



Special Library Association
Washington, DC
June 15, 2009



Challenges and Promises of the Semantic Web in Health Care and Life Sciences



Olivier Bodenreider

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

Outline

- ◆ Semantic Web
- ◆ Semantic Web *for Health Care and Life Sciences*
- ◆ Promises
- ◆ Challenges
- ◆ Semantic Web and medical libraries



Semantic Web



Quick introduction

Defining the Semantic Web

- ◆ “The Semantic Web
 - ... is an extension of the current web in which
 - ... information is given well-defined meaning,
 - ... better enabling computers and people to work in cooperation.”



The Semantic Web

Tim Berners-Lee, James Hendler and Ora Lassila

Scientific American, May 2001

<http://www.scientificamerican.com/article.cfm?id=the-semantic-web>



From linking documents to linking data

◆ Original WWW

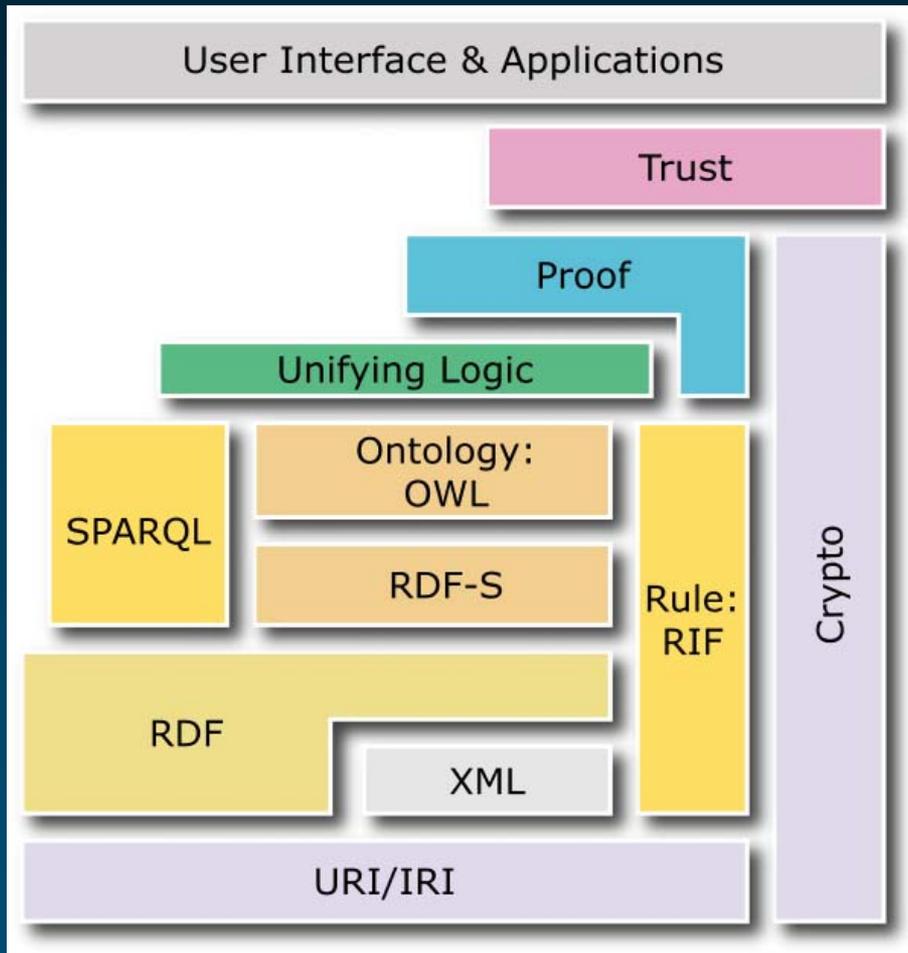
- Web of documents
- Links among documents
- No semantics in the links
- For humans

◆ Semantic Web

- Web of data
- Links among concepts in resources
- Links have an explicit semantics
- For humans and machines



Set of enabling technologies and tools



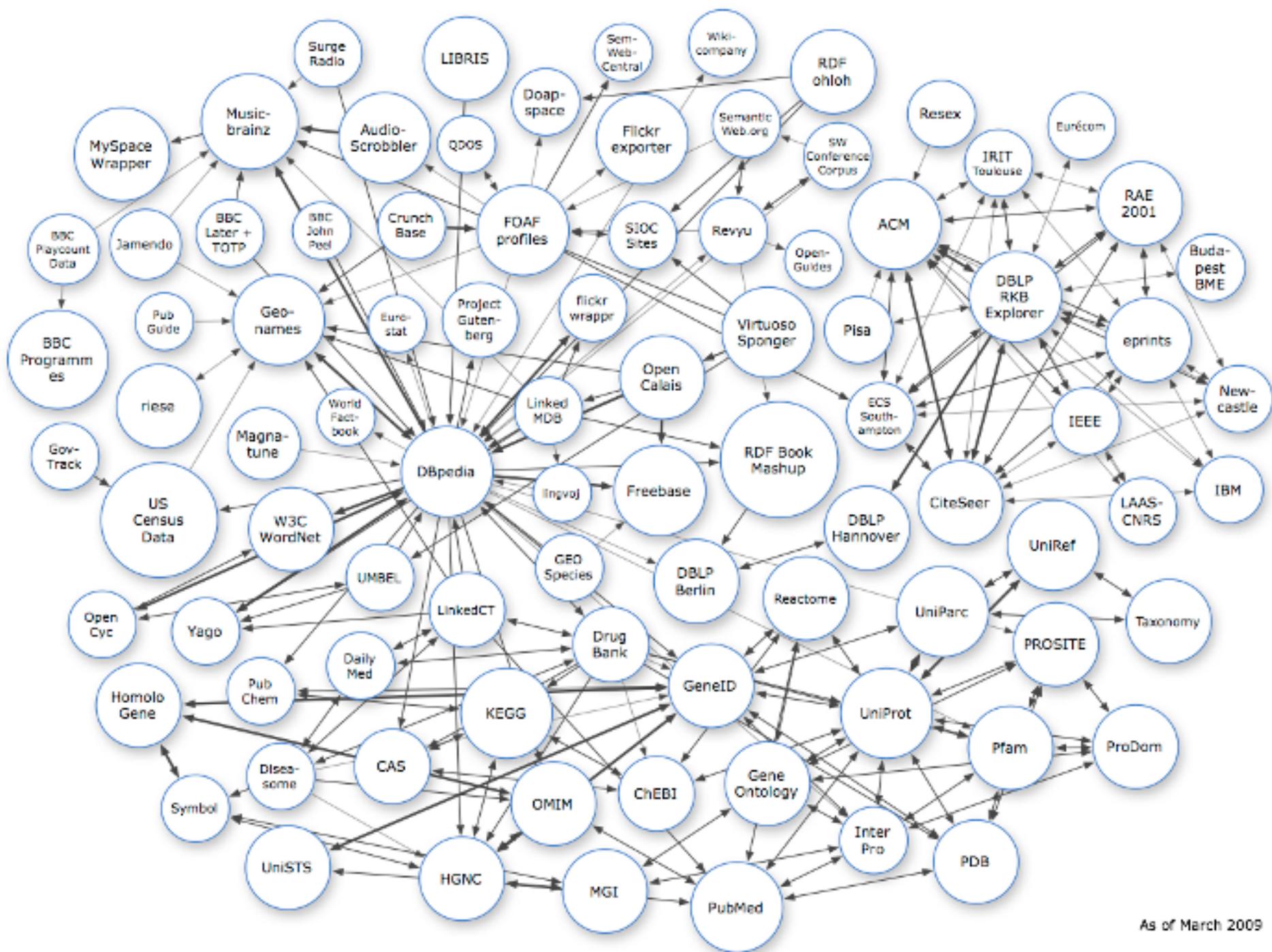
- ◆ Unambiguous names for resources and entities
- ◆ Standard representation for explicit links among entities
- ◆ Shared conceptualization of the domain
- ◆ Standard tools for querying data
- ◆ Standard mechanisms for reasoning over data

Semantic Web in 2009 Standards

◆ Standards

- Uniform Resource Identifiers (URI)
- Resource Description Framework (RDF)
- Web Ontology Language (OWL)
- SPARQL query language
- Rule Interchange Format (RIF)





Linked data

- ◆ Resources available in RDF
 - Unique, unambiguous identifiers for entities
 - Explicit relations among entities
- ◆ Links across resources (federation)
 - Enabled by
 - Shared identifiers across resources
 - Global identifiers, resolvable on the web

Semantic Web
for Health Care and Life Sciences

The original Semantic Web scenario

- ◆ “Mom needs to see a specialist and then has to have a series of physical therapy sessions. [...] I'm going to have my agent set up the appointments.”
- ◆ “The agent promptly retrieved information about Mom's *prescribed treatment* from the doctor's agent, looked up several lists of *providers*, and checked for the ones *in-plan* for Mom's insurance within a *20-mile radius* of her *home* and with a *rating of excellent* or *very good* on trusted rating services.”
- ◆ ...

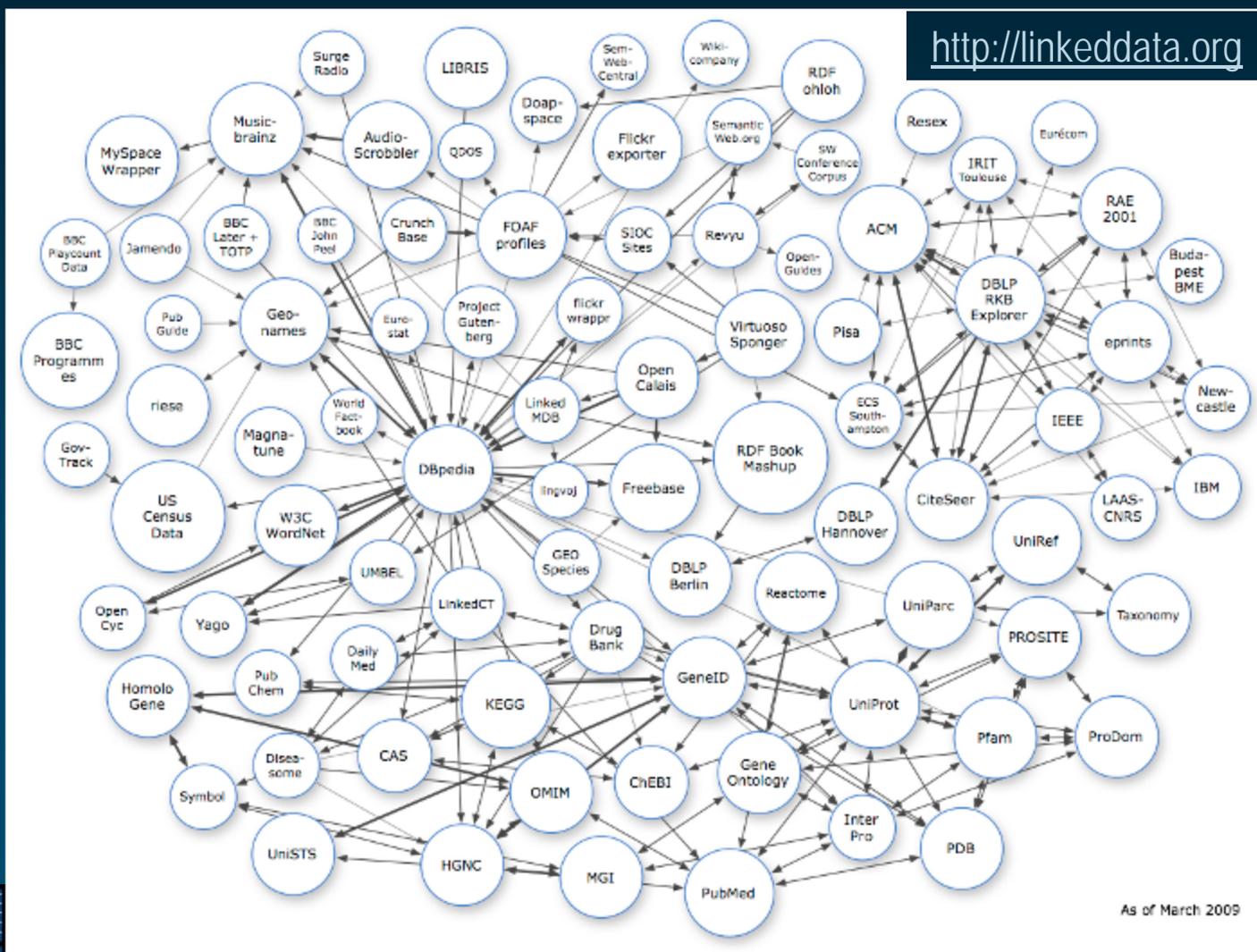


The Semantic Web

Tim Berners-Lee, James Hendler and Ora Lassila

Scientific American, May 2001

From linked data...



W3C Health Care and Life Sciences IG



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

Semantic Web Health Care and Life Sciences (HCLS) Interest Group

Introduction

The **mission** of the Semantic Web Health Care and Life Sciences Interest Group, part of the [Semantic Web Activity](#), is to develop, advocate for, and support the use of Semantic Web technologies for biological science, translational medicine and health care. These domains stand to gain tremendous benefit by adoption of Semantic Web technologies, as they depend on the interoperability of information from many domains and processes for efficient decision support.

The group will:

- ◆ Document use cases to aid individuals in understanding the business and technical benefits of using Semantic Web technologies.
- ◆ Document guidelines to accelerate the adoption of the technology.
- ◆ Implement a selection of the use cases as proof-of-concept demonstrations.
- ◆ Explore the possibility of developing high level vocabularies.
- ◆ Disseminate information about the group's work at government, industry, and academic events.



W3C Health Care and Life Sciences IG

- ◆ Formed in 2005
 - Re-chartered in 2008
- ◆ Broad industry participation
 - Over 100 members
 - Mailing list of over 600
- ◆ Get involved!
 - Team contact: Eric Prud'hommeaux <eric@w3.org>



HCLSIG activities

- ◆ Document use cases to aid individuals in understanding the business and technical benefits of using Semantic Web technologies
- ◆ Document guidelines to accelerate the adoption of the technology
- ◆ Implement a selection of the use cases as proof-of-concept demonstrations
- ◆ Develop high-level vocabularies
- ◆ Disseminate information about the group's work at government, industry, and academic events



HCLSIG task forces

- ◆ **BioRDF**
 - Use of RDF (and OWL) to represent biomedical data
- ◆ **Clinical Observations Interoperability**
 - Enable the reuse of clinical research and clinical practice data
- ◆ **Linking Open Drug Data**
 - Linking drug information sources
- ◆ **Translational Medicine Ontology**
 - Development of an ontology for biomedical data integration
- ◆ **Scientific Discourse**
 - Annotate information extracted from the scientific literature
- ◆ **Terminology**
 - Use of SW technologies for representing existing biomedical terminologies



Promises of the Semantic Web for Health Care and Life Sciences

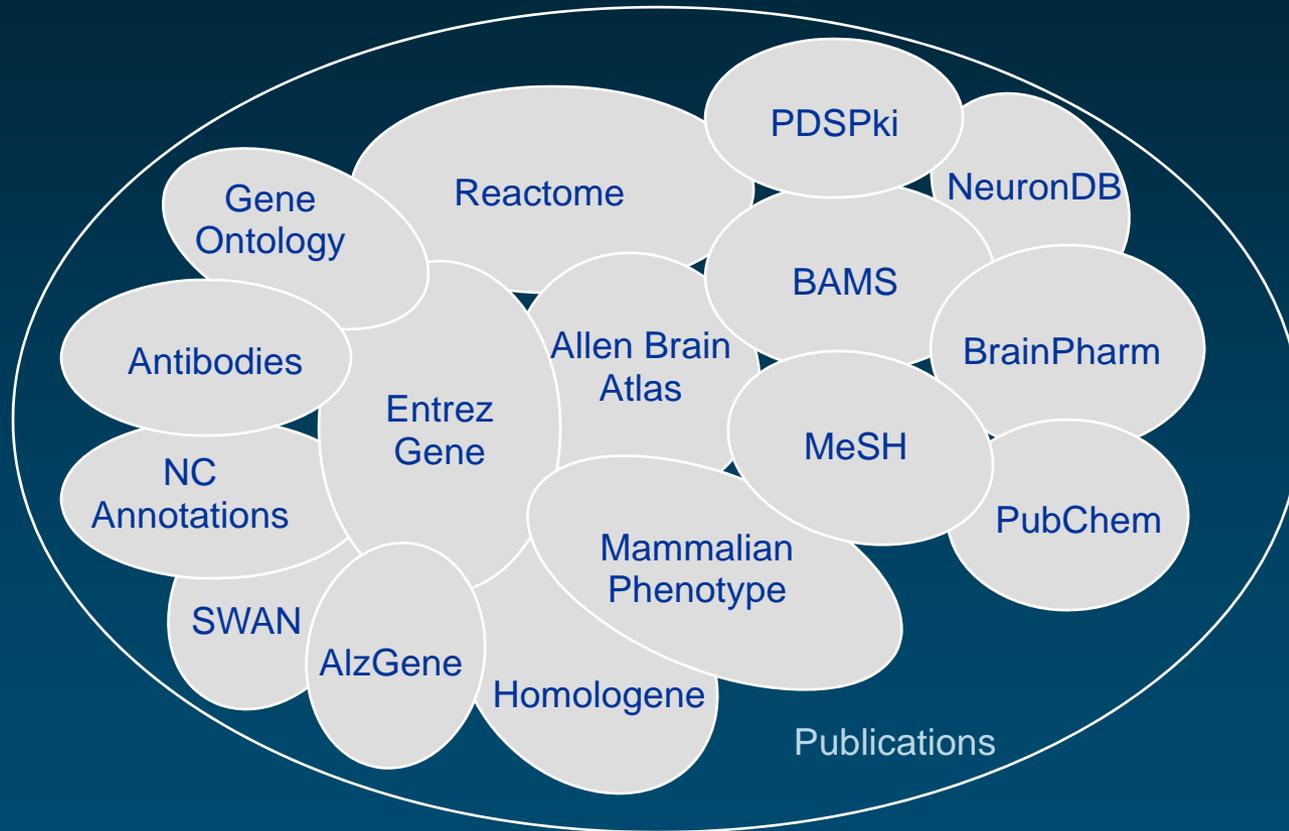
Biomedical Semantic Web

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

[Ruttenberg, BMC Bioinf. 2007]



HCLS mashup of biomedical sources



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



Shared identifiers Example

Entrez Gene

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]

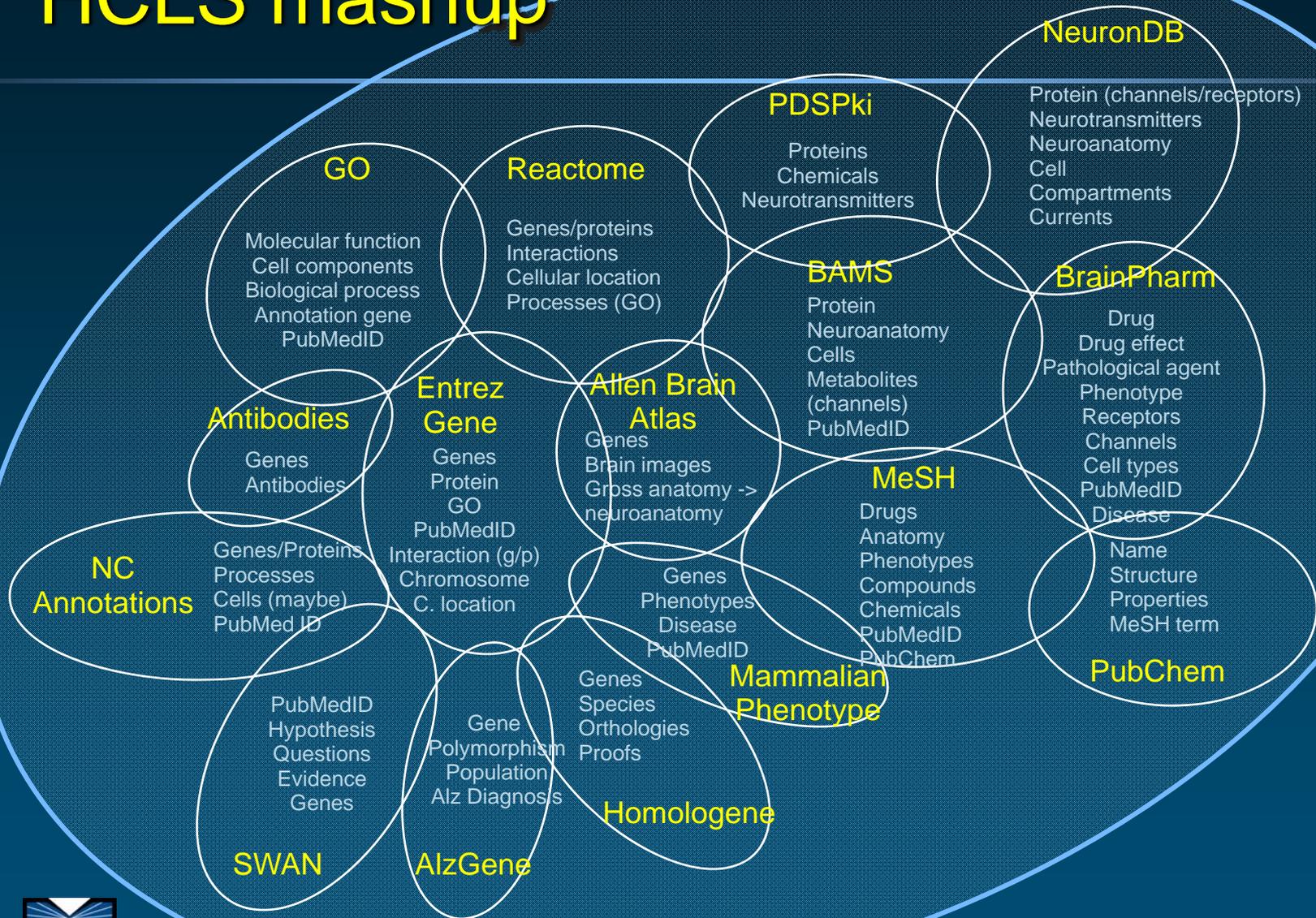


| | |
|--|--|
| Name | Cholesterol 25-hydroxylase CH25H_HUMAN CH25H |
| Stable identifier | REACT_10656.1 ENSEMBL:ENSG00000138135 Entrez Gene 9023 |
| Links to corresponding entries in other databases | HapMap:NM_003956 KEGG Gene:9023 MIM:604551 RefSeq:NM_003956 RefSeq:NP_003947 UCSC:O95992 UniProt:O95992 |
| Other identifiers related to this sequence | CH25H_HUMAN, ENSG00000138135, ENST00000371852, ENSP00000360918, ENST00000260706, ENSP00000260706, 206932_at, 3236_at, 45019_at, g4502498_3p_at, A_14_P139081, A_23_P86470, CCDS87400, GE6210, AF059212, AF059214, AL513533, BC017843, BC072430, EntrezGene:9023, GI_31542304-S, LMN_8057, IP100022560, MIM:604551, OTTHUMT00000049291, AAC97481, AAC97483, CAI13519, SAH17843, 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 |
| | <ul style="list-style-type: none"> - Steroid metabolism <ul style="list-style-type: none"> - Metabolism of bile acids and bile salts <ul style="list-style-type: none"> - Synthesis of bile acids and bile salts <ul style="list-style-type: none"> - Cholesterol is hydroxylated to 25-hydroxycholesterol [Homo sapiens] |

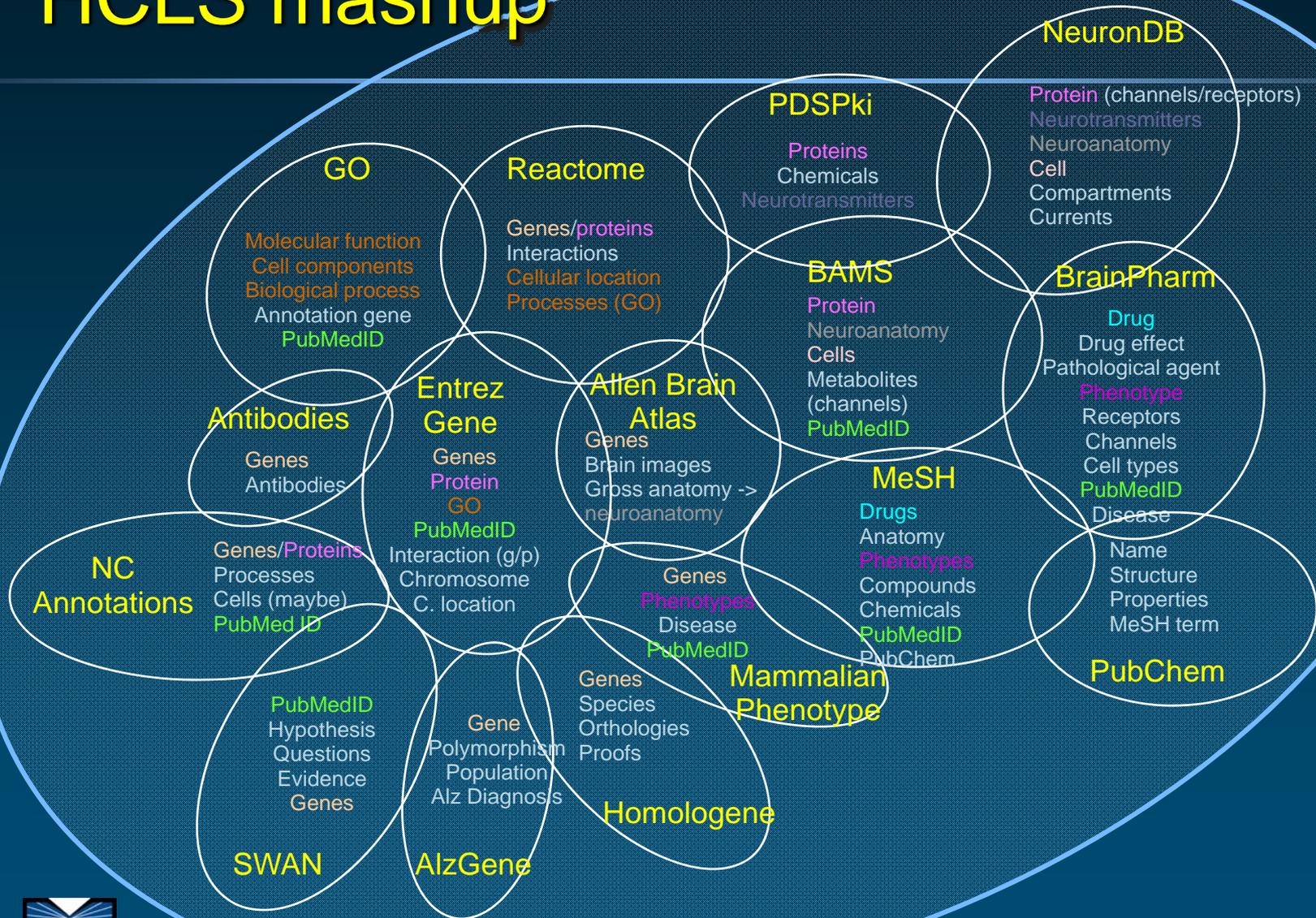


Lister Hill National

HCLS mashup



HCLS mashup



HCLS mashups

- ◆ Based on RDF/OWL
- ◆ Based on shared identifiers
 - “Recombinant data” (E. Neumann)
- ◆ Ontologies used in some cases
- ◆ Support applications (SWAN, SenseLab, etc.)

- ◆ Journal of Biomedical Informatics
special issue on Semantic bio-mashups
[[J. Biomedical Informatics 41\(5\) 2008](#)]



Semantic bio-mashups

- ◆ Bio2RDF: Towards a mashup to build bioinformatics knowledge systems
- ◆ Identifying disease-causal genes using Semantic Web-based representation of integrated genomic and phenomic knowledge
- ◆ Schema driven assignment and implementation of life science identifiers (LSIDs)
- ◆ The SWAN biomedical discourse ontology
- ◆ An ontology-driven semantic mashup of gene and biological pathway information: Application to the domain of nicotine dependence
- ◆ Towards an ontology for sharing medical images and regions of interest in neuroimaging
- ◆ yOWL: An ontology-driven knowledge base for yeast biologists
- ◆ Dynamic sub-ontology evolution for traditional Chinese medicine web ontology
- ◆ Ontology-centric integration and navigation of the dengue literature
- ◆ Infrastructure for dynamic knowledge integration—Automated biomedical ontology extension using textual resources
- ◆ An ontological knowledge framework for adaptive medical workflow
- ◆ Semi-automatic web service composition for the life sciences using the BioMoby semantic web framework
- ◆ Combining Semantic Web technologies with Multi-Agent Systems for integrated access to biological resources

[J. Biomedical Informatics 41(5) 2008]



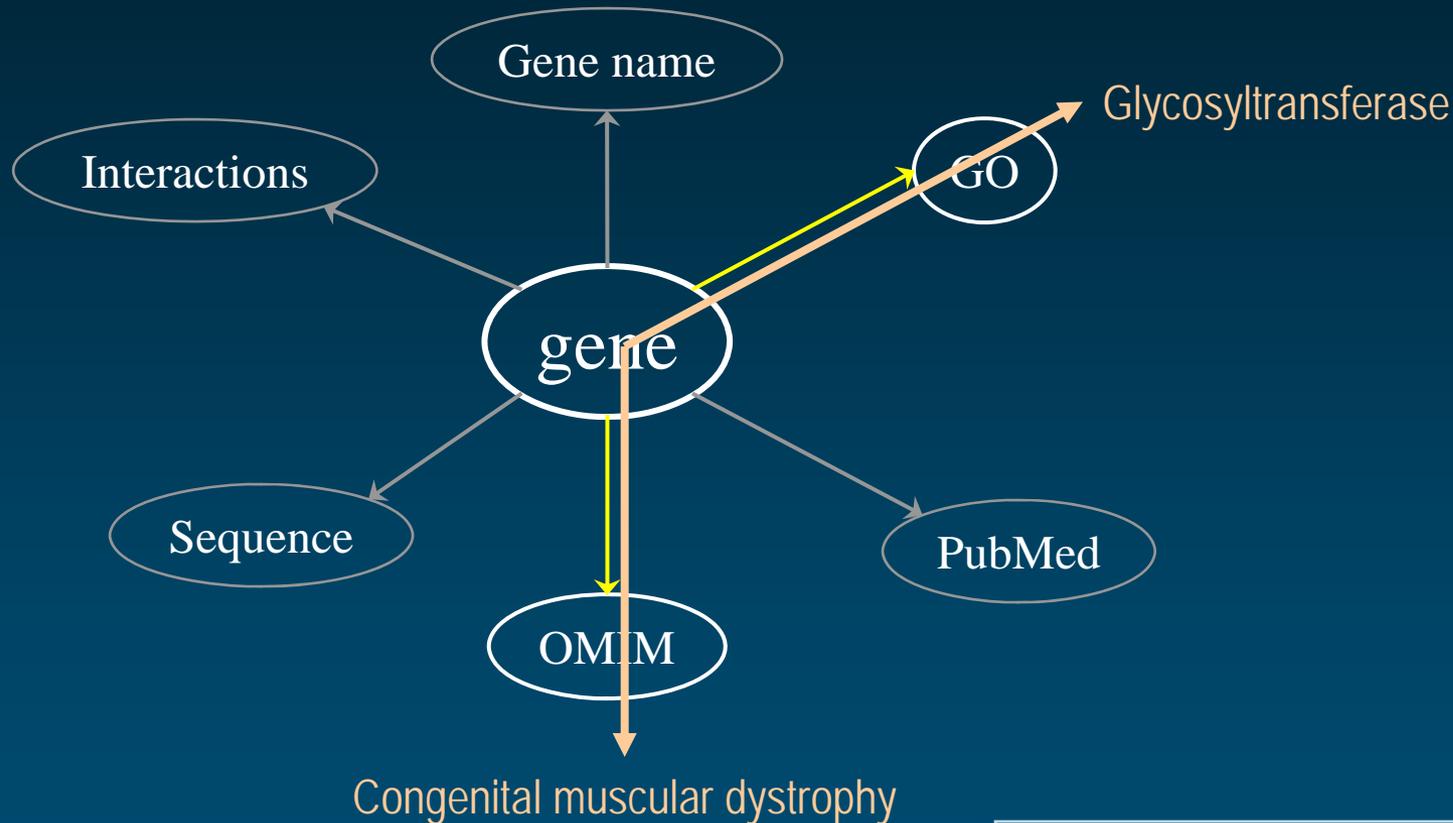
Bio2RDF

- ◆ “Semantic web atlas of postgenomic knowledge”
- ◆ 65M “triples”
- ◆ 34 biomedical sources
- ◆ Publicly available
- ◆ Distributed environment

Bi  ***2RDF.org***



Another mashup example



Link between glycosyltransferase activity and congenital muscular dystrophy?

Search Gene for APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) Go Clear

Limits Preview/Index History Clipboard Details

Display Full Report Show 5 Send to

All: 1 Current Only: 1 Genes Genomes: 1 SNP GeneView: 1

APP
(GeneID: 351)

1: APP amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) [Homo sapiens]
GeneID: 351 Primary source: [HGNC:620](#) updated 26-Jul-2006

Entrez Gene Home

- Table Of Contents
- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- HIV-1 protein interactions
- Interactions
- General gene information
- General protein information
- Reference Sequences
- Related Sequences
- Additional Links
- Links

Summary

Official Symbol: APP **and Name:** amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease) provided by [HUGO Gene Nomenclature Committee](#)
See related: [HPRD:00100](#), [MIM:104760](#)
Gene type: protein coding
Gene name: APP
Gene description: amyloid beta (A4) precursor protein (peptidase nexin-II, Alzheimer disease)

has_protein_name

amyloid beta A4 protein

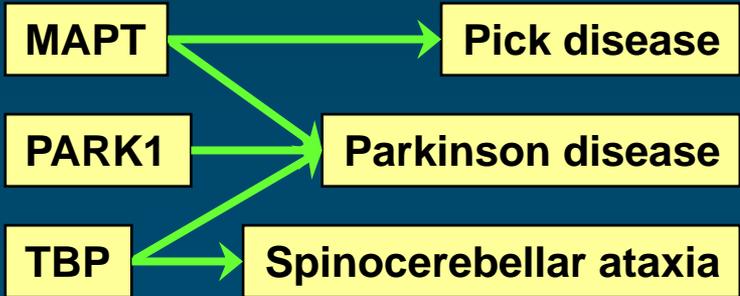
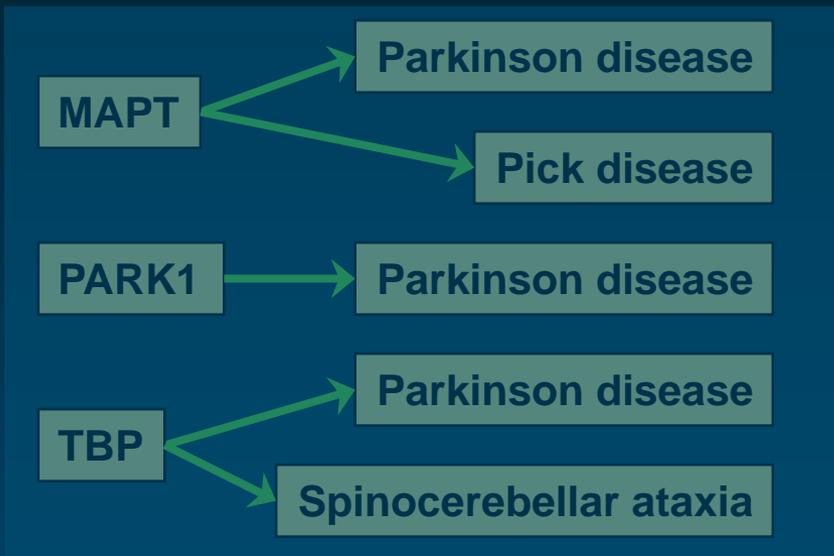
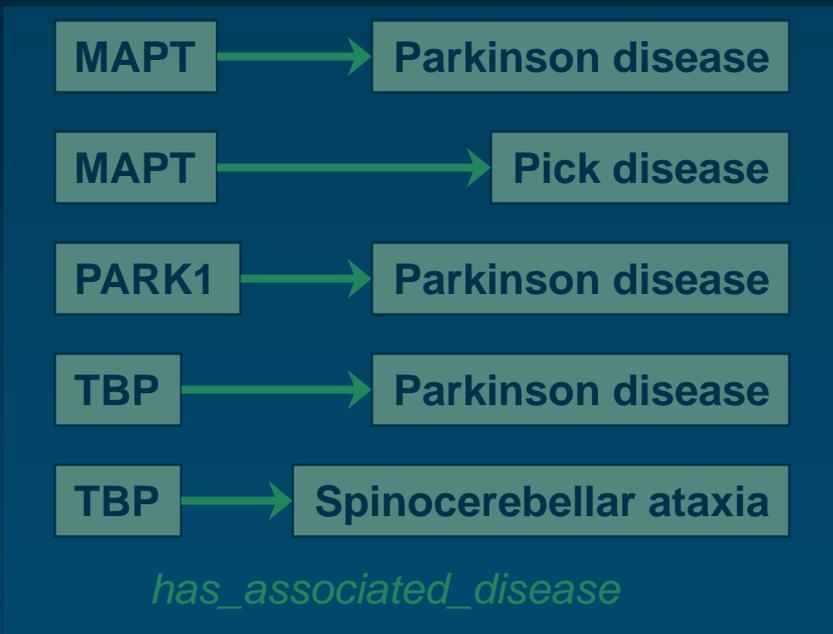
General protein information

Names: amyloid beta A4 protein
protease nexin-II; A4 amyloid protein; amyloid-beta protein; beta-amyloid peptide; cerebral vascular amyloid peptide; amyloid beta (A4) precursor protein (protease nexin-II, Alzheimer disease)

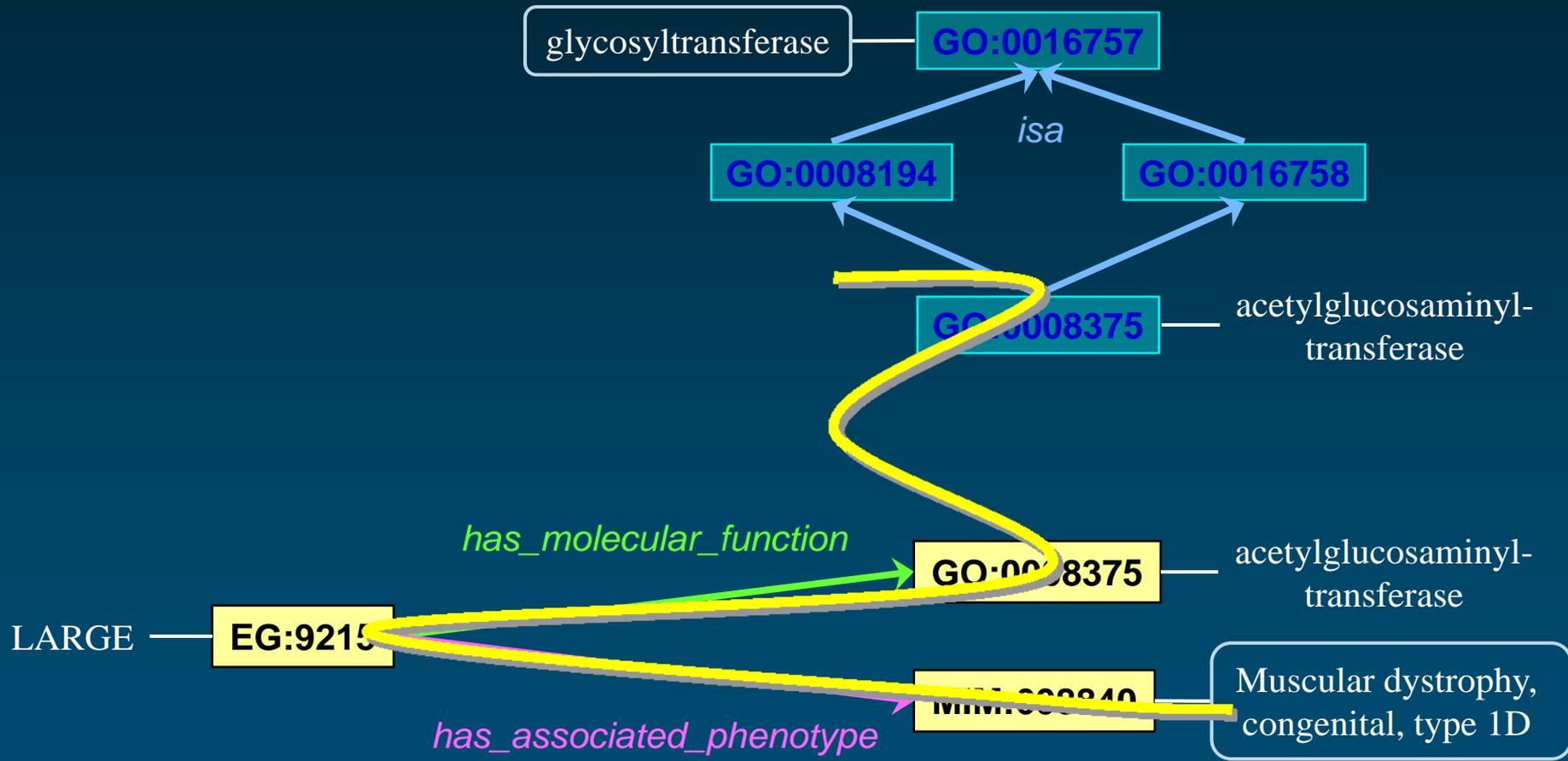
RDF triple Gene property



RDF graph Connecting several genes



From *glycosyltransferase* to congenital muscular dystrophy



Summary of promises

- ◆ Flexibility
 - Easier to federate than relational databases
 - Inherently distributed
- ◆ Based on standards
 - Tooling available
- ◆ Platform for data integration
 - Enabling technology for semantic interoperability, hypothesis generation and knowledge discovery

Challenges of the Semantic Web for Health Care and Life Sciences

Challenging issues

- ◆ Permanent identifiers for biomedical entities
- ◆ Bridges across ontologies
- ◆ Other issues
 - Availability of biomedical datasets
 - Discoverability
 - Formalism
 - Ontology integration
 - Scalability

Challenging issues

Permanent identifiers for biomedical entities

Identifying biomedical entities

- ◆ Multiple identifiers for the same entity in different ontologies
- ◆ Barrier to data integration in general
 - Data annotated to different ontologies cannot “recombine”
 - Need for mappings across ontologies
- ◆ Barrier to data integration in the Semantic Web
 - Multiple possible identifiers for the same entity
 - Depending on the underlying representational scheme (URI vs. LSID)
 - Depending on who creates the URI

Possible solutions

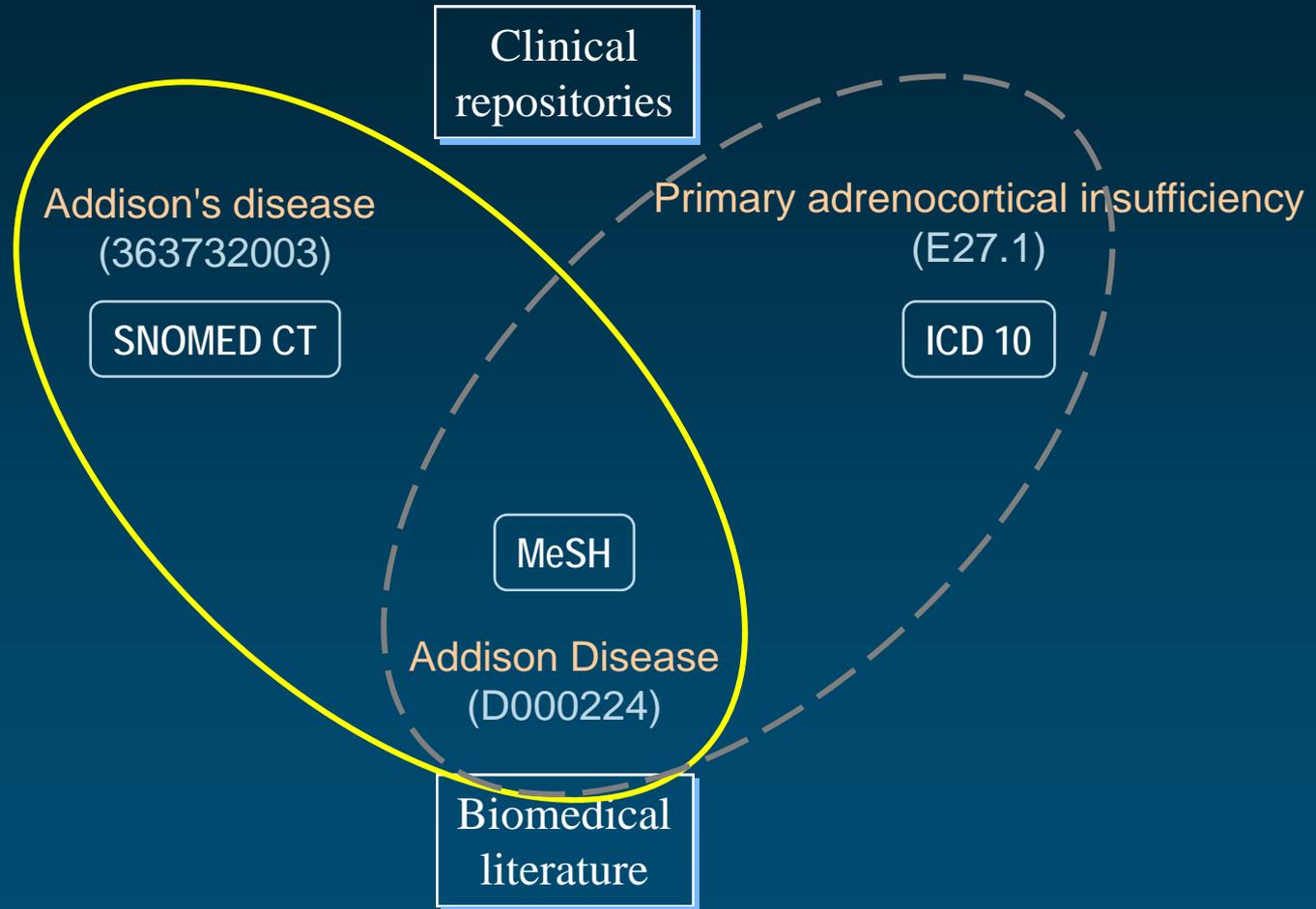
- ◆ PURL <http://purl.org>
 - One level of indirection between developers and users
 - Independence from local constraints at the developer's end
- ◆ The institution creating a resource is also responsible for minting URIs
 - E.g., URI for genes in Entrez Gene
- ◆ Guidelines: “URI note”
 - W3C Health Care and Life Sciences Interest Group
- ◆ Shared names initiative [\[http://sharedname.org/\]](http://sharedname.org/)
 - Identify resources vs. entities



Challenging issues

Bridges across ontologies

Trans-namespace integration



(Integrated) concept repositories

- ◆ Unified Medical Language System

<http://umlsks.nlm.nih.gov>

- ◆ NCBO's BioPortal

<http://www.bioontology.org/tools/portal/bioportal.html>

- ◆ caDSR

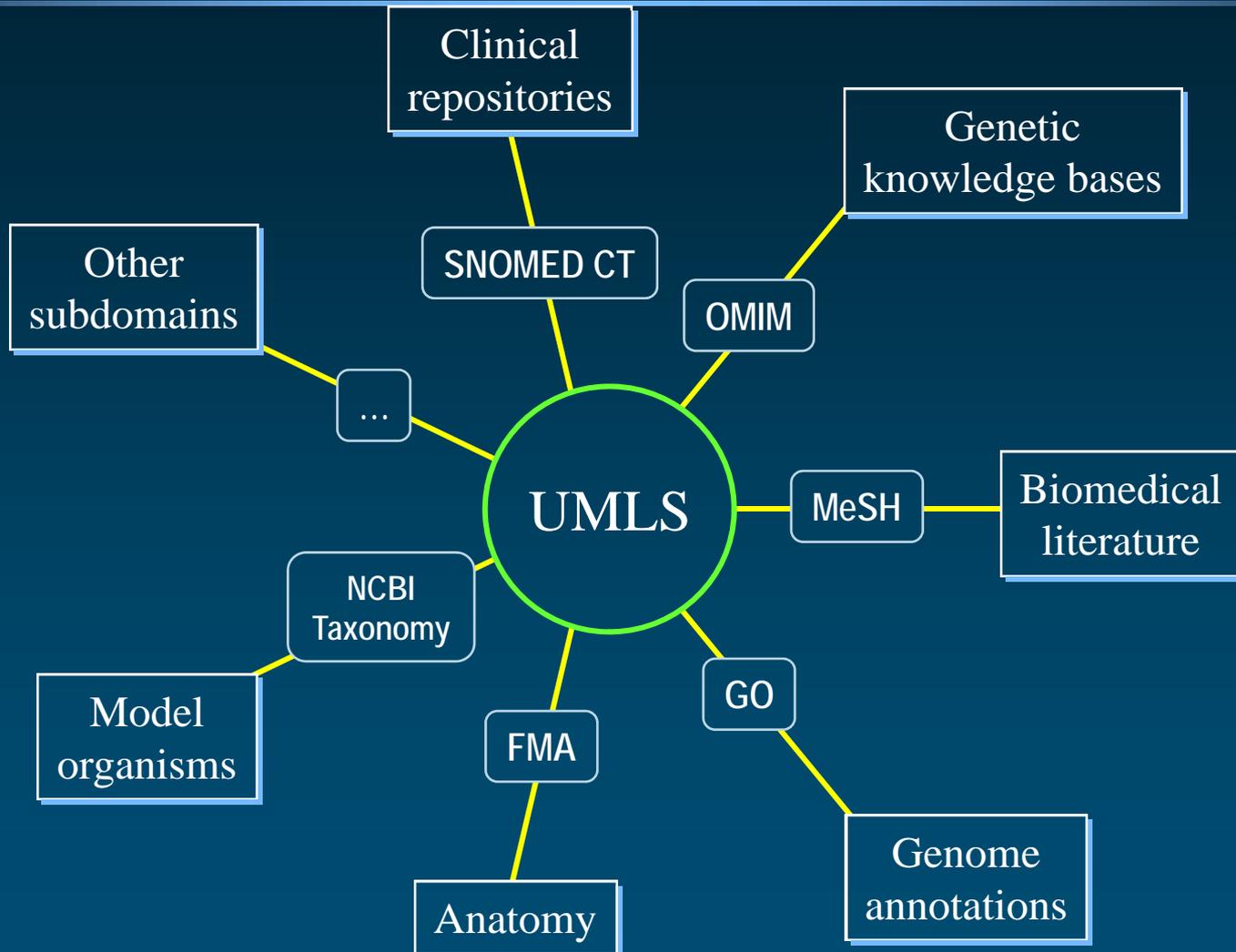
http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr

- ◆ Open Biomedical Ontologies (OBO)

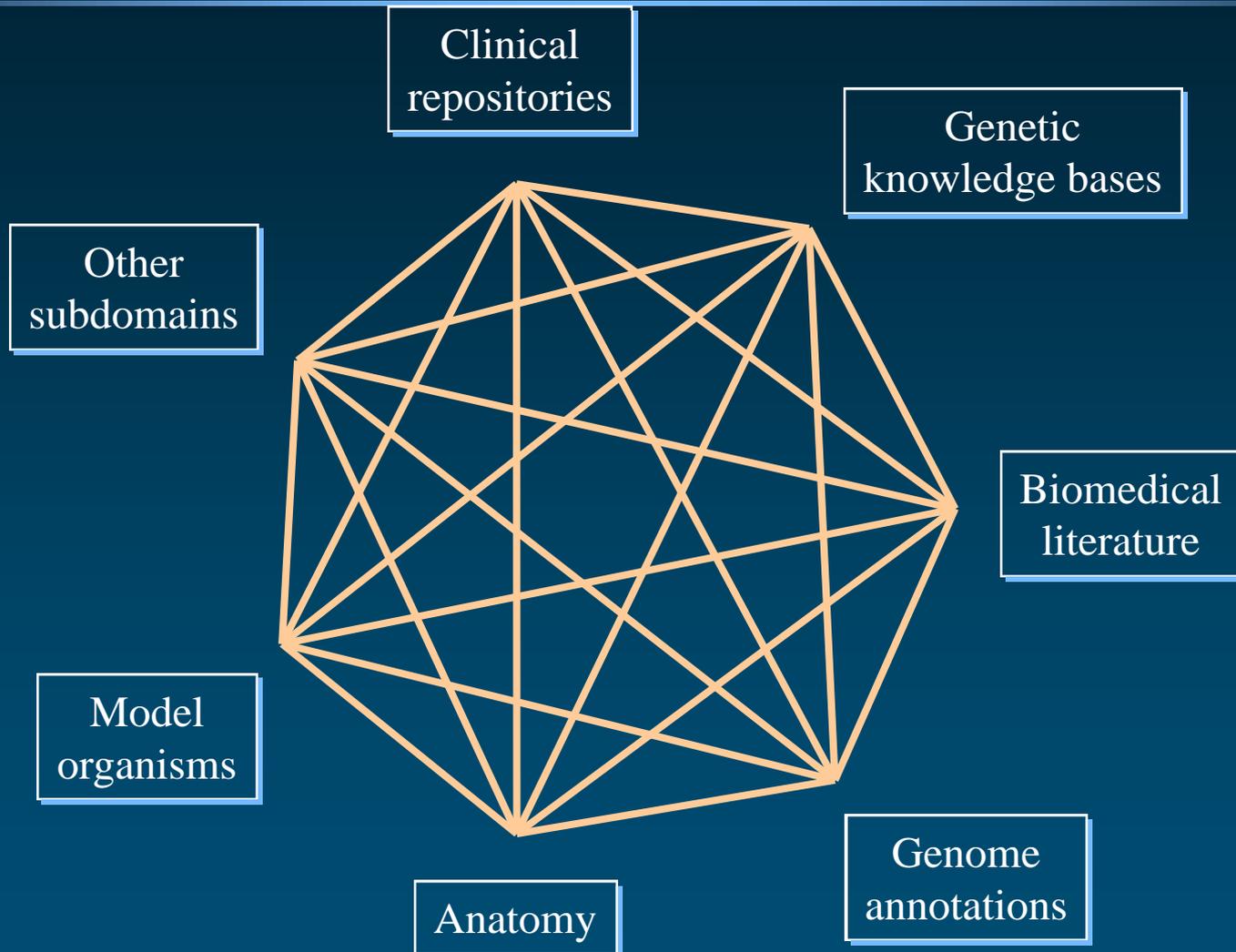
<http://obofoundry.org/>



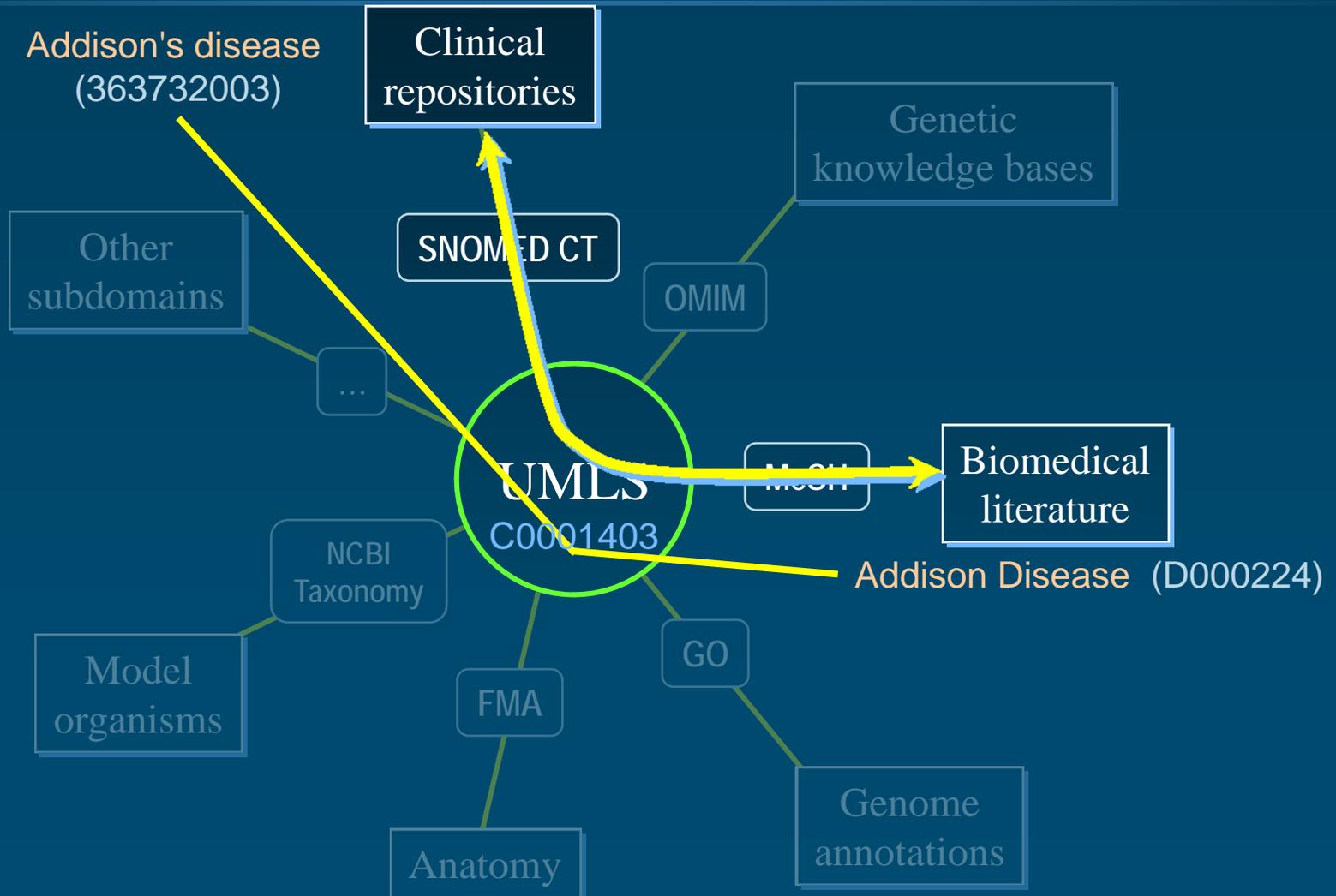
Integrating subdomains



Integrating subdomains



Trans-namespace integration



Mappings

- ◆ Created manually (e.g., UMLS)
 - Purpose
 - Directionality
- ◆ Created automatically (e.g., BioPortal)
 - Lexically: ambiguity, normalization
 - Semantically: lack of / incomplete formal definitions
- ◆ Key to enabling semantic interoperability
- ◆ Enabling resource for the Semantic Web

Challenging issues

Other issues

Availability

- ◆ Publicly available datasets
 - Is a *de facto* prerequisite for Semantic Web applications
 - Is a requirement from some funding agencies (“sharing plan”)
- ◆ Many datasets are freely available
 - Model organisms databases annotated to the Gene Ontology
 - NCBI knowledge bases in Entrez
 - Open Access journals
 - Public archives (e.g., PubMedCentral)
 - Ontologies from the Open Biomedical Ontology (OBO) family
- ◆ Limited availability
 - Intellectual property issues
 - Some ontologies and terminologies
 - Confidentiality and privacy issues
 - Personally identifiable medical information



Discoverability

- ◆ No universal repositories for biomedical datasets
 - Some datasets made available through portals (NCBI, EBI, NCBO)
- ◆ Ontology repositories
 - UMLS: 152 source vocabularies (biased towards healthcare applications)
 - NCBO BioPortal: ~141 ontologies (biased towards biological applications)
 - Limited overlap between the two repositories
- ◆ Need for discovery services
 - Metadata for ontologies and biomedical datasets



Formalism

- ◆ Several major formalism for ontologies
 - Web Ontology Language (OWL) – NCI Thesaurus
 - OBO format – most OBO ontologies
 - UMLS Rich Release Format (RRF) – UMLS, RxNorm
- ◆ Biomedical datasets
 - RDF/OWL
 - Legacy: free text, Excel spreadsheets, PowerPoint
- ◆ Conversion mechanisms
 - OBO to OWL
 - LexGrid (import/export to LexGrid internal format)

Ontology integration

- ◆ *Post hoc* integration , form the bottom up
 - UMLS approach
 - Integrates ontologies “as is”, including legacy ontologies
 - Facilitates the integration of the corresponding datasets
- ◆ Coordinated development of ontologies
 - OBO Foundry approach
 - Ensures consistency *ab initio*
 - Excludes legacy ontologies

Scalability

- ◆ Billions of triples for raw data in biomedicine
- ◆ Need for query systems to access distributed repositories (“data cloud”)
- ◆ Technical solutions have emerged
 - Traditional vendors have embraced RDF/OWL
 - New vendors/products



Semantic Web and medical libraries

Why additional library services?

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

What additional services?

- ◆ Refined information retrieval
 - Indexing on relations in addition to concepts
 - *Find articles asserting that **IL-13 inhibits COX-2***
- ◆ Multi-document summarization
 - Extract and visualize facts from the literature
 - *Summarize the top 300 papers on **panic disorder***
- ◆ Question answering
 - Clinical and biological questions
 - *What drugs **interact** with **imipramine**?*
- ◆ Knowledge discovery
 - Reasoning with facts from heterogeneous resources
 - *From MEDLINE and UMLS together*



Normalized and integrated knowledge

- ◆ Normalized knowledge
 - Common format
 - Common identification mechanism
- ◆ Integrated knowledge
 - Single repository
 - Seamless environment
 - *Phenotype and genotype information together*

Biomedical Knowledge Repository



Sources of knowledge

- ◆ Biomedical literature
 - Predications extracted from **MEDLINE** abstracts and full-text publicly available articles using text mining techniques
 - Other corpora (e.g., **ClinicalTrials.gov**)
- ◆ Terminological knowledge
 - **UMLS**
- ◆ Structured knowledge bases
 - NCBI resources (e.g., **Entrez Gene**)
 - Functional annotations from model organism databases
 - ...
- ◆ Contributed knowledge
 - The repository is open to collaborators outside NLM



Formalism Triples



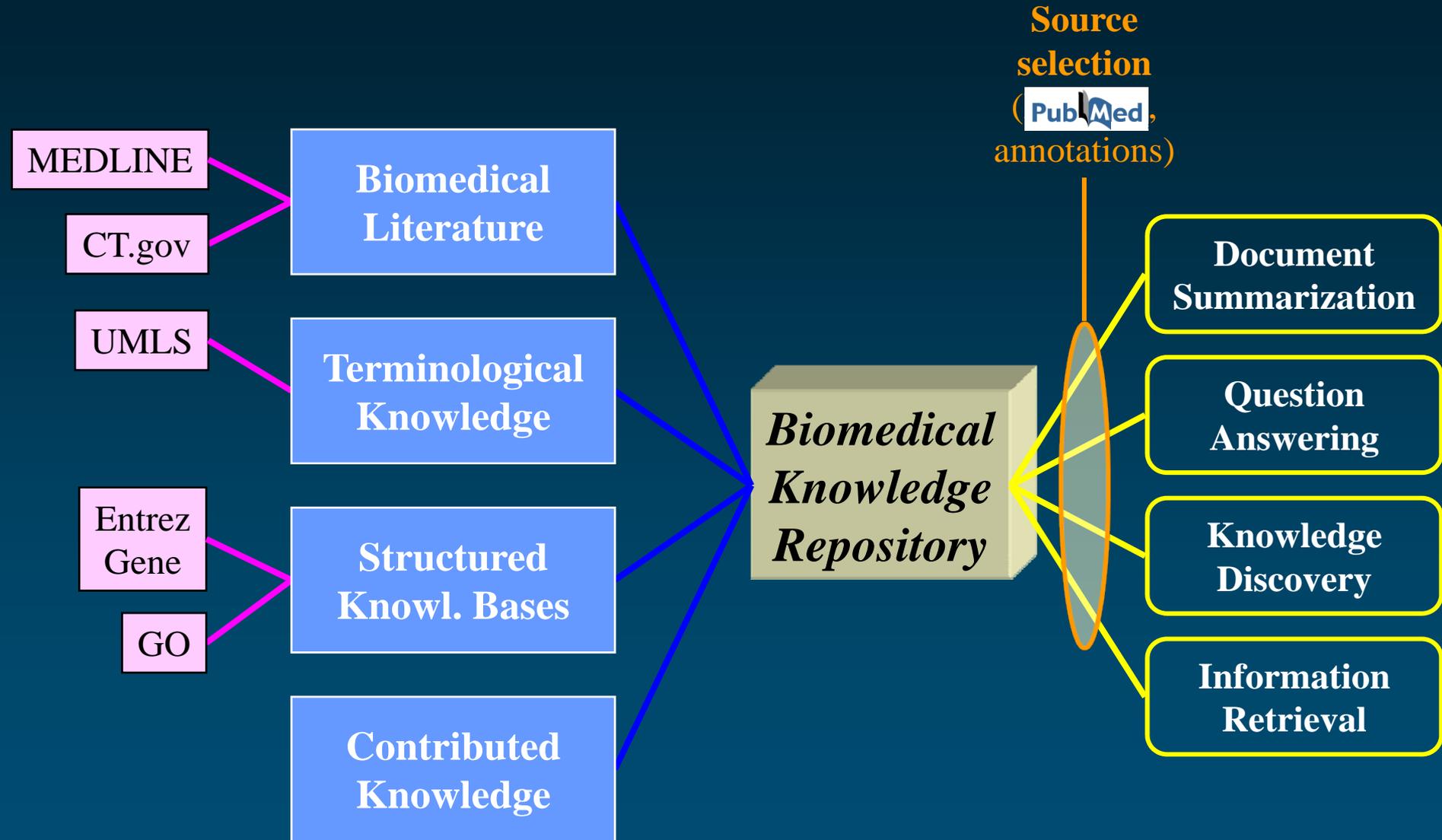
- ◆ Facts
- ◆ Assertions
- ◆ Relations
- ◆ Semantic predications
- ◆ RDF triples

Annotated knowledge Metadata

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



Advanced Library Services A research project at NLM



Wrapping up

Take home points

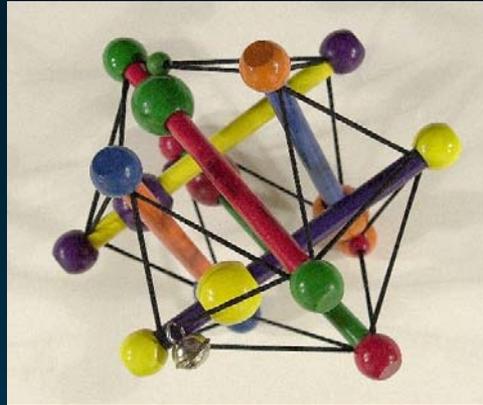
◆ Biomedical Semantic Web

- Data integration
- Enable applications
 - Hypothesis generation
 - Knowledge discovery

◆ Paradigm shift

- Migration towards data-driven research
(as opposed to hypothesis-driven research)
- Translational research





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