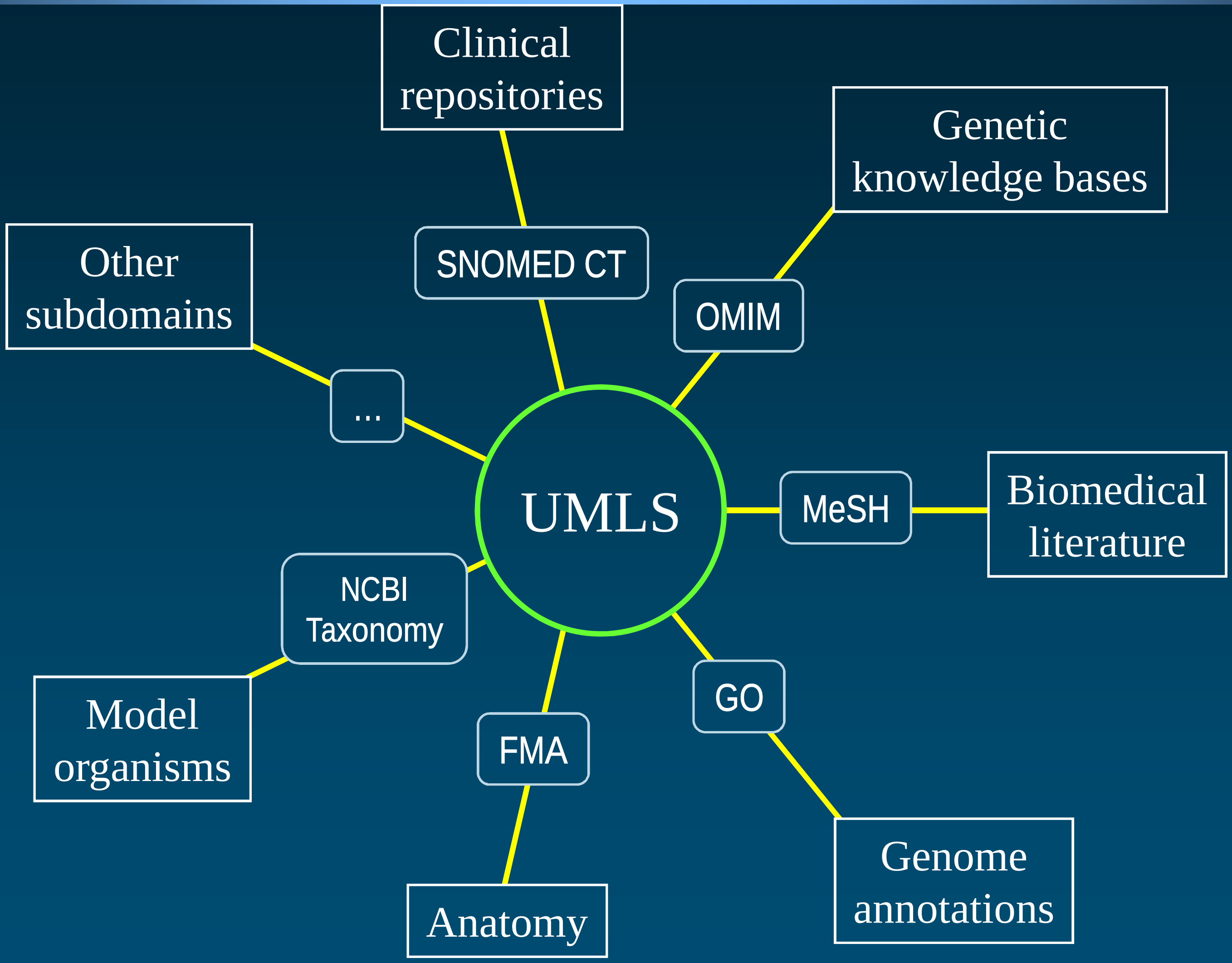
Terminology integration in the UMLS

SNOMED CT:	363732003	Addison's disease
MeSH:	D000224	Addison Disease
NCI Thesaurus:	C26689	Addison's Disease
ICD 9-CM:	255.41	Addison's disease NOS
ICD 10:	E27.1	Primary adrenocortical insufficiency
MedDRA:	10001130	Addison's disease
…		

C0001403

- Identified as synonyms (semi-automatically)
- Clustered into a UMLS concept
- Assigned a permanent identifier (CUI)

Integrating subdomains


Clinical repositories

SNOMED CT

Genetic knowledge bases

OMIM

Other subdomains

...

UMLS

MeSH

Biomedical literature

NCBI Taxonomy

Model organisms

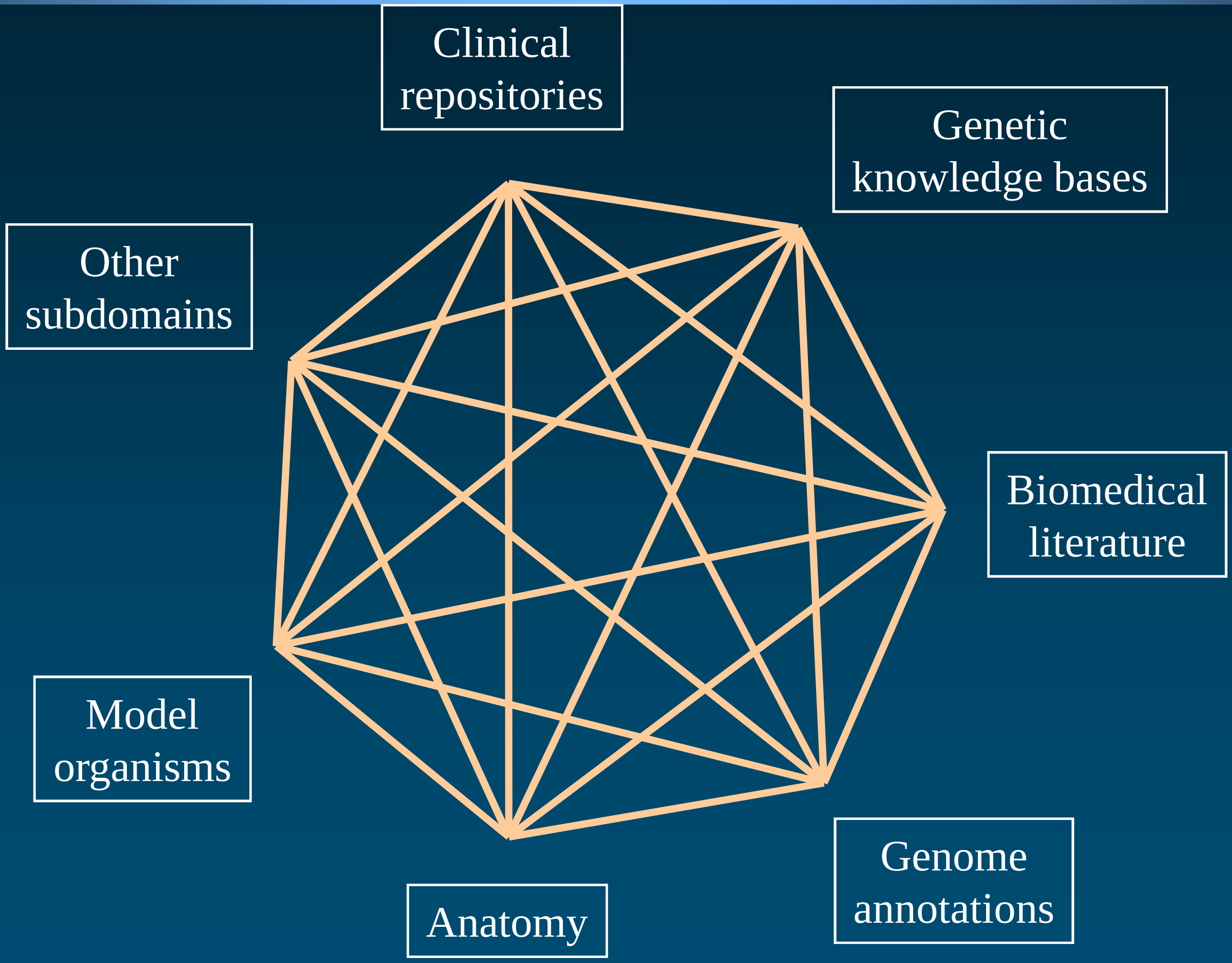
FMA

Anatomy

GO

Genome annotations

Integrating subdomains



Clinical repositories

Genetic knowledge bases

Biomedical literature

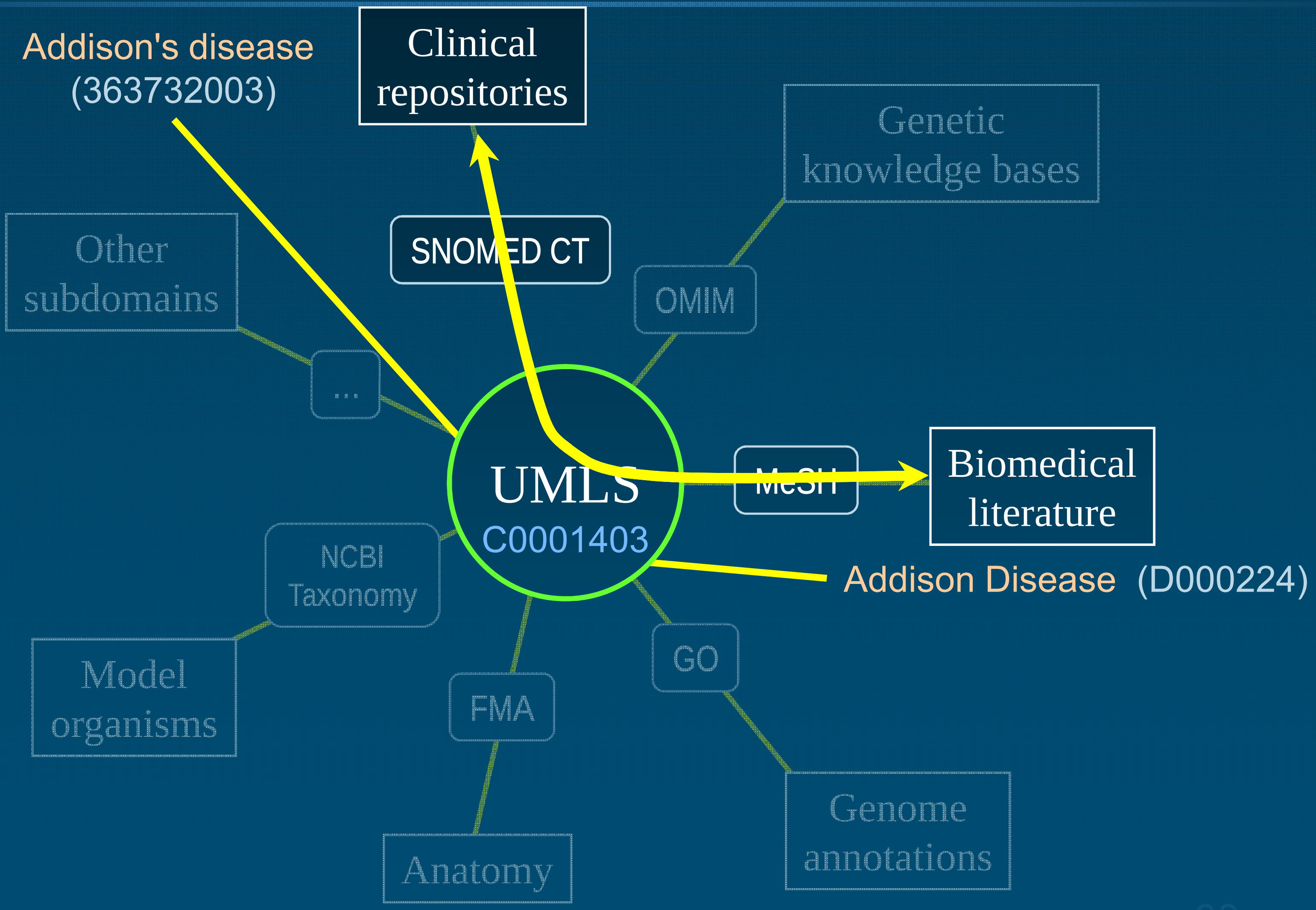
Genome annotations

Anatomy

Model organisms

Other subdomains

Trans-namespace integration



Ontologies for mining biomedical data

Normalization

Integration

Aggregation

Issues

- Various levels of granularity
  - Upper limb
    - Hand
      - Index finger
        - Diaphysis of distal phalanx of left index finger
- Fine-grained may not be appropriate for high-level analysis
  - Reduce statistical power
- Need to abstract away from details
  - Aggregate into a more generic concept
  - Corollary: Apply to more specific concepts

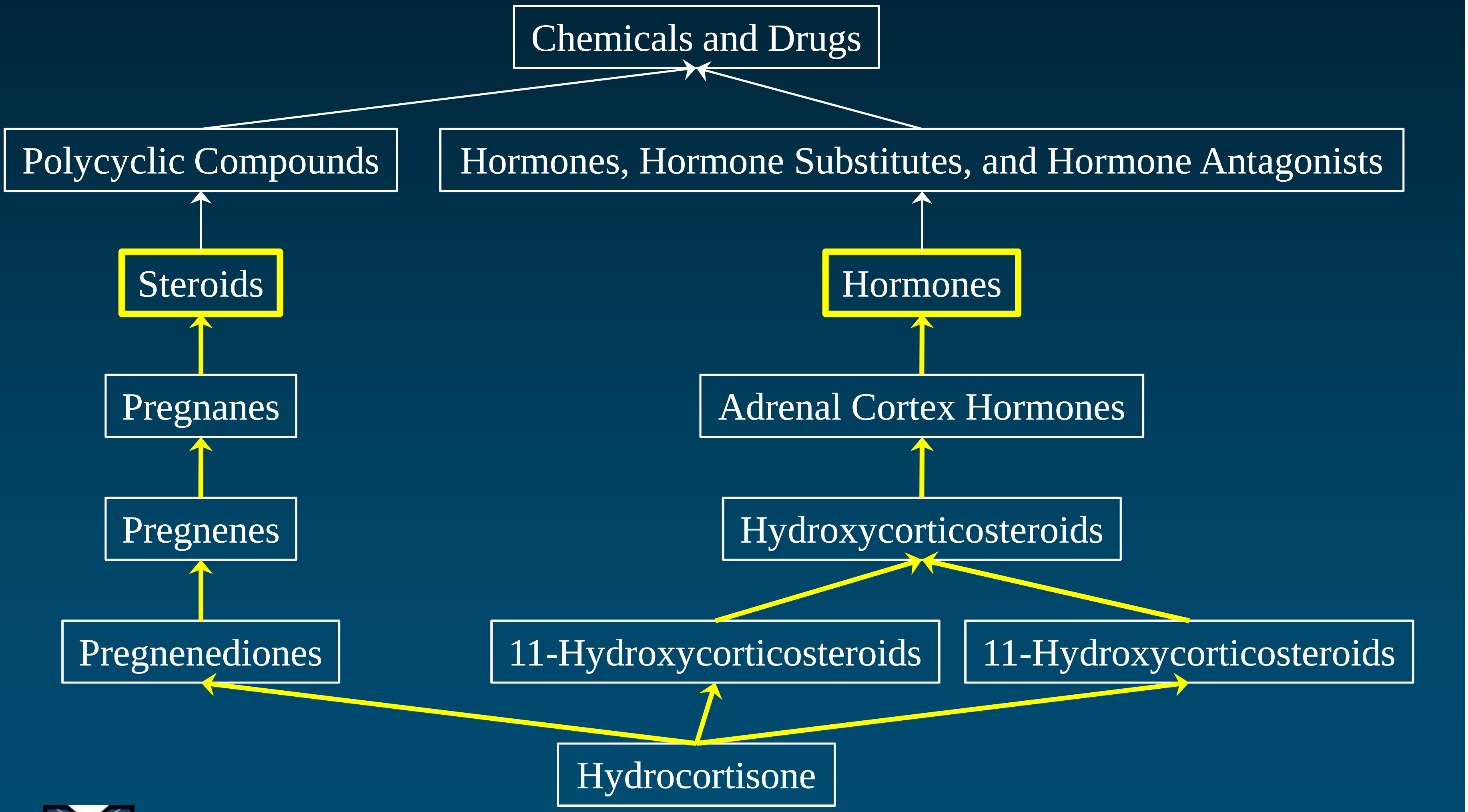
NLM

Solutions with ontologies

- Aggregate along subsumption hierarchies
- Helps enrich feature sets for data mining purposes

- Examples
  - GO Slims (analysis of functional annotations)
  - Categorization of adverse events based on high-level disease categories
- Corollary
  - Patient selection based on high-level ICD 9-CM codes
  - MeSH term “explosion” (information retrieval)

Aggregation with MeSH



Chemicals and Drugs

Polycyclic Compounds

Hormones, Hormone Substitutes, and Hormone Antagonists

Steroids

Hormones

Pregnanes

Adrenal Cortex Hormones

Pregnenes

Hydroxycorticosteroids

Pregnenediones

11-Hydroxycorticosteroids

11-Hydroxycorticosteroids

Hydrocortisone

Ontologies for mining biomedical data

Text mining

Analysis of gene expression data

Mining adverse drug reactions

Ontologies for mining biomedical data

Text mining

Analysis of gene expression data

Mining adverse drug reactions

Ontological resources for text mining

- Lexical resources
  - SPECIALIST lexicon (UMLS) / LVG
  - Lexico-syntactic analysis, normalization
- Terminological resources
  - UMLS Metathesaurus / MetaMap
  - Named entity recognition, semantic normalization
- Ontological resources
  - UMLS Semantic Network / SemRep
  - Relation extraction, semantic interpretation

[Ananiadou, Text mining for biology and biomedicine 2006]

Ontologies for mining biomedical data

Text mining

Analysis of gene expression data

Mining adverse drug reactions

Traditional approach

Analysis of gene expression data

- Cluster analysis
  - Genes
  - Genes and samples
- Elicitation of clusters using external knowledge
  - Functional annotations
  - Participation in pathways

Clustering constraints from ontologies

- Use ontologies as a source of prior knowledge
- Ontologies used to constrain the clustering process
  - Several variants of the clustering algorithms
- Tends to result in more meaningful clusters

[Liu, CSB 2004]
[Kustra, CBMS 2006]
[Huang, Omics 2006]
[Chabalier, BMC Bioinfo 2007]

Ontologies for mining biomedical data

Text mining

Analysis of gene expression data

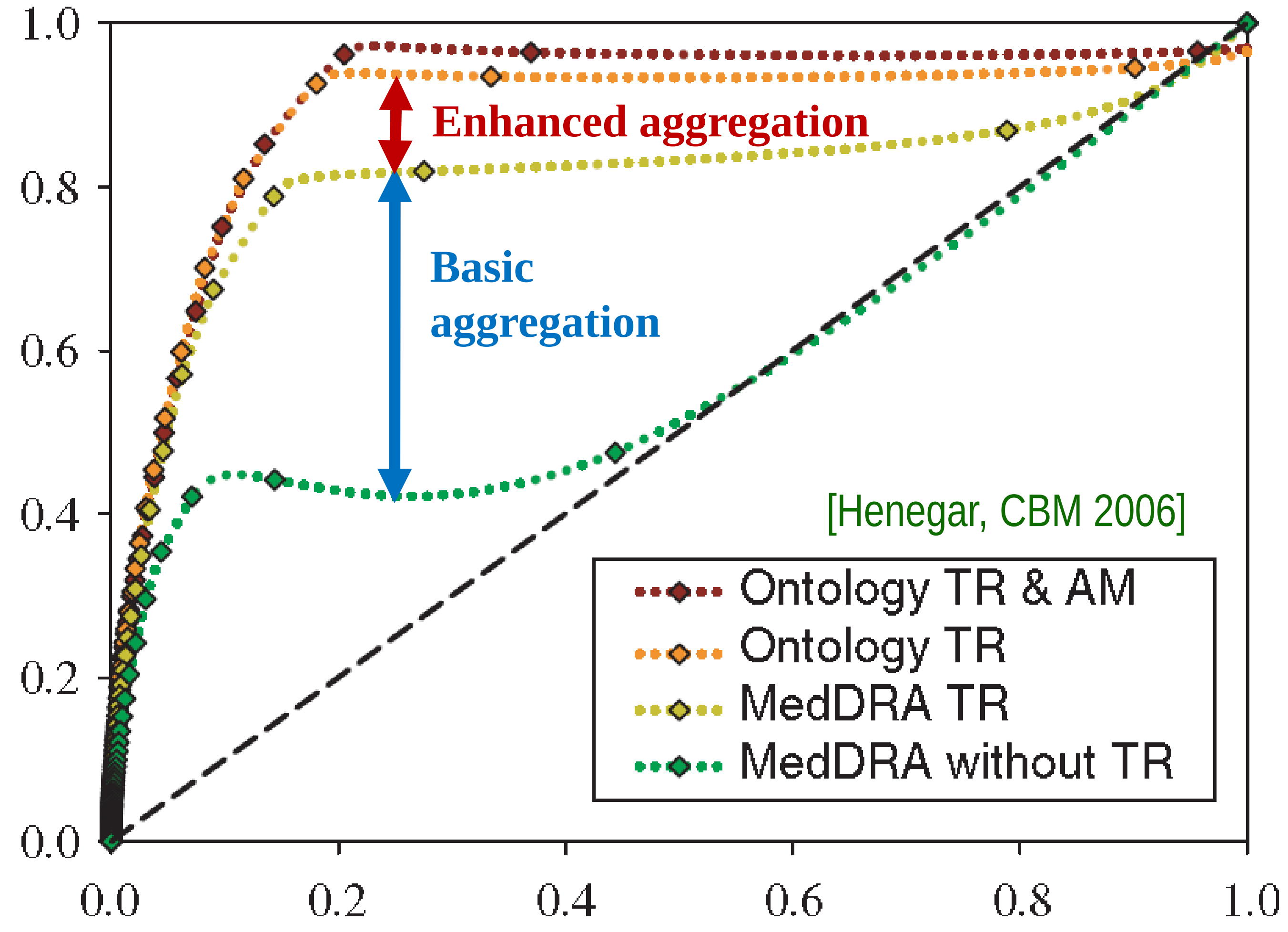
Mining adverse drug reactions

Identifying adverse drug reactions

- Pharmacovigilance of self-reported ADR cases
  - Coded with MedDRA
  - Manually curated
- Bayesian analysis of the drug-ADR associations
- 4 variants
  - MedDRA without subsumption links
  - MedDRA with original subsumption links
  - MedDRA with enhanced subsumption links
  - MedDRA with enhanced subsumption links and approximate matching

[Henegar, CBM 2006]

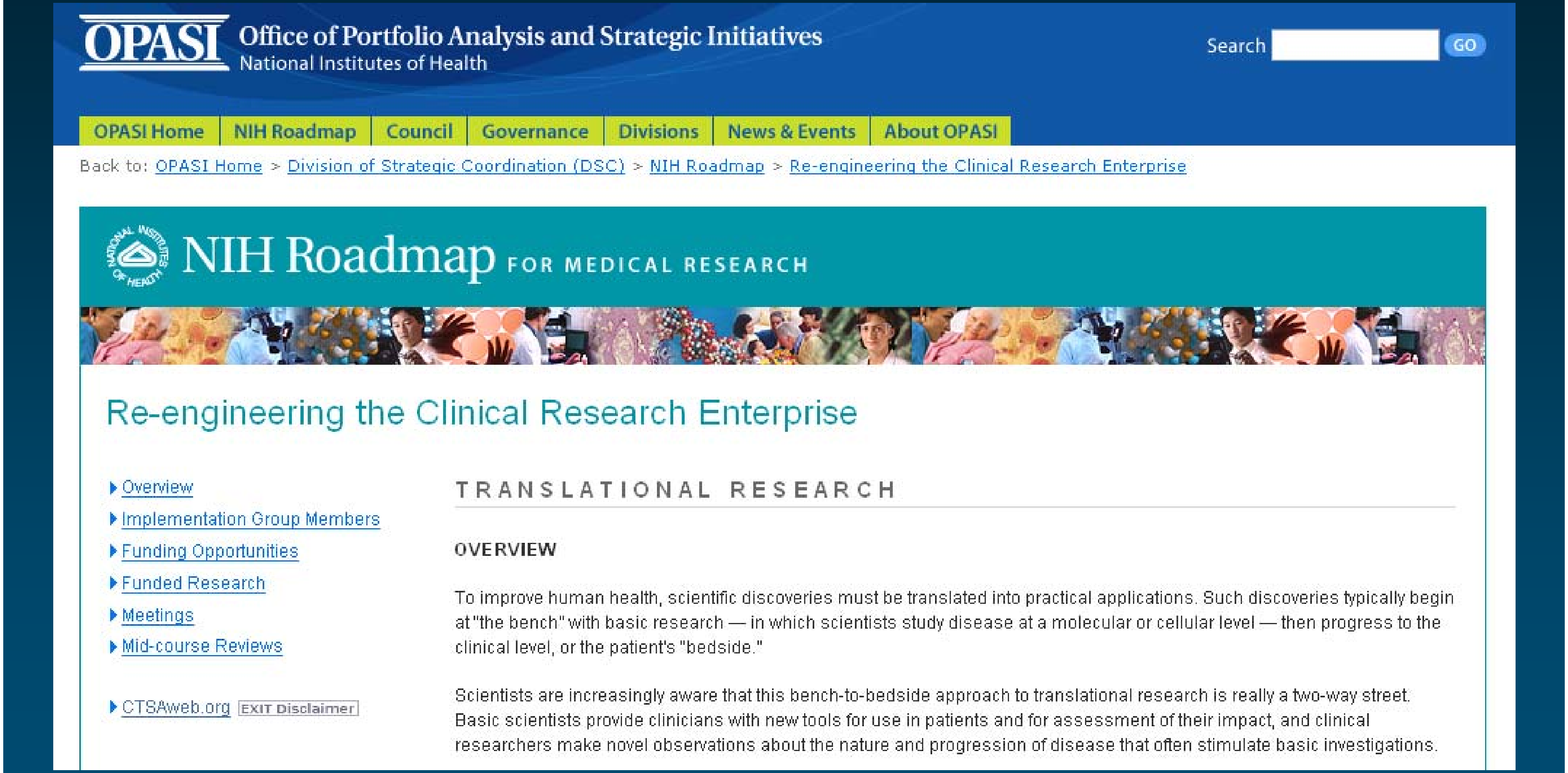
Using subsumption links increases the signal



[Henegar, CBM 2006]

Conclusions

Translational research NIH Roadmap


OPASI Office of Portfolio Analysis and Strategic Initiatives
National Institutes of Health
Search
GO
OPASI Home
NIH Roadmap
Council
Governance
Divisions
News & Events
About OPASI
Back to: OPASI Home > Division of Strategic Coordination (DSC) > NIH Roadmap > Re-engineering the Clinical Research Enterprise
NIH Roadmap FOR MEDICAL RESEARCH
Overview
Implementation Group Members
Funding Opportunities
Funded Research
Meetings
Mid-course Reviews
CTSAweb.org
EXIT Disclaimer

Re-engineering the Clinical Research Enterprise

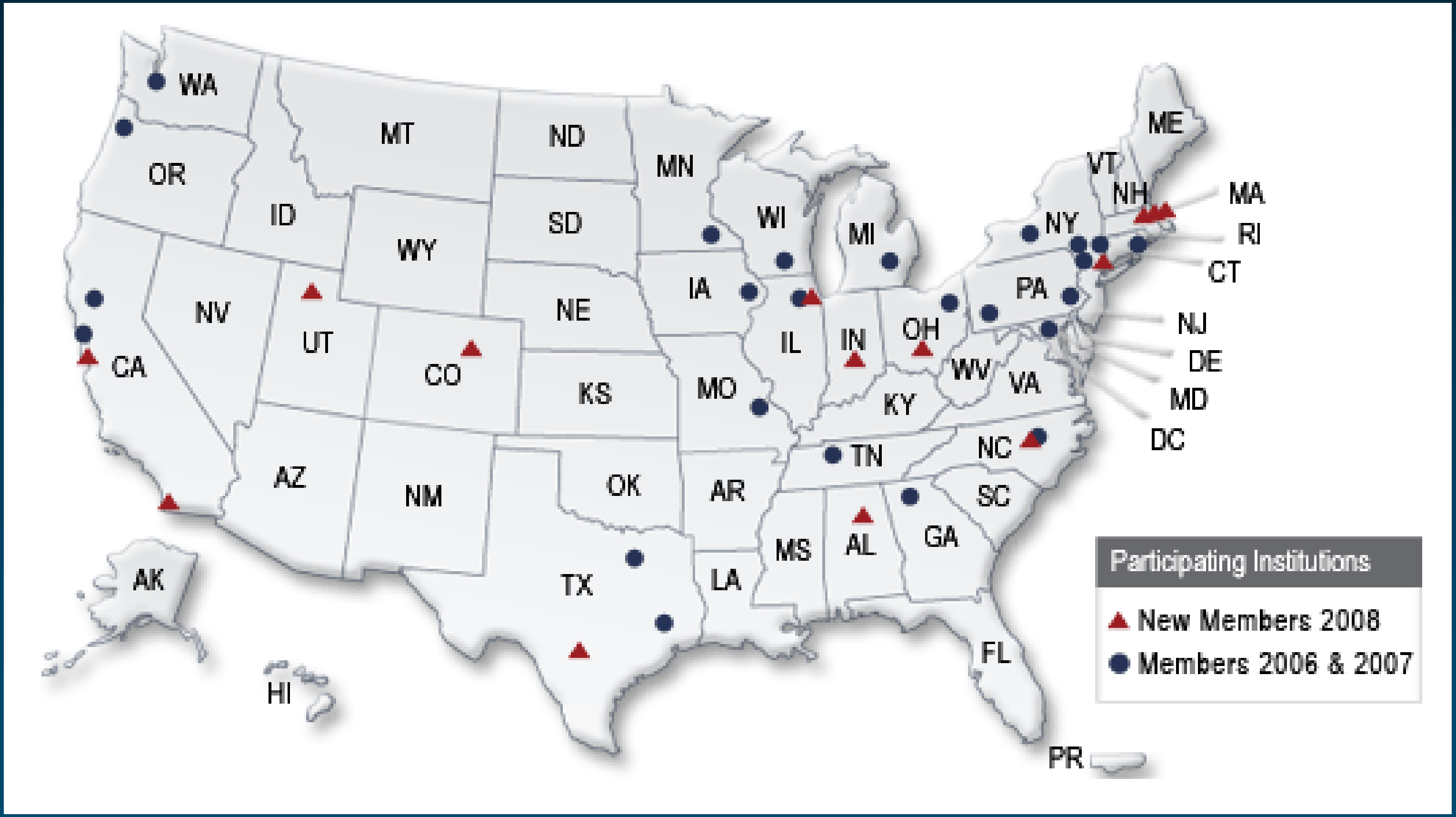
TRANSLATIONAL RESEARCH

OVERVIEW

To improve human health, scientific discoveries must be translated into practical applications. Such discoveries typically begin at "the bench" with basic research — in which scientists study disease at a molecular or cellular level — then progress to the clinical level, or the patient's "bedside."

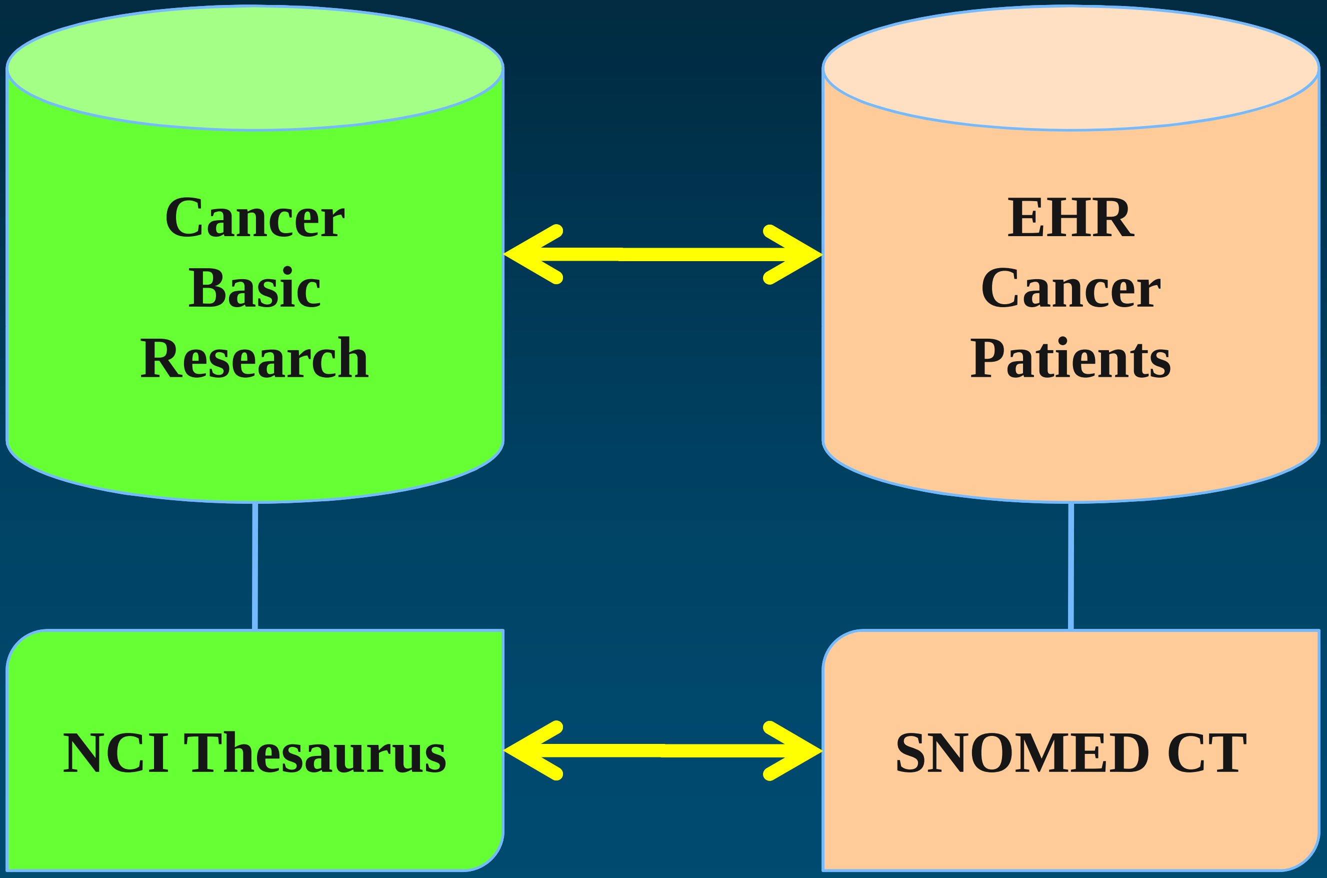
Scientists are increasingly aware that this bench-to-bedside approach to translational research is really a two-way street. Basic scientists provide clinicians with new tools for use in patients and for assessment of their impact, and clinical researchers make novel observations about the nature and progression of disease that often stimulate basic investigations.

Clinical and Translational Science Awards



http://www.ctsaweb.org/

Ontology and translational research



Ontologies for data mining

- Ontologies
  - Normalize datasets
  - Aggregate data of different granularity

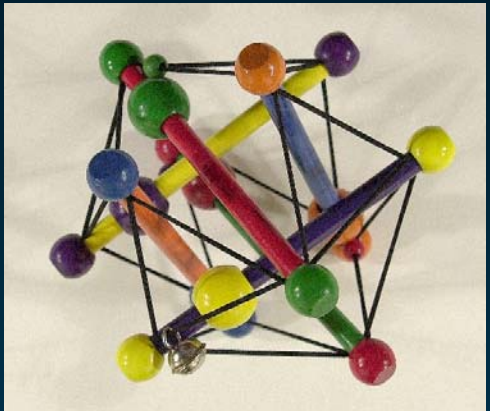
increase signal

- Ontology integration systems
  - Integrate datasets
- Challenges
  - Permanent identifiers for biomedical entities
  - Availability
  - Quality

Data mining with ontologies

- Ontologies are increasingly used in biological data mining
  - Text mining
    - Named entity recognition
    - Relation extraction
  - In combination with other features
  - To enhance feature sets

- Few data mining algorithms natively take advantage of ontologies



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