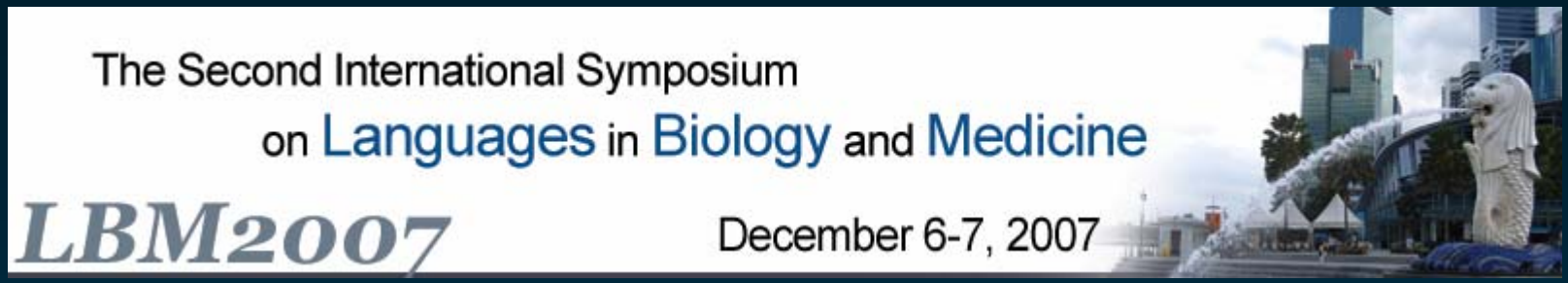


Singapore - December 6, 2007

Terminological systems in biomedicine

From terminology integration to information integration

Olivier Bodenreider

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

Outline

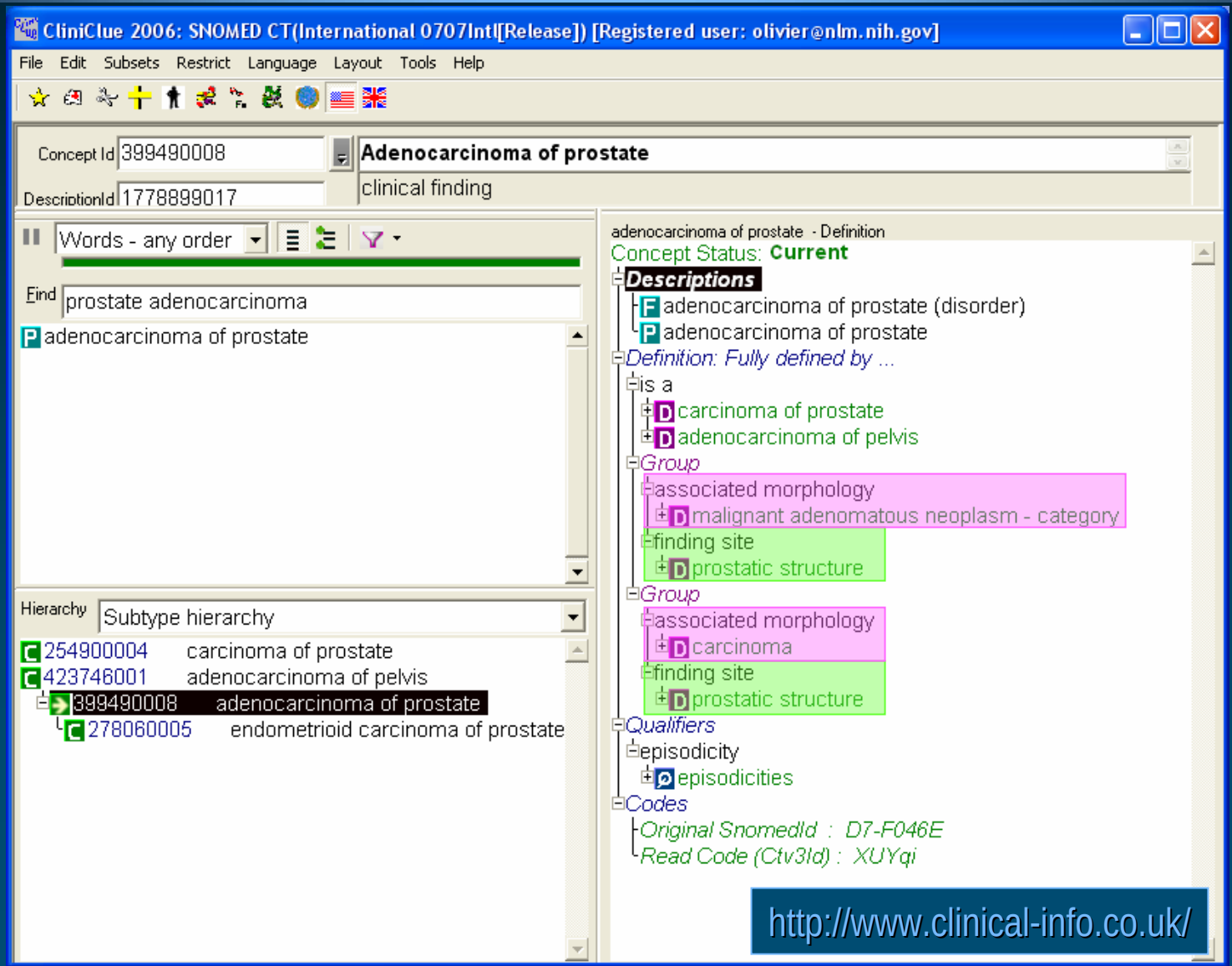
- Information integration in biomedicine
  - Use case: Oncology
  - Some issues: naming, normalization, mapping
- Terminology integration in biomedicine
  - Bottom-up
    Unified Medical Language System
  - Top-down
    OBO Foundry ontologies
- Applications
  Biomedical Semantic Web

Information integration in oncology

Information sources and terminologies

- Multiple terminologies for oncology
  - International Classification of Diseases-Oncology (ICD-O-3)
    - Cancer registries
    - Epidemiology, Public health
  - SNOMED CT
    - Patient records
    - Clinical care
  - NCI Thesaurus
    - Annotation of research data

SNOMED CT


CliniClue 2006: SNOMED CT(International 0707Intl[Release]) [Registered user: olivier@nlm.nih.gov]
File Edit Subsets Restrict Language Layout Tools Help
Concept Id 399490008
Adenocarcinoma of prostate
DescriptionId 1778899017
clinical finding
Words - any order
Find prostate adenocarcinoma
P adenocarcinoma of prostate
Hierarchy Subtype hierarchy
C 254900004 carcinoma of prostate
C 423746001 adenocarcinoma of pelvis
399490008 adenocarcinoma of prostate
C 278060005 endometrioid carcinoma of prostate
adenocarcinoma of prostate - Definition
Concept Status: Current
Descriptions
F adenocarcinoma of prostate (disorder)
P adenocarcinoma of prostate
Definition: Fully defined by ...
is a
D carcinoma of prostate
D adenocarcinoma of pelvis
Group
associated morphology
D malignant adenomatous neoplasm - category
finding site
D prostatic structure
Group
associated morphology
D carcinoma
finding site
D prostatic structure
Qualifiers
episodicity
episodicities
Codes
Original SnomedId : D7-F046E
Read Code (Ctv3Id) : XUYqi
http://www.clinical-info.co.uk/

NCI Thesaurus

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

Information about this concept:

DEFINITION

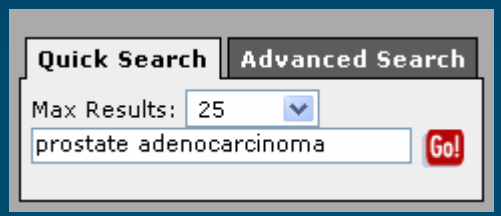
Synonym with source data
Synonym with source data
Synonym with source data
Preferred_Name
Semantic_Type
Synonym
Synonym
Synonym
Unified Medical Language System Concept Identifier

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



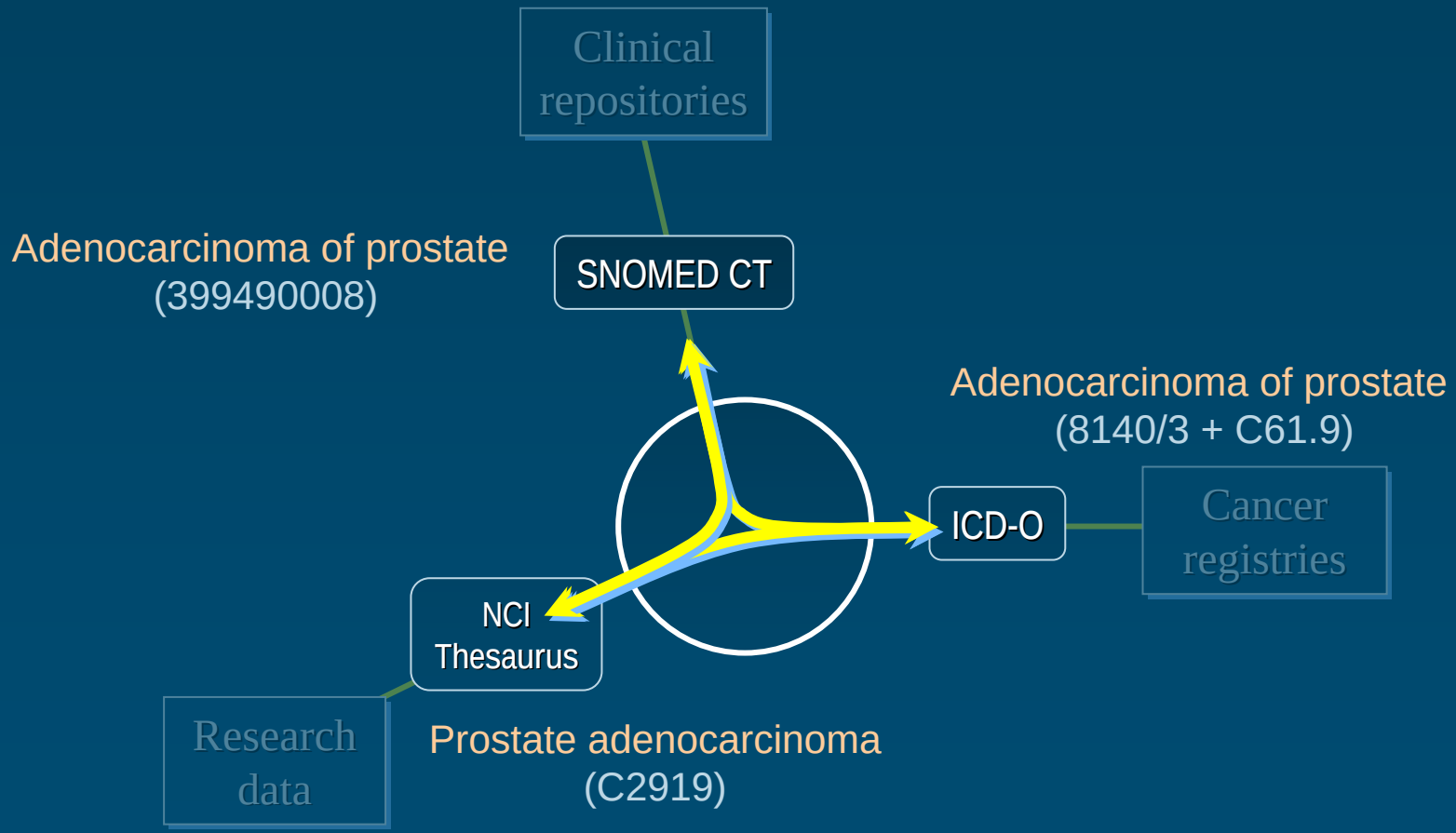
http://nciterms.nci.nih.gov/

ICD-O-3

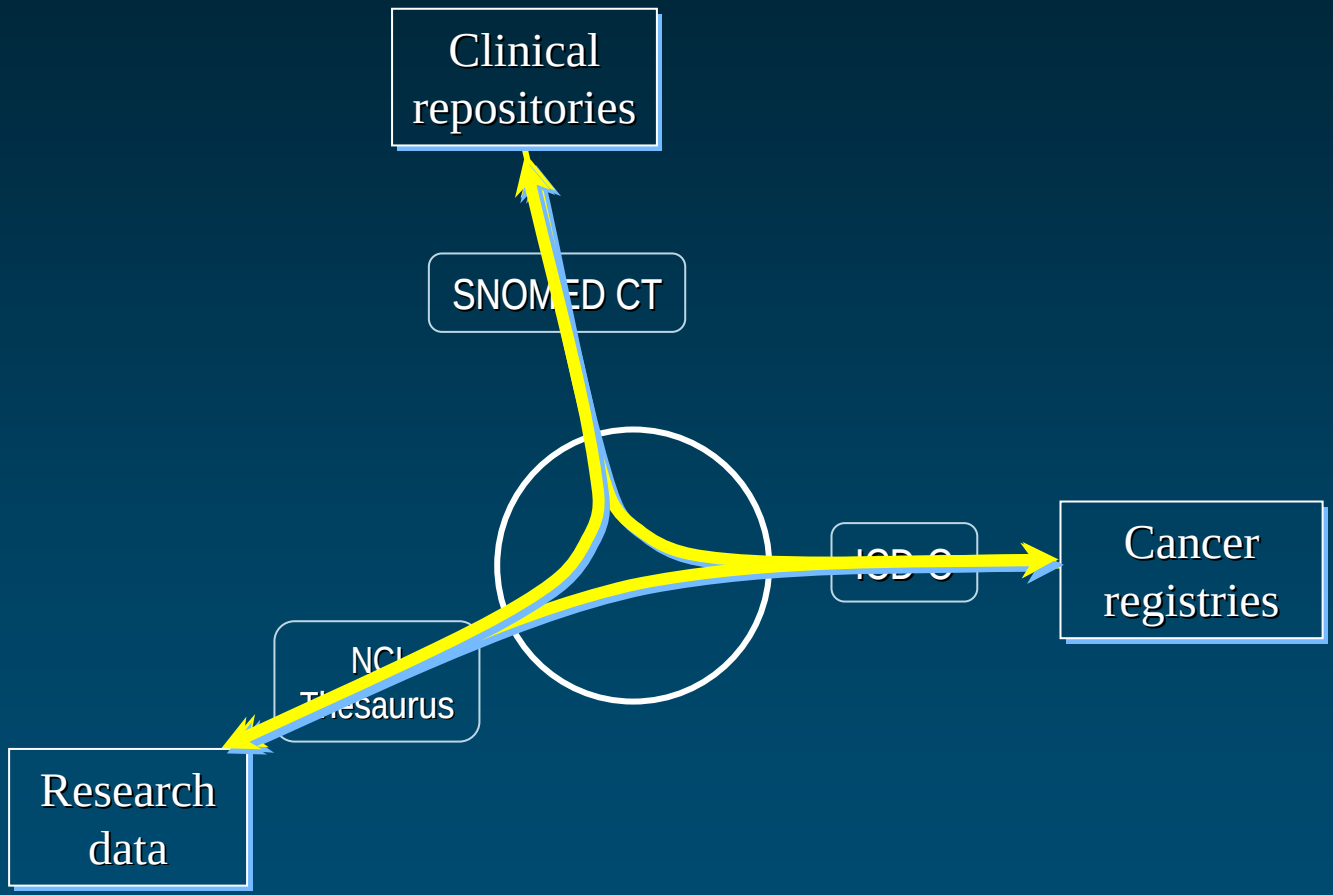
- Morphology
  - […]
  - 814-838 Adenomas and adenocarcinomas
    - 8140/3 Adenocarcinoma, NOS
- Anatomy
  - […]
  - C60-C63 Male genital organs
    - C61 Prostate gland
      - C61.9 Prostate, NOS
        Prostate gland

Adenocarcinoma of prostate

Integrating terminologies



Integrating subdomains



Information integration in biomedicine

Some issues: naming, normalization, mapping

❶ Naming

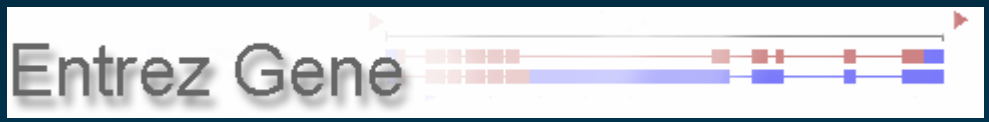
- Many biomedical entities have several names (synonymy)
  - Drug names
  - Gene names
  - Disease names
  - …
- A given name may refer to several different entities (polysemy)
  - Nail (body part)
  - Nail (medical device)

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

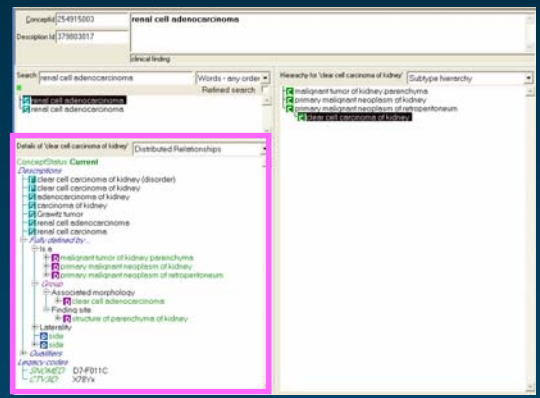
Fully defined by...

- 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
  - Findinq site
    - D structure of parenchyma of kidney
- Laterality
  - R side
  - R side

Qualifiers

Legacy codes

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



http://www.clinical-info.co.uk/

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

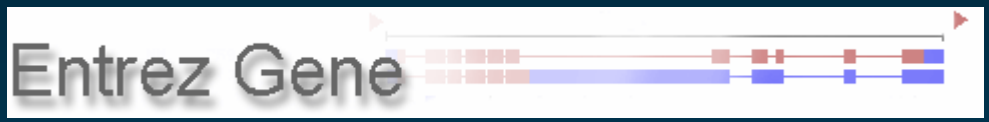
- Biomedical entities are identified by unique identifiers in various terminology systems
- Resolve names into identifiers (in a given namespace)
- Supported (in part) by terminology resources
- Example: BioCreative
  - 1B and 2GN – Gene Normalization

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

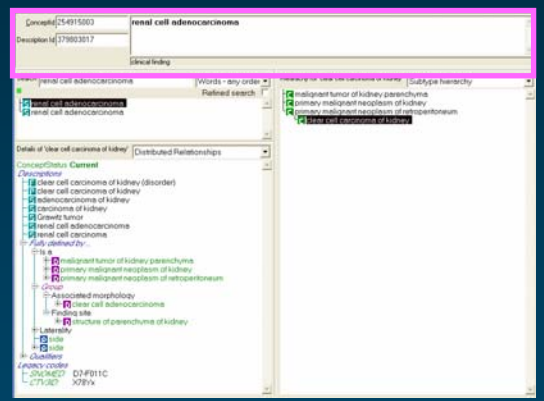
Fully defined by...

- 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
  - R side
  - R side

Qualifiers

Legacy codes

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



ConceptId 254915003 | renal cell adenocarcinoma

Description Id 379803017 | clinical finding

http://www.clinical-info.co.uk/

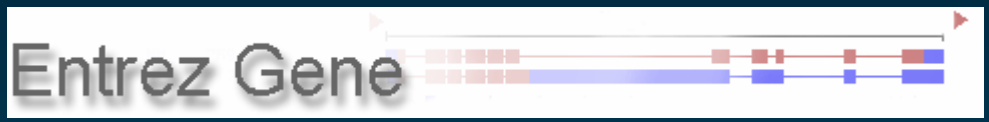
❸ Mapping / Integration

- 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

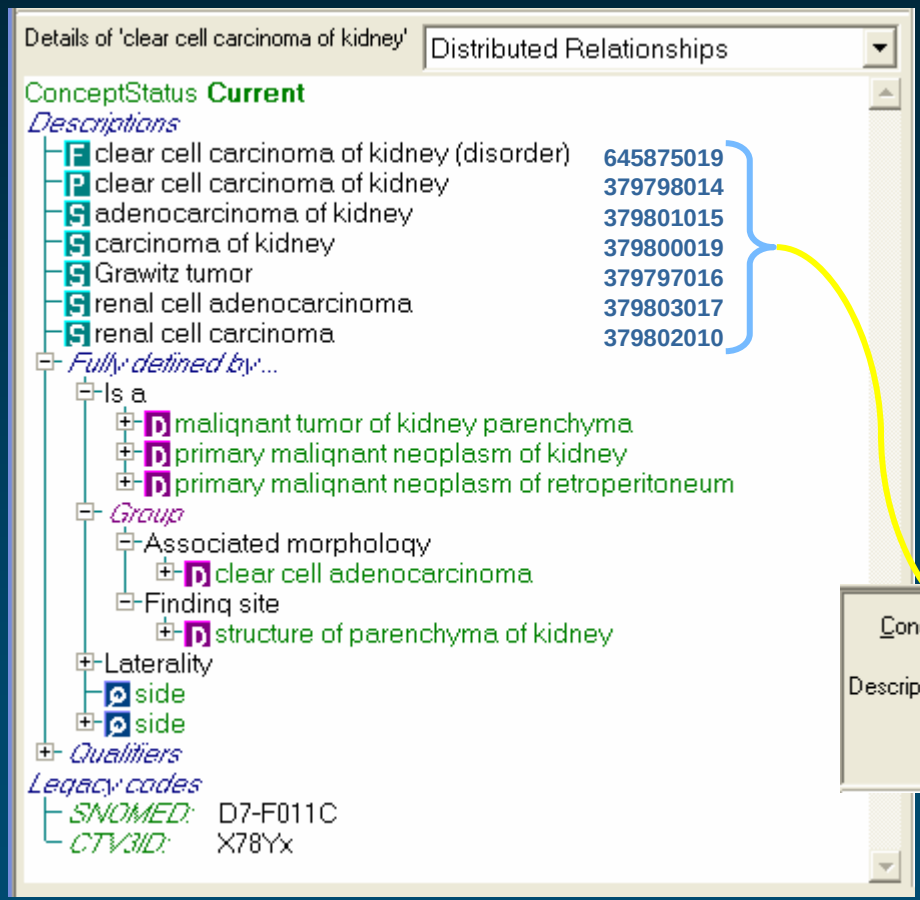


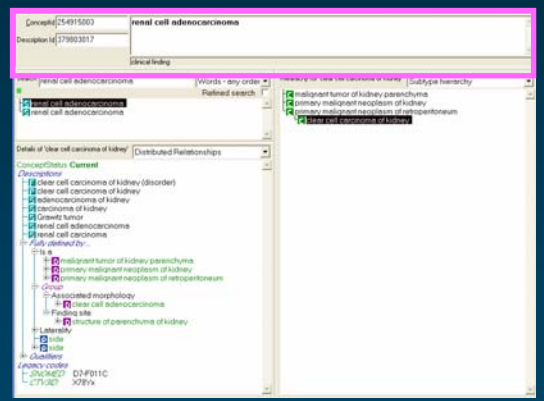
http://www.ncbi.nlm.nih.gov/sites/entrez

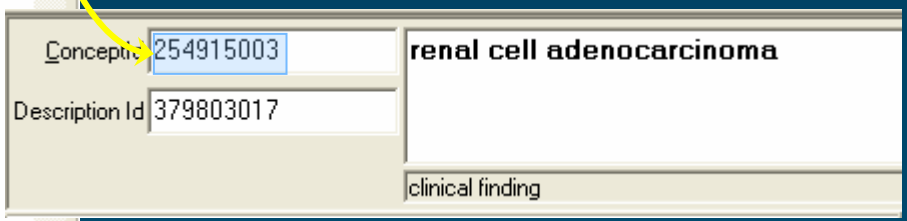
DMD — Order cDNA clone, Links

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







http://www.clinical-info.co.uk/

Bottom-up terminology integration

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.»

Source Vocabularies (2007AC)

- 141 source vocabularies
  - 17 languages
- Broad coverage of biomedicine
  - 6.1M names
  - 1.5M concepts
  - 8M relations
- Common presentation

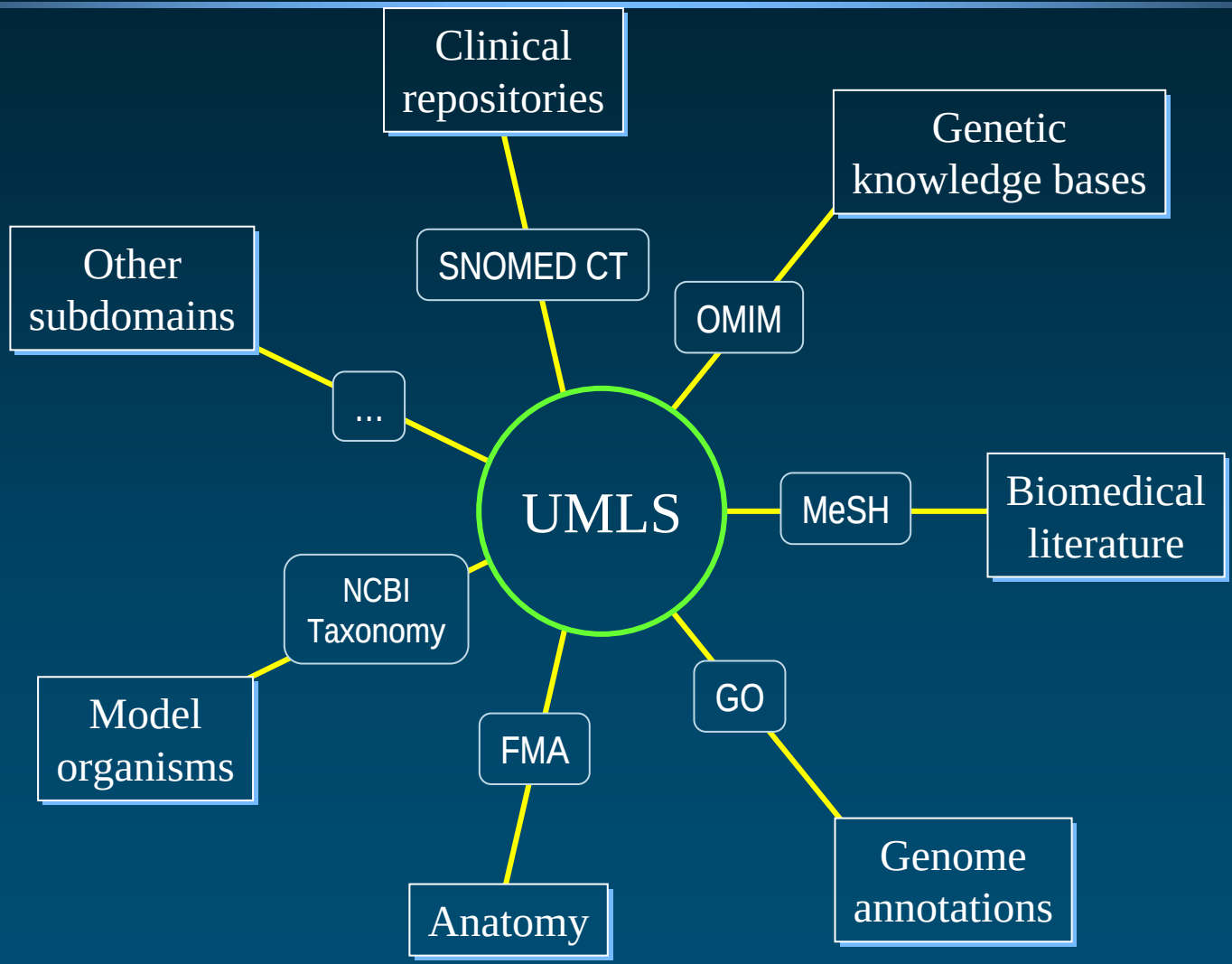
Biomedical terminologies in UMLS

- 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)

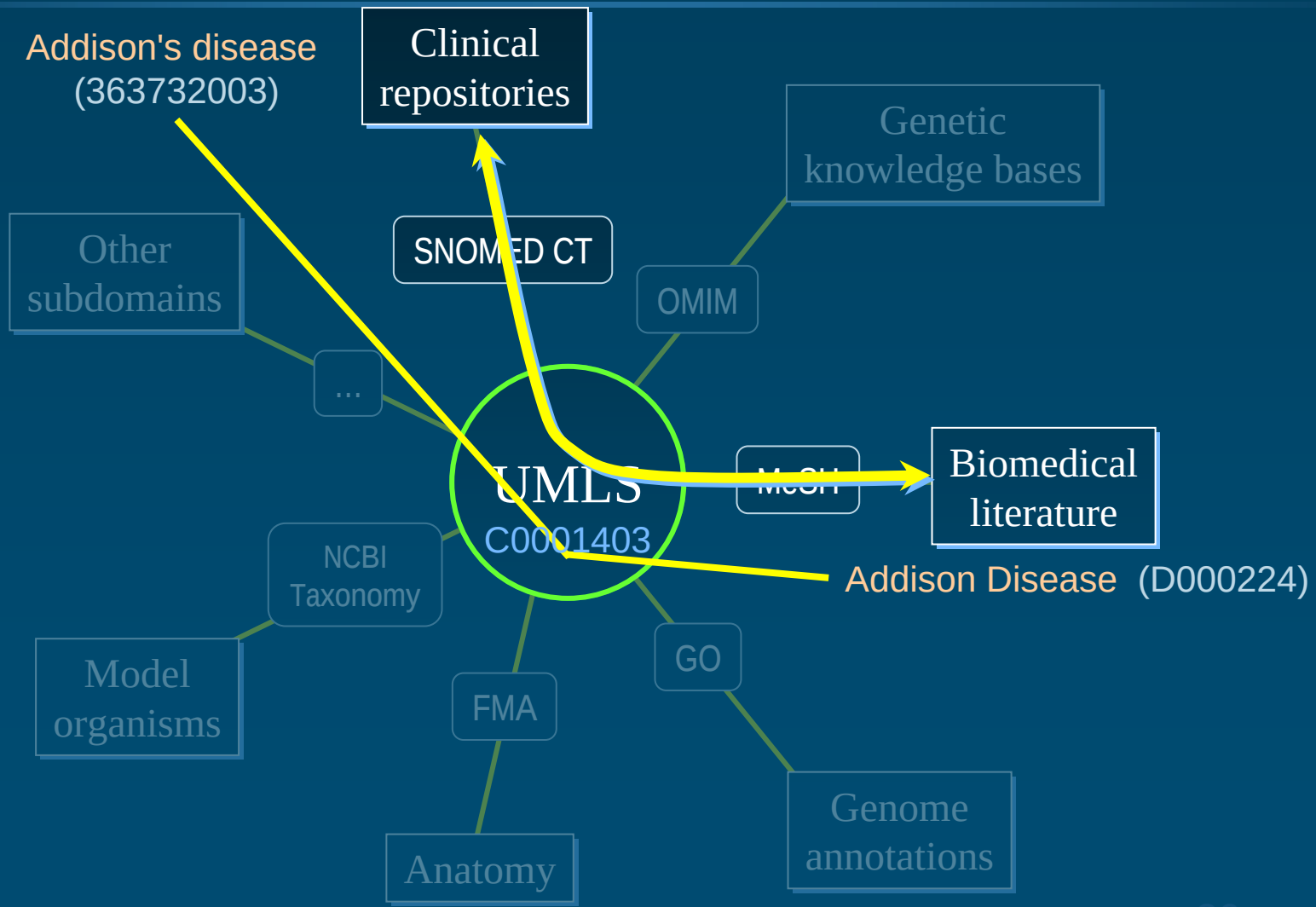
Biomedical terminologies in UMLS

- 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)

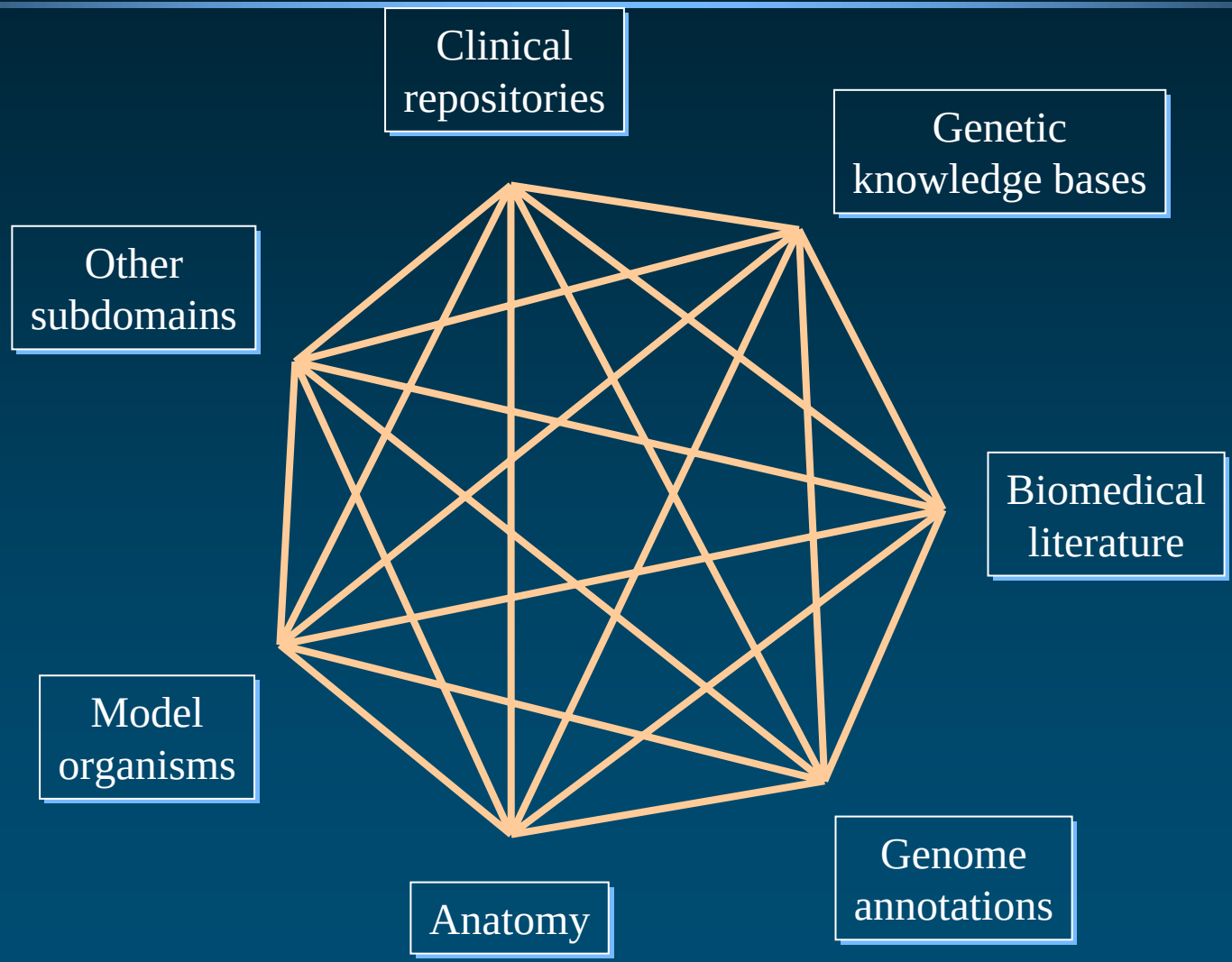
Integrating subdomains



Trans-namespace integration



Integrating subdomains



Top-down terminology integration

OBO Foundry ontologies

Open Biological Ontologies

- Extended family of the Gene Ontology (GO)
- Collaborative development
  - http://obo.sourceforge.net/
- National Center for Biomedical Ontology
  - http://bioontology.org/
- OBO Foundry
  - http://obofoundry.org/
  - Promote best practices in ontology development
  - 10 inclusion criteria

Open Biological Ontologies (OBO)

Domain	Prefix	File	Format
Biological imaging methods	FBbi	image.obo	obo
Biological process	GO	gene ontology.obo	obo
BRENDA tissue / enzyme source	BTO	BrendaTissue.obo	obo
C. elegans development	WBls	worm development.obo	obo
C. elegans gross anatomy	WBbt		obo
C. elegans phenotype	WBPhenotype	phenotype_ontology_obo.cgi	obo
Cell type	CL	cell.obo	obo
Cellular component	GO	gene ontology.obo	obo
Cereal plant development	GRO	cereals development.obo	obo
Cereal plant trait	TO	plant trait.obo	obo
Chemical entities of biological interest	CHEBI	chebi.obo	obo
Common Anatomy Reference Ontology	CARO	caro.obo	obo
Dictyostelium discoideum anatomy	DDANAT	dictyostelium anatomy.obo	obo
Drosophila development	FBdv	fly development.obo	obo
Drosophila gross anatomy	FBbt	fly anatomy.obo	obo
Environment Ontology	ENVO	envo.obo	obo

http://obo.sourceforge.net/

Integrating subdomains

RELATION TO TIME / GRANULARITY	CONTINUANT				OCCURRENT
	INDEPENDENT		DEPENDENT		
ORGAN AND ORGANISM	**Organism** (NCBI Taxonomy?)	**Anatomical Entity** (FMA, CARO)	**Organ Function** (FMP, CPRO)	**Phenotypic Quality** (PaTO)	**Biological Process** (GO)
CELL AND CELLULAR COMPONENT	**Cell** (CL)	**Cellular Component** (FMA, GO)	**Cellular Function** (GO)		
MOLECULE	**Molecule** (ChEBI, SO, RnaO, PrO)		**Molecular Function** (GO)		Molecular Process (GO)

(Barry Smith)

OBO ontologies Examples

- Gene Ontology
- Cell types
- Sequence Ontology
- ChEBI
- Foundational Model of Anatomy
- PATO – phenotypic qualities
- Relationship types
- Ontology for Biomedical Investigations

Applications

Biomedical Semantic Web

W3C Health Care and Life Sciences IG

W3C Technology and Society domain | Semantic Web Activity

W3C Semantic Web Health Care and Life Sciences Interest Group

The Semantic Web Health Care and Life Sciences Interest Group is designed to improve collaboration, research and development, and innovation adoption in the health care and life science industries. Aiding decision-making in clinical research, Semantic Web technologies will bridge many forms of biological and medical information across institutions.

Contents: Mission and Scope | Membership and Joining | Charter /History | Resources | Presentations | Articles | New and Events | Conferences | Task Forces

Nearby: Discussion archive | HCLS WIKI | Applications and Demonstrations | OWL | RDF Data Access | Rules | Semantic Web Best Practices and Deployment

Introduction

Both Life Science Research and Health Care are areas undergoing phenomenal growth, holding much promise for our future as long as we can manage and apply the new knowledge gained without driving up costs. Key to their success is the implementation of new informatics models that will unite many forms of biological and medical information across all institutions, through the encoding of meaning into the data and their interpretations. By focusing on the semantics of information, researchers will have more access to the knowledge required to effectively find cures to diseases, while doctors will have better tools for individualized clinical management of patients.

Mission and Scope

The Semantic Web for Health Care and Life Sciences Interest Group (HCLSIG) is chartered to develop and support the use of Semantic Web technologies and practices to improve collaboration, research and development, and innovation adoption in the of Health Care and Life Science domains. Success in these domains depends on a foundation of semantically rich system, process and information interoperability. (more).

News and Events

- Last Call: SPARQL Query Language for RDF 2007-03-27: : Comments are due by 18 April.(Permalink)
- HCLS demo, planned for WWW2007 in Banff. To help participate in the demo, please contact Alan Ruttenberg.
- FIRST INTERNATIONAL WORKSHOP ON HEALTH CARE AND LIFE SCIENCES DATA INTEGRATION FOR THE SEMANTIC WEB, May 8, WWW2007 in Banff.
- Eric Prud'hommeaux, new W3C staff contact for HCLS.
- GRDDL links Microformats and Semantic Web: Working Draft xmlns="http://www.w3.org/2000/svg"(Permalink)

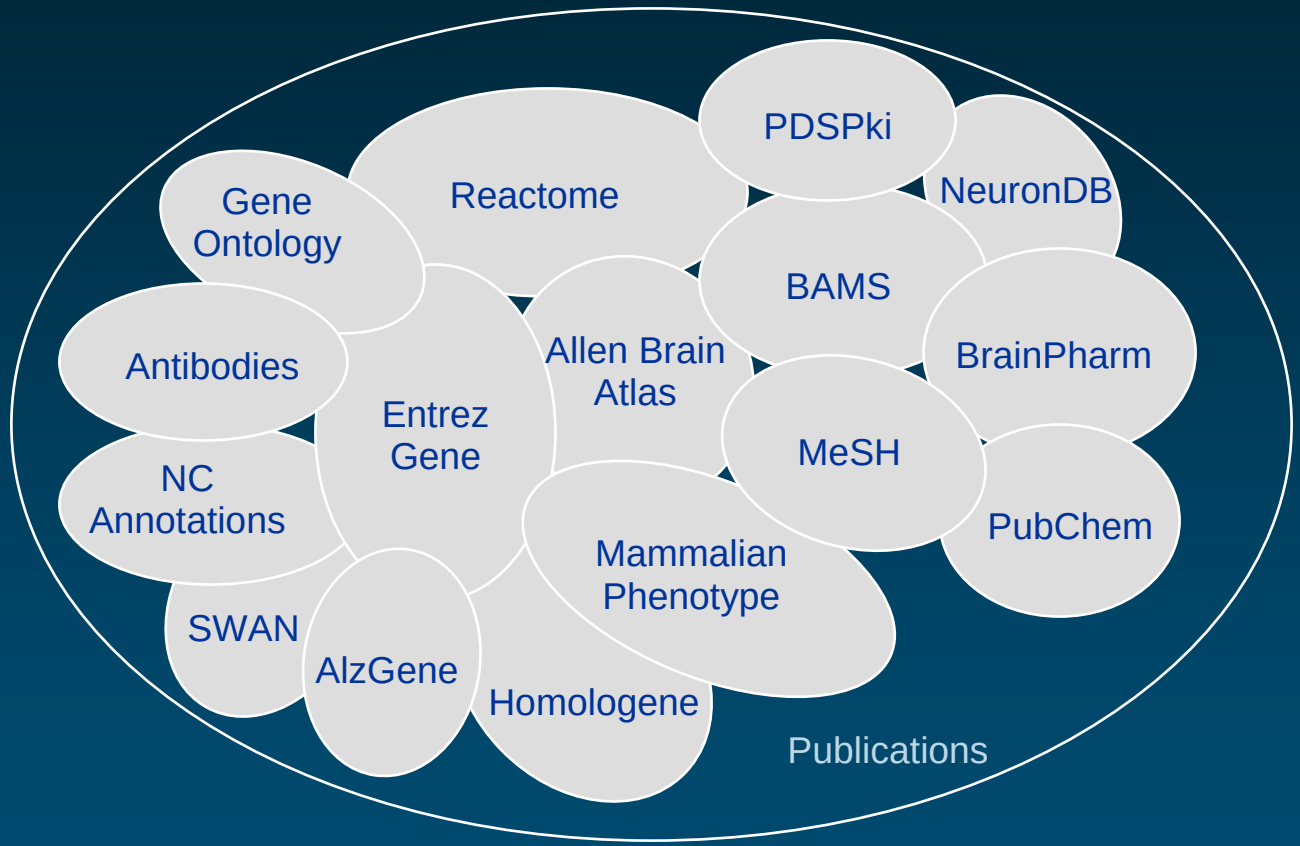
http://www.w3.org/2001/sw/hcls/

Biomedical Semantic Web

- Integration
  - Data/Information
  - E.g., translational research
- Hypothesis generation
- Knowledge discovery

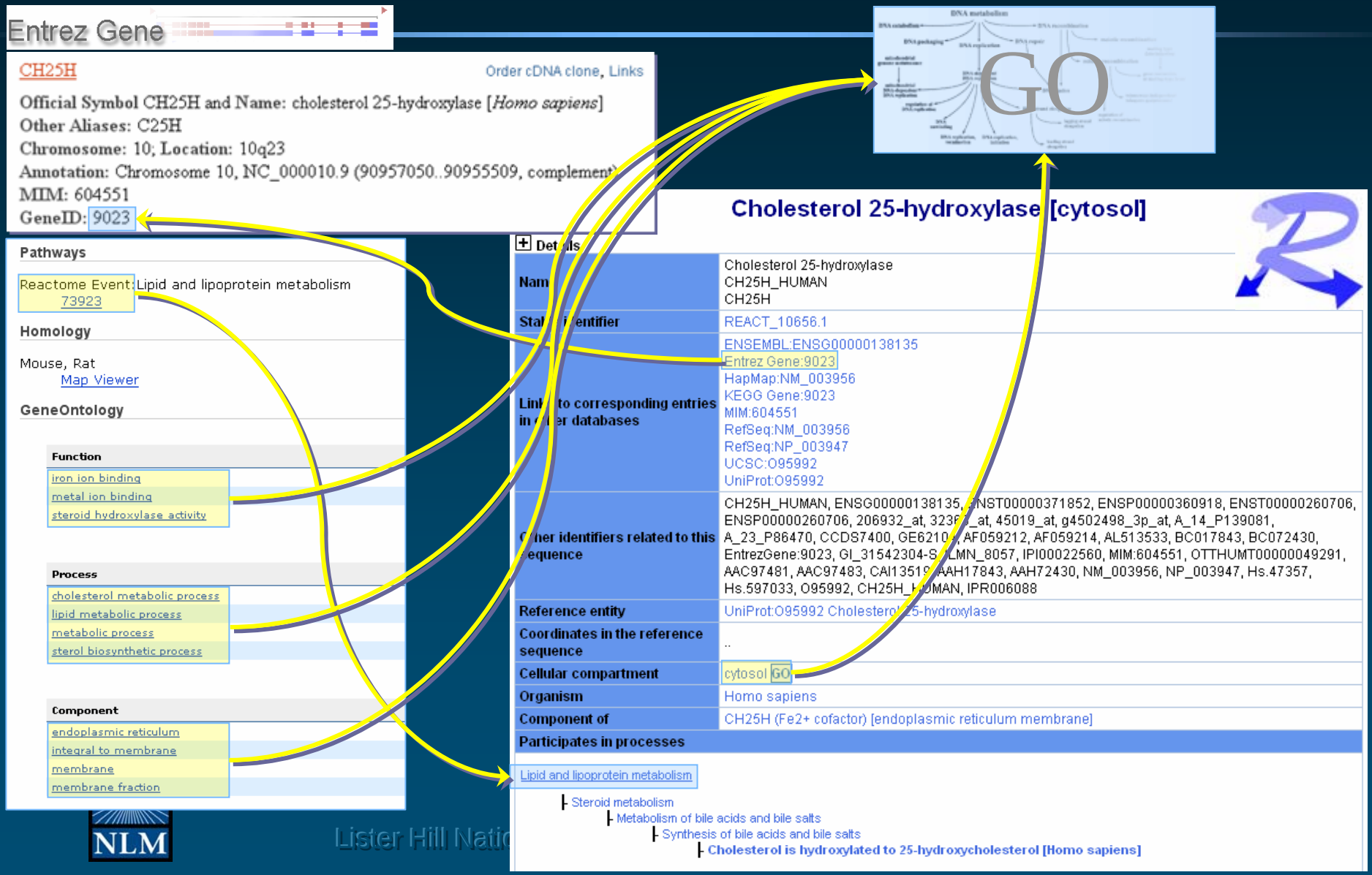
- Clinical data
  - Aggregation, sharing, exchange
  - Support for clinical decision

HCLS mashup of biomedical sources

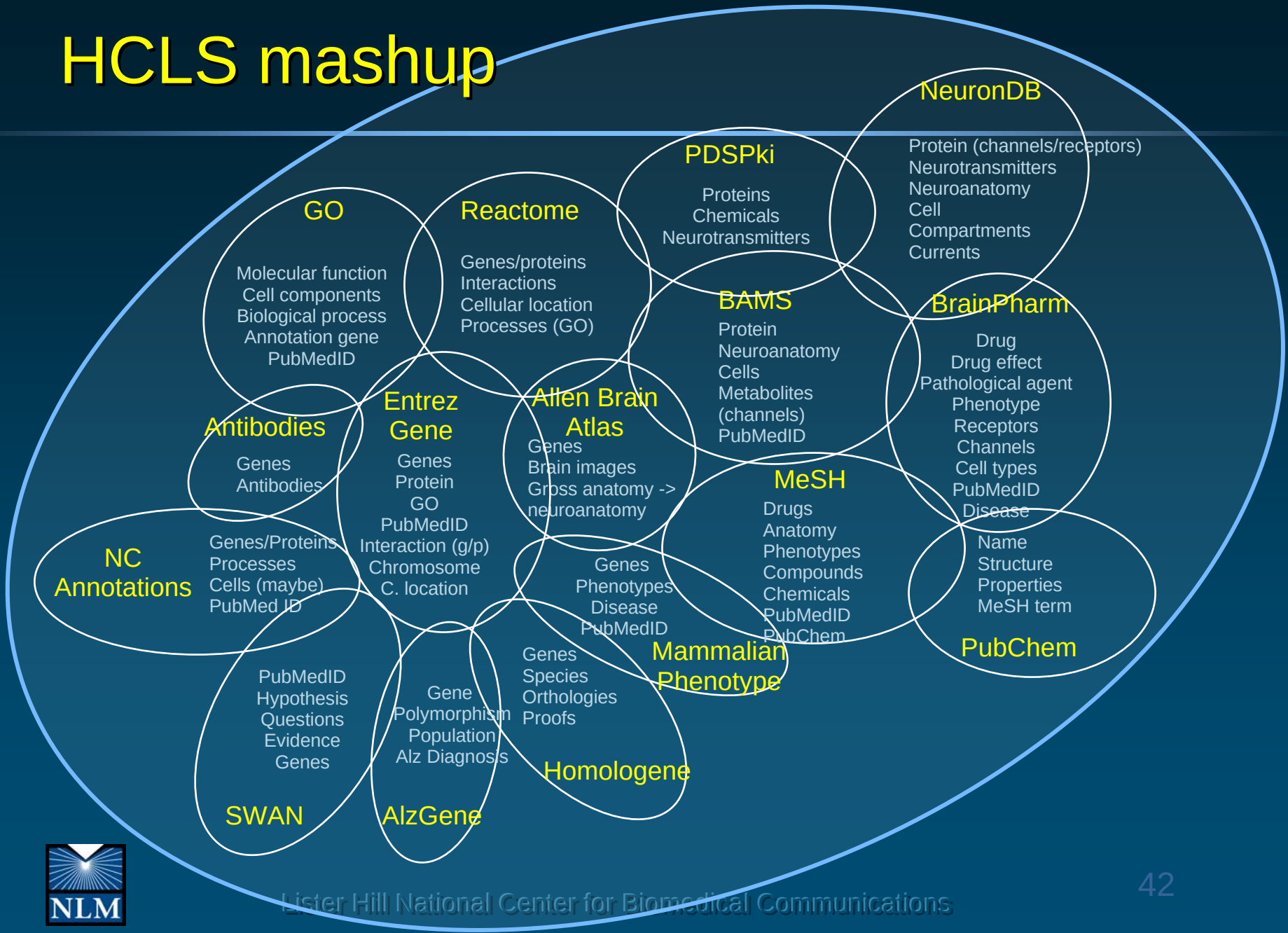


http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo

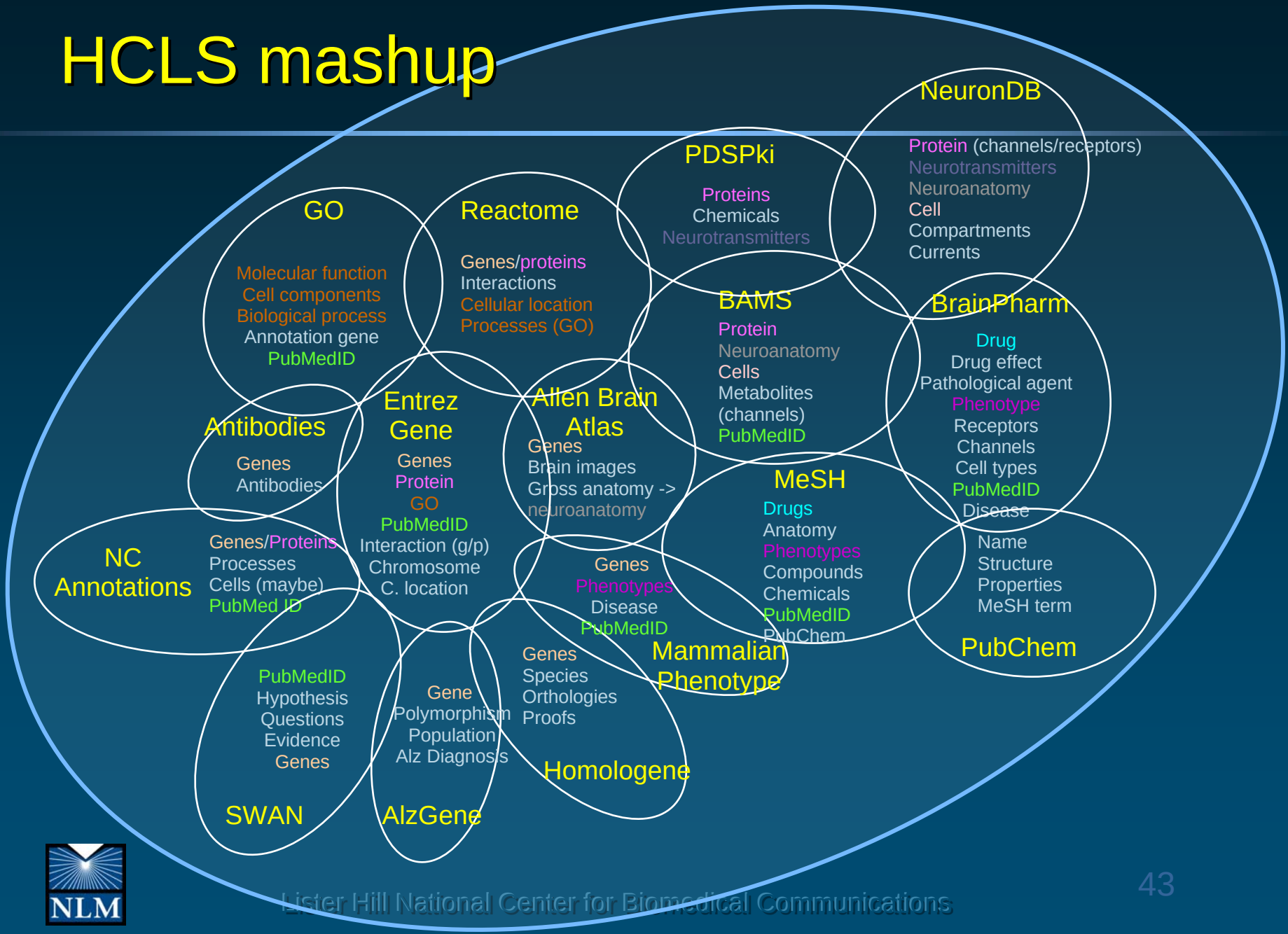
Shared identifiers Example



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
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

Some unresolved issues

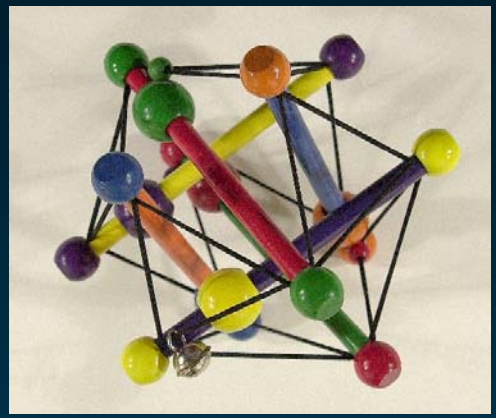
- Format
  - RDF/S, OWL, SKOS vs. OBO, RRF, etc.
  - Converters
- Permanent identification of biomedical entities
  - Syntax: URI vs. LSID
  - Semantic: Trans-namespace identification
- Availability, openness
- Governance, trust

Summary

- Terminologies/Ontologies provide
  - Lists of entities
  - Names for entities
  - Identifiers for entities
- Additionally
  - Information model for integration
  - Trans-namespace resolution
  - Support for inference

Future directions

- Information integration
  - Knowledge extracted from text
  - Knowledge in structured knowledge bases
- Ontologies for relations
  - In complement to ontologies for entities
  - To support reasoning



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