



*Kno.e.sis*

Wright State University, Dayton, Ohio  
May 27, 2009

## Ontologies and data integration in biomedicine



*Olivier Bodenreider*

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

# Outline

---

- ◆ Why integrate data?
- ◆ Ontologies and data integration
- ◆ Examples
- ◆ Challenging issues



*Why integrate data?*

# Why integrate data?

- ◆ Sources of information
  - Created by
    - Independent researchers
    - Separate workflows
  - Heterogeneous
  - Scattered
  - “Silos”
- ◆ To identify patterns in integrated datasets
  - Hypothesis generation
  - Knowledge discovery



# Motivation Translational research

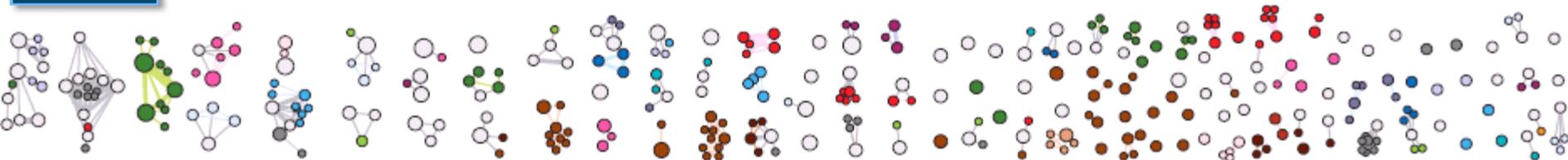
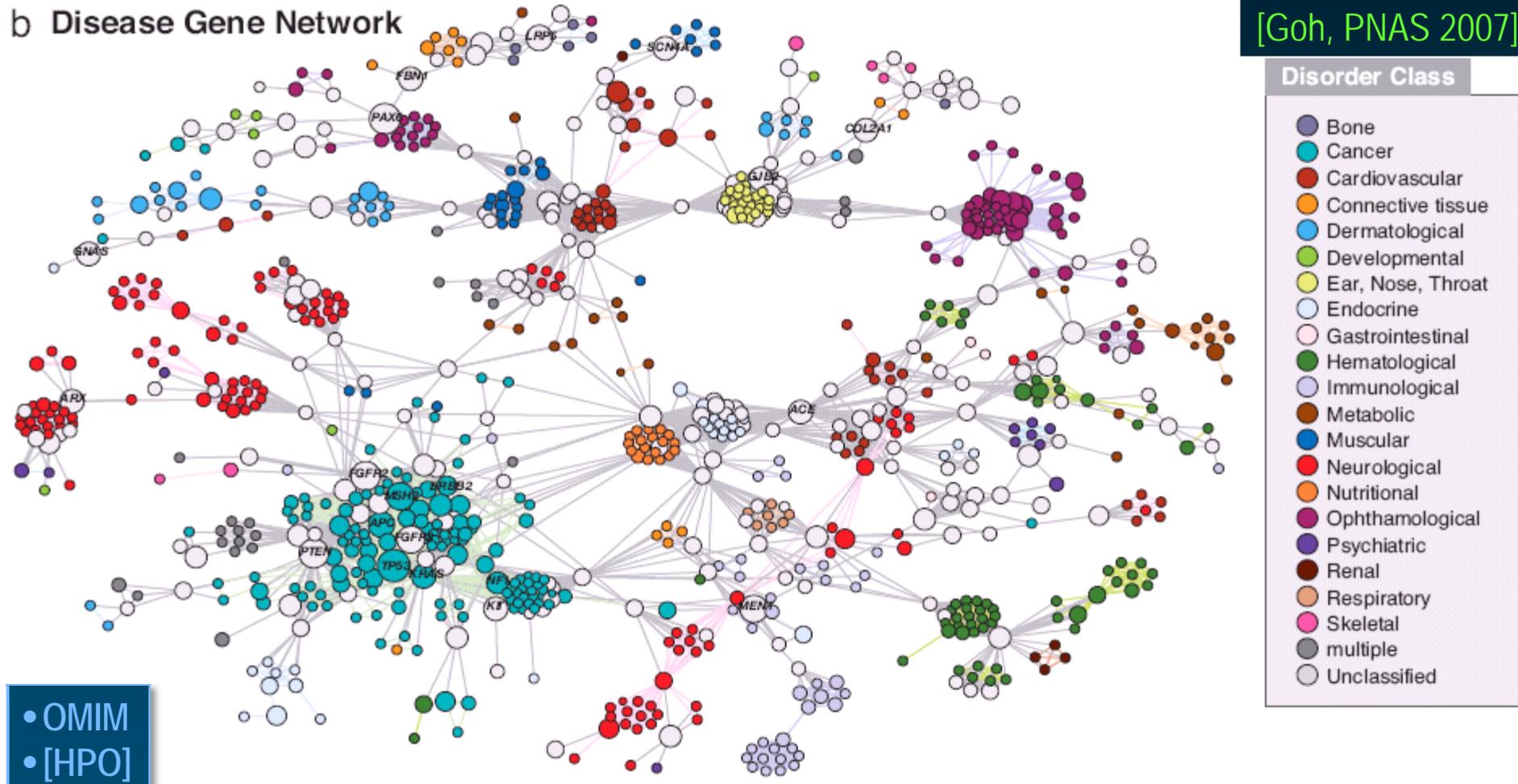
- ◆ “Bench to Bedside”
- ◆ Integration of clinical and research activities and results
- ◆ Supported by research programs
  - NIH Roadmap
  - Clinical and Translational Science Awards (CTSA)
- ◆ Requires the effective integration and exchange and of information between
  - Basic research
  - Clinical research



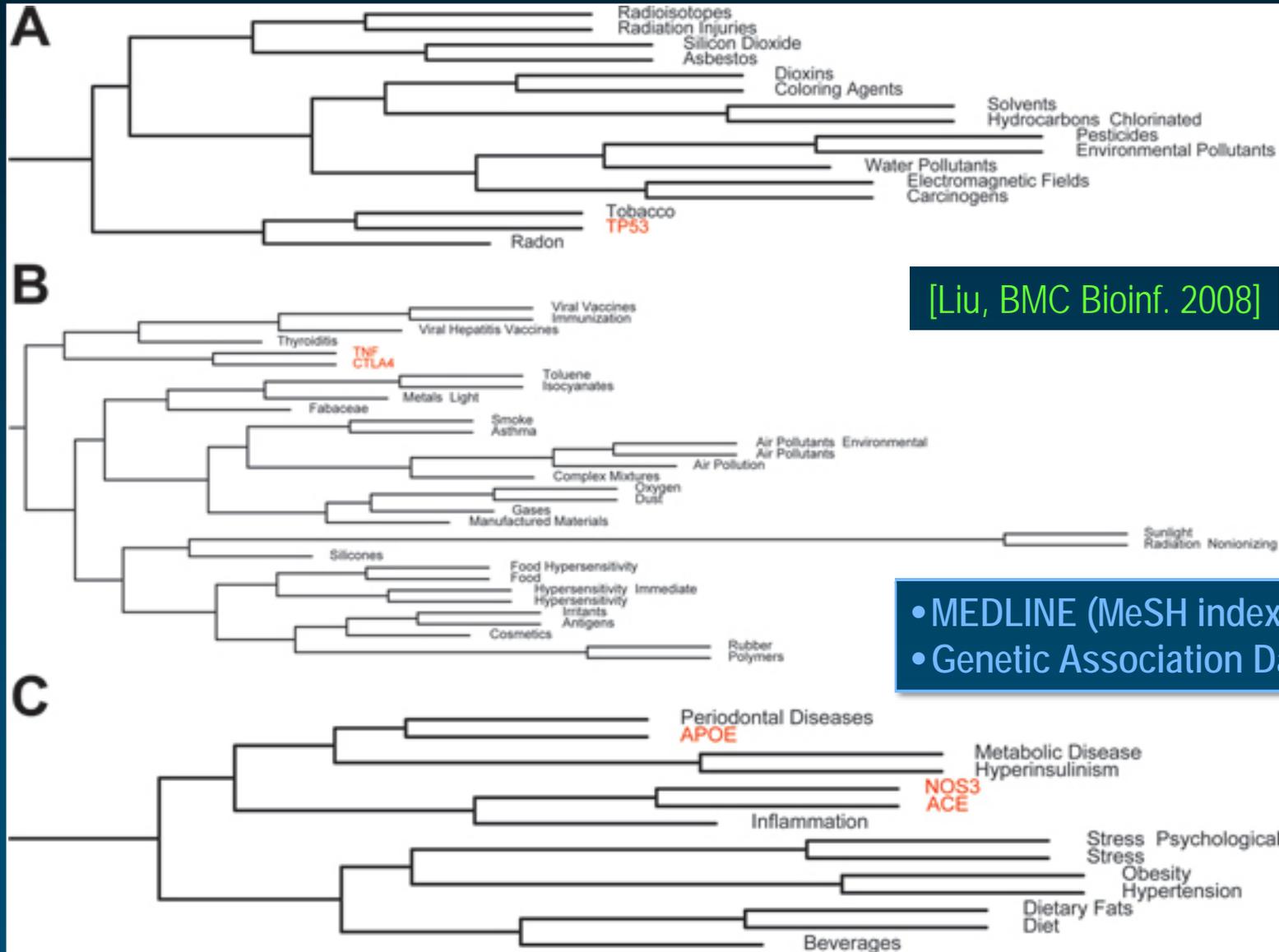
# Genotype and phenotype

b Disease Gene Network

[Goh, PNAS 2007]



# Genes and environmental factors



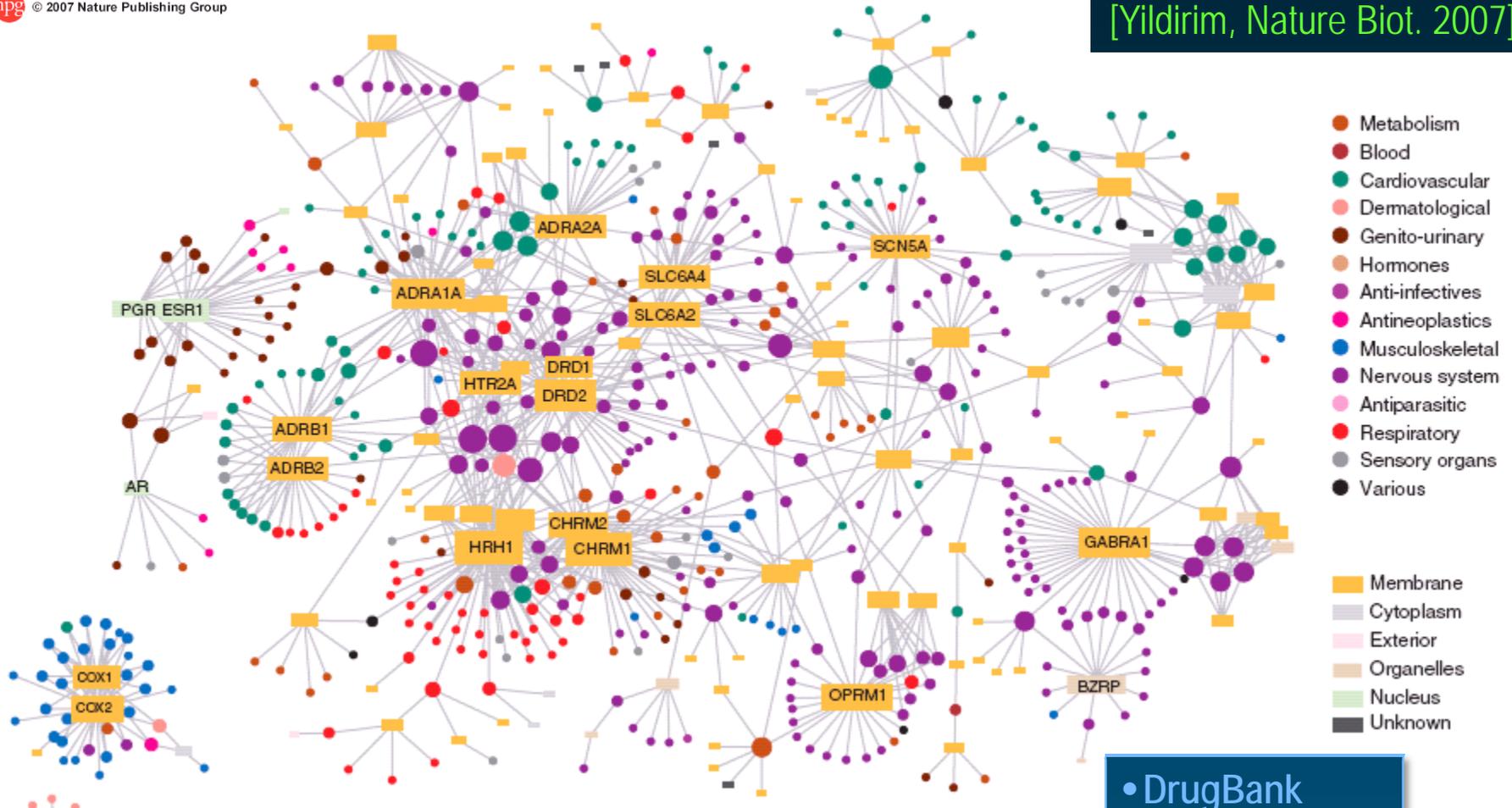
[Liu, BMC Bioinf. 2008]

- MEDLINE (MeSH index terms)
- Genetic Association Database

# Integrating drugs and targets

mpg © 2007 Nature Publishing Group

[Yildirim, Nature Biot. 2007]



- DrugBank
- ATC
- Gene Ontology



Why ontologies?

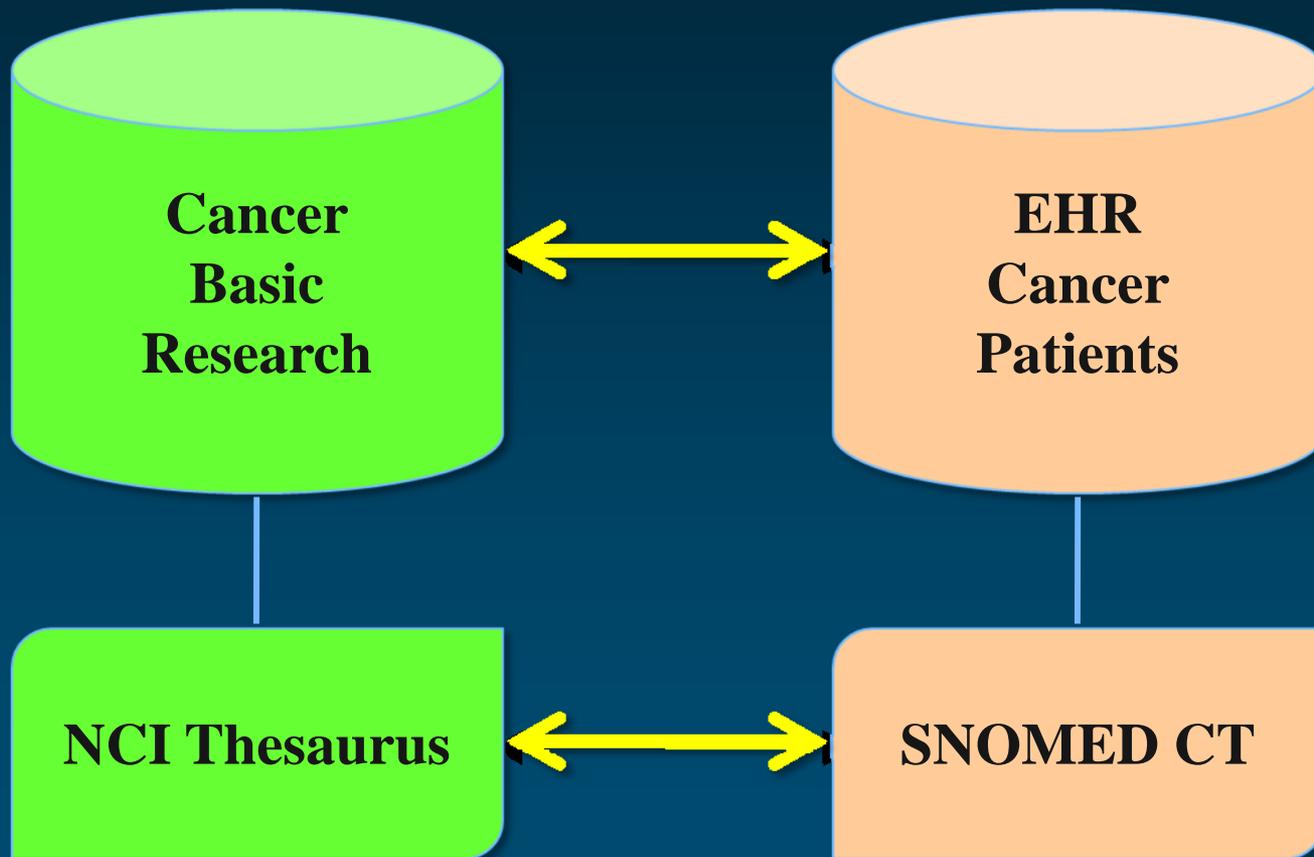
# Uses of biomedical ontologies

- ◆ Knowledge management
  - Annotating data and resources
  - Accessing biomedical information
  - Mapping across biomedical ontologies
- ◆ Data integration, exchange and semantic interoperability
- ◆ Decision support
  - Data selection and aggregation
  - Decision support
  - NLP applications
  - Knowledge discovery

[Bodenreider, YBMI 2008]



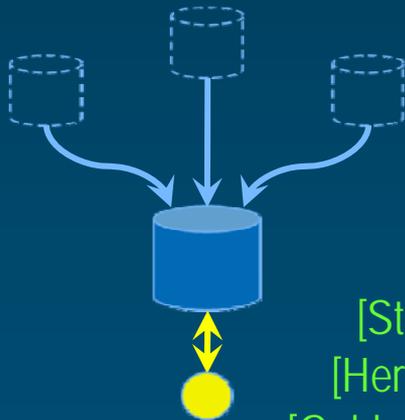
# Terminology and translational research



# Approaches to data integration (1)

## ◆ Warehousing

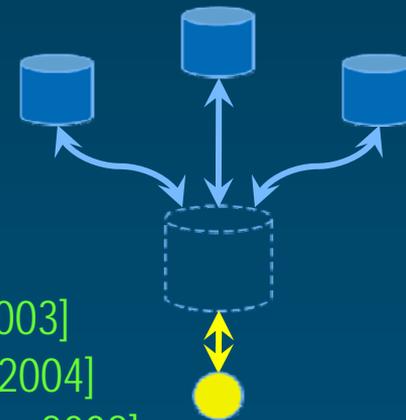
- Sources to be integrated are transformed into a common format and converted to a common vocabulary



[Stein, Nature Rev. Gen. 2003]  
[Hernandez, SIGMOD Rec. 2004]  
[Goble J. Biomedical Informatics 2008]

## ◆ Mediation

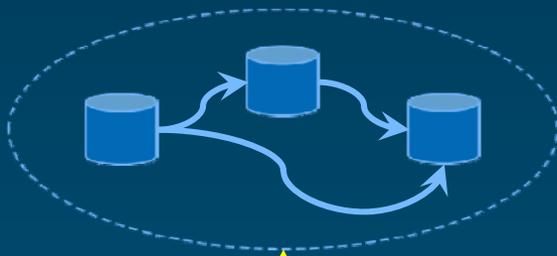
- Local schema (of the sources)
- Global schema (in reference to which the queries are made)



# Approaches to data integration (2)

## ◆ Linked data

- Links among data elements
- Enable navigation by humans



[Stein, Nature Rev. Gen. 2003]

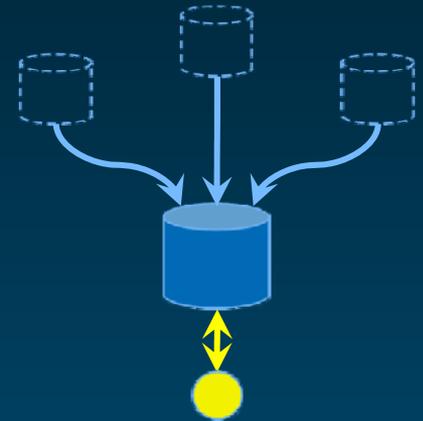
[Hernandez, SIGMOD Rec. 2004]

[Goble J. Biomedical Informatics 2008]

# Ontologies and warehousing

## ◆ Role

- Provide a conceptualization of the domain
  - Help define the schema
  - Information model vs. ontology
- Provide value sets for data elements
- Enable standardization and sharing of data



## ◆ Examples

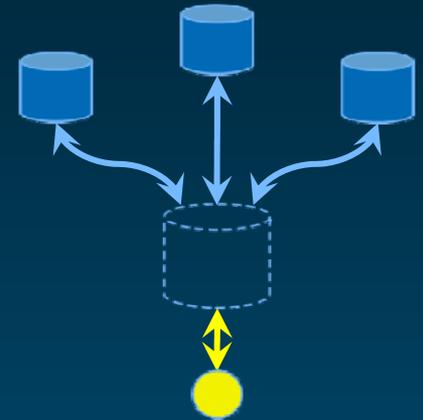
- Annotations to the Gene Ontology
- BioWarehouse
- Clinical information systems

<http://biowarehouse.ai.sri.com/>

# Ontologies and mediation

## ◆ Role

- Reference for defining the global schema
- Map between local and global schemas
  - Query reformulation
  - Local-as-view vs. Global-as-view



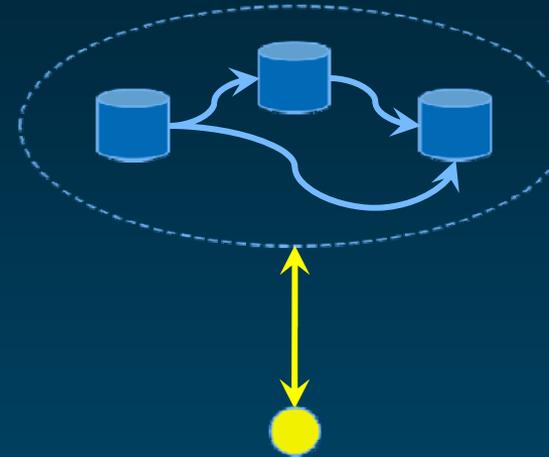
## ◆ Examples

- TAMBIS [Stevens, Bioinformatics 2000]
- BioMediator [Louie, AMIA 2005]
- OntoFusion [Perez-Rey, Comput Biol Med 2006]

# Ontologies and linked data

## ◆ Role

- Explicit conceptualization of the domain
- Semantic normalization of data elements



## ◆ Examples

- Entrez
- Semantic Web mashups
- Bio2RDF

[<http://www.ncbi.nlm.nih.gov/>]

[J. Biomedical informatics 41(5) 2008]

[<http://bio2rdf.org/>]



# Ontologies and data integration

- ◆ Source of identifiers for biomedical entities
  - Semantic normalization
  - *Warehouse approaches*
- ◆ Source of reference relations for the global schema
  - Mapping between local and global schemas
  - *Mediator-based approaches*
- ◆ Source of identifiers for biomedical entities
  - Semantic normalization
  - Explicit conceptualization of the domain
  - *Linked data approaches*



# Ontologies and data aggregation

- ◆ Source of hierarchical relations
  - Aggregate data into coarser categories
  - Abstract away from low-frequency, fine grained data points
  - Increase power
  - Improve visualization



# Examples

*Gene Ontology*

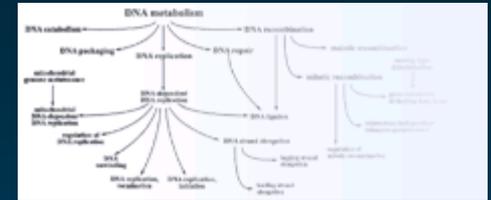
<http://www.geneontology.org/>



# Annotating data

## ◆ Gene Ontology

- Functional annotation of gene products in several dozen model organisms



## ◆ Various communities use the same controlled vocabularies

## ◆ Enabling comparisons across model organisms

## ◆ Annotations

- Assigned manually by curators
- Inferred automatically (e.g., from sequence similarity)

# GO Annotations for Aldh2 (mouse)

GO Annotations in Tabular Form

(Text View)

(GO Graph)



Category	Classification Term	Evidence
Molecular Function	<a href="#">aldehyde dehydrogenase (NAD) activity</a>	IEA
Molecular Function	<a href="#">oxidoreductase activity</a>	IEA
Molecular Function	<a href="#">oxidoreductase activity</a>	IEA
Cellular Component	<a href="#">mitochondrion</a>	IDA
Biological Process	<a href="#">metabolic process</a>	IEA
Biological Process	<a href="#">oxidation reduction</a>	IEA

[http:// www.informatics.jax.org/](http://www.informatics.jax.org/)

# GO ALD4 in Yeast

## GO Annotations

## Molecular Function

Manually curated

## Biological Process

Manually curated

## Cellular Component

Manually curated

High-throughput

All **ALD4** GO evidence and references

*View Computational GO annotations for **ALD4***

- aldehyde dehydrogenase (NAD) activity (IDA, IMP, ISS)
- aldehyde dehydrogenase [NAD(P)+] activity (IDA)
  
- ethanol metabolic process (IMP)
  
- mitochondrial nucleoid (IDA)
- mitochondrion (IMP, ISS)
- mitochondrion (IDA)



<http://db.yeastgenome.org/>



# GO Annotations for ALDH2 (Human)



Function						
GO:0016491	oxidoreductase activity	interpro	IEA	IPR015590	UniProt	9606
GO:0016491	oxidoreductase activity	interpro	IEA	IPR016160	UniProt	9606
GO:0016491	oxidoreductase activity	interpro	IEA	IPR016162	UniProt	9606
GO:0016491	oxidoreductase activity	interpro	IEA	IPR016161	UniProt	9606
GO:0016491	oxidoreductase activity	spkw	IEA	KW-0560	UniProt	9606
GO:0004029	aldehyde dehydrogenase (NAD) activity	1306115	TAS		PINC	9606
GO:0004030	aldehyde dehydrogenase [NAD(P)+] activity	8903321	TAS		PINC	9606
GO:0009055	electron carrier activity	8903321	TAS		UniProt	9606
GO:0004029	aldehyde dehydrogenase (NAD) activity	enzyme	IEA	1.2.1.3	UniProt	9606

<http://www.ebi.ac.uk/GOA/>



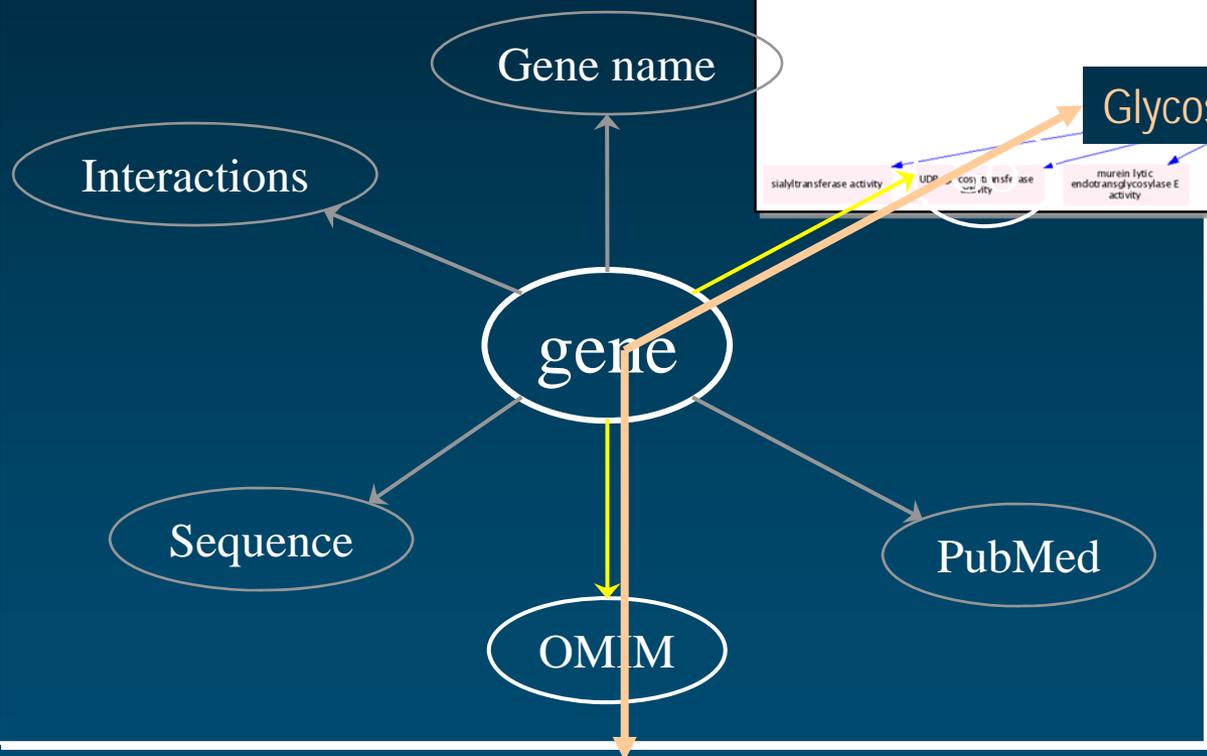
# Integration applications

- ◆ Based on shared annotations
  - Enrichment analysis (within/across species)
  - Clustering (co-clustering with gene expression data)
- ◆ Based on the structure of GO
  - Closely related annotations
  - Semantic similarity [Lord, PSB 2003]
- ◆ Based on associations between gene products and annotations [Bodenreider, PSB 2005]
- ◆ Leveraging reasoning [Sahoo, Medinfo 2007]

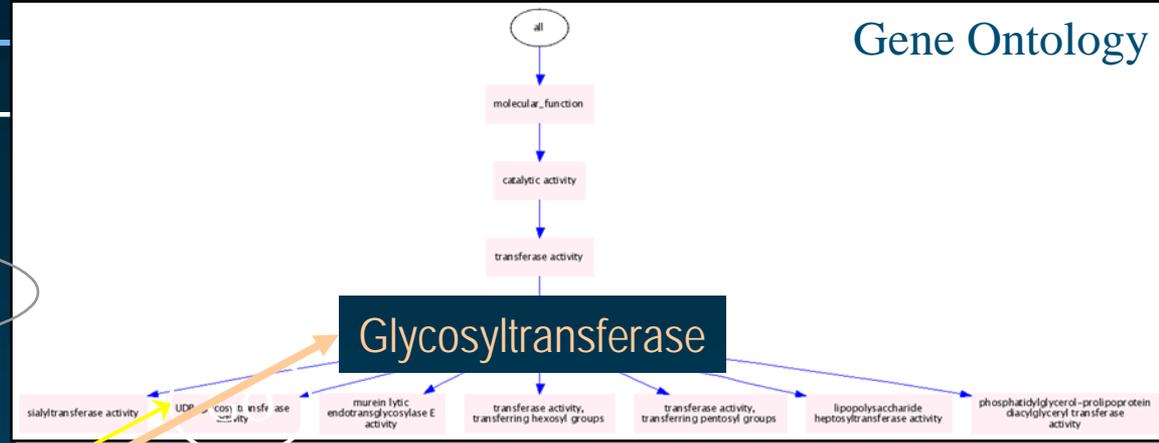


# Integration Entrez Gene + GO

Entrez Gene



Gene Ontology

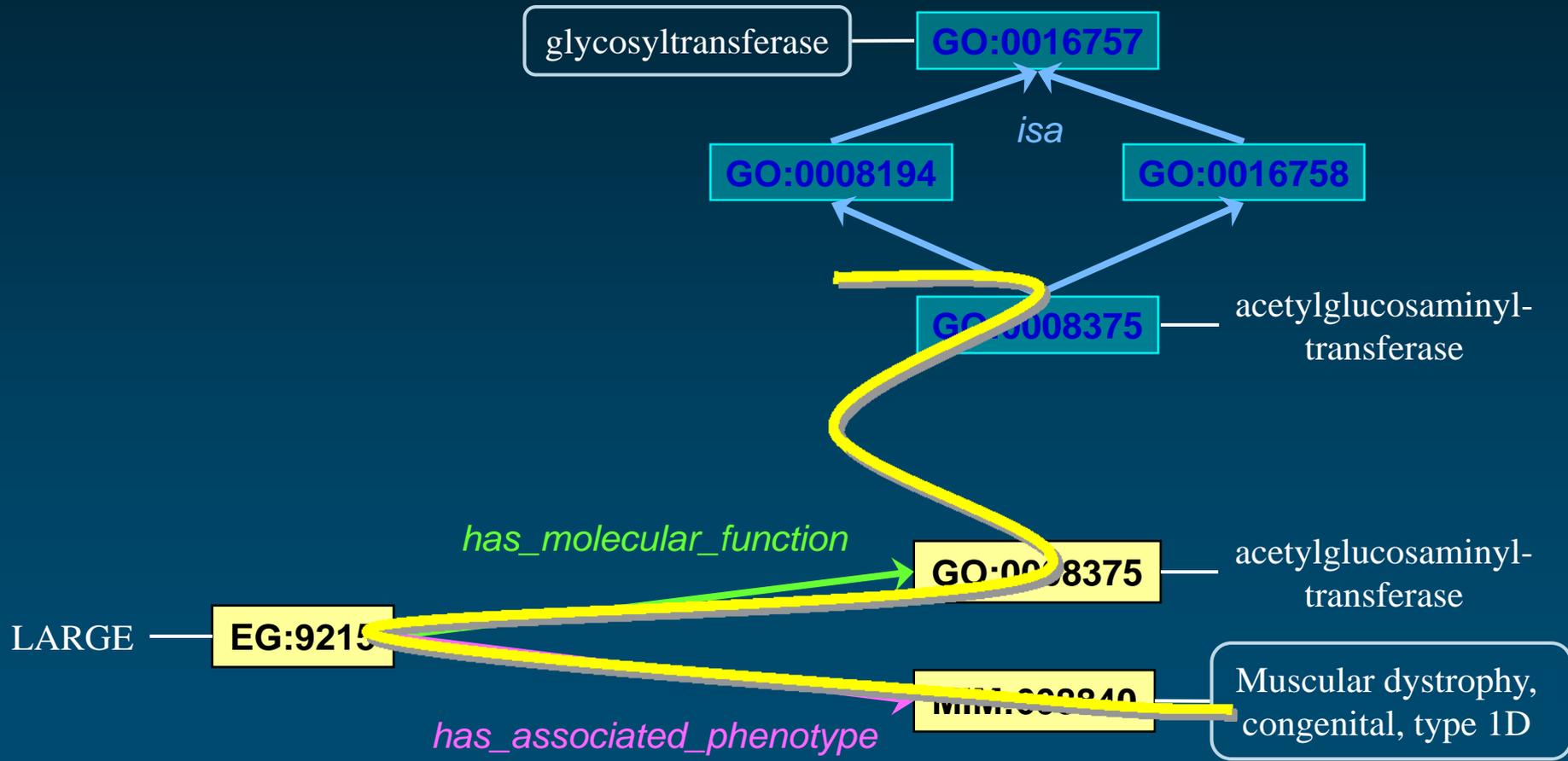


[Sahoo, Medinfo 2007]

Congenital muscular dystrophy



# From *glycosyltransferase* to congenital muscular dystrophy



# Examples

*caBIG*

<http://cabig.nci.nih.gov/>



National Cancer Institute



**caBIG™**

Cancer Biomedical  
Informatics Grid™

# Cancer Biomedical Informatics Grid

- ◆ US National Cancer Institute
- ◆ Common infrastructure used to share data and applications across institutions to support cancer research efforts in a grid environment
- ◆ Service-oriented architecture
- ◆ Data and application services available on the grid
- ◆ Supported by ontological resources



# caBIG services

- ◆ caArray
  - Microarray data repository
- ◆ caTissue
  - Biospecimen repository
- ◆ caFE (Cancer Function Express)
  - Annotations on microarray data
- ◆ ...
  
- ◆ caTRIP
  - Cancer Translational Research Informatics Platform
  - Integrates data services



# Ontological resources

## ◆ NCI Thesaurus

- Reference terminology for the cancer domain
- ~ 60,000 concepts
- OWL Lite

## ◆ Cancer Data Standards Repository (caDSR)

- Metadata repository
- Used to bridge across UML models through Common Data Elements
- Links to concepts in ontologies



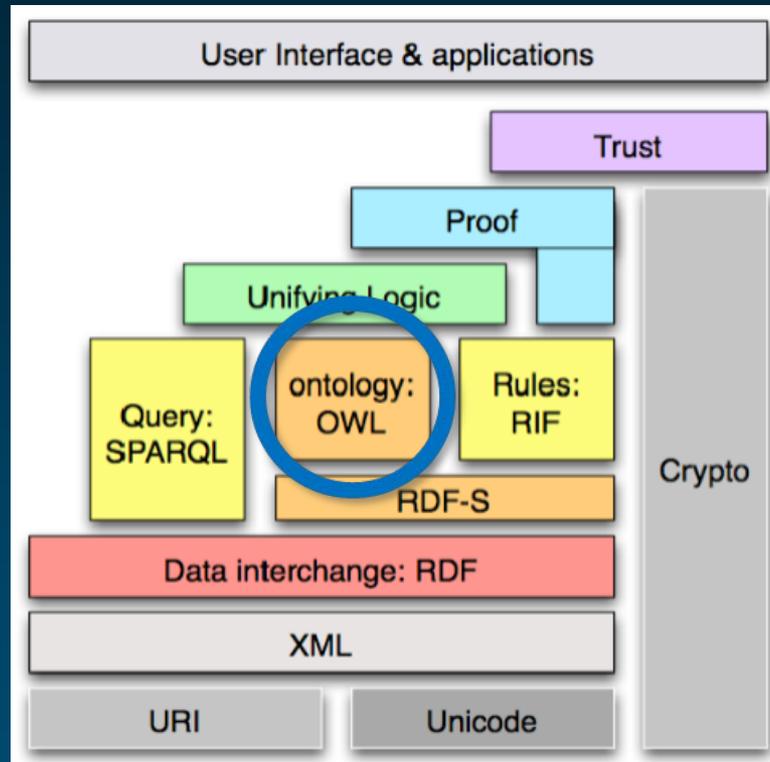
# Examples

*Semantic Web  
for Health Care and Life Sciences*

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



# Semantic Web layer cake







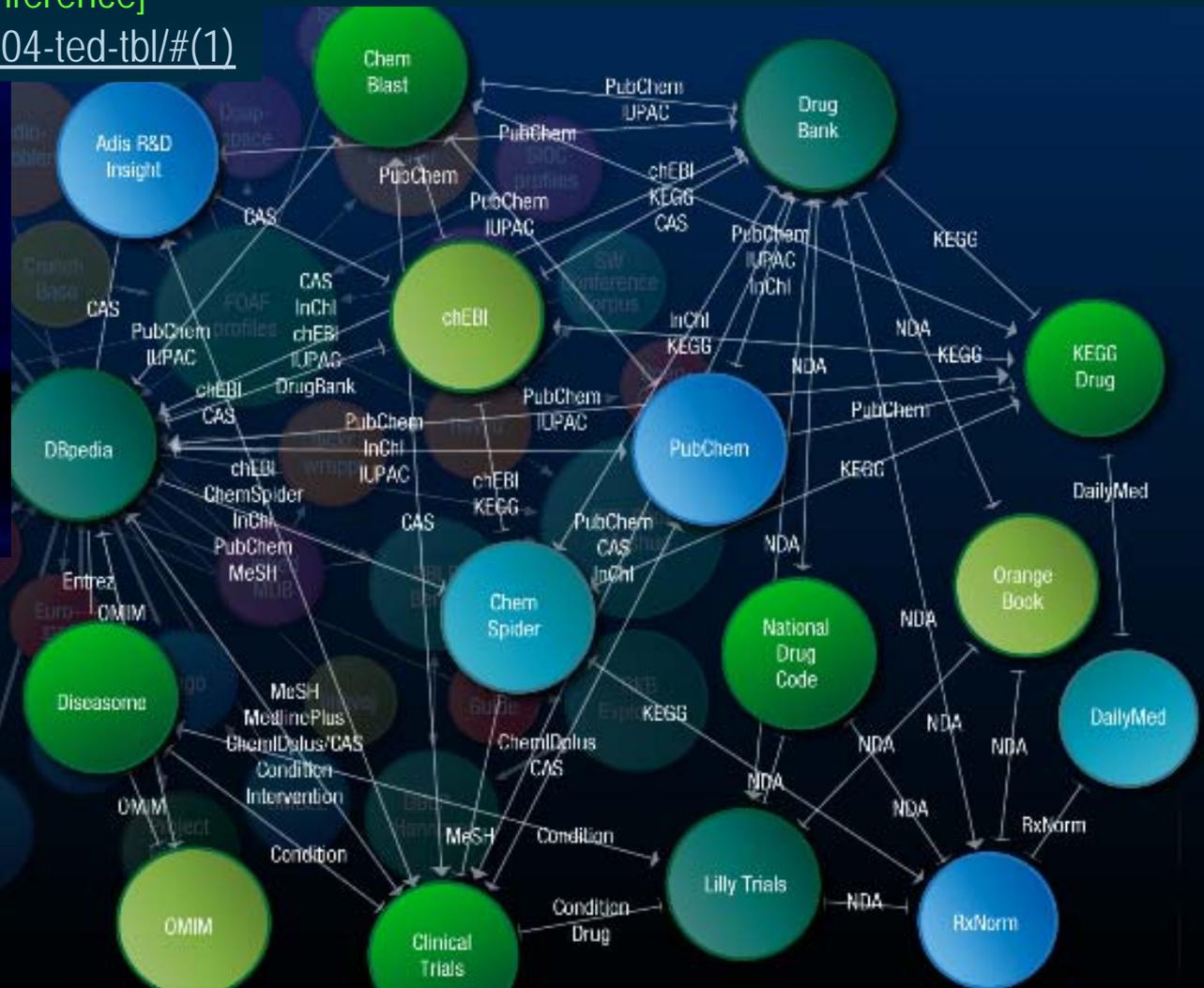
# Linked biomedical data

[Tim Berners-Lee TED 2009 conference]

[http://www.w3.org/2009/Talks/0204-ted-tbl/#\(1\)](http://www.w3.org/2009/Talks/0204-ted-tbl/#(1))



**TIMBERNERS-LEE**



# W3C Health Care and Life Sciences IG



## 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.

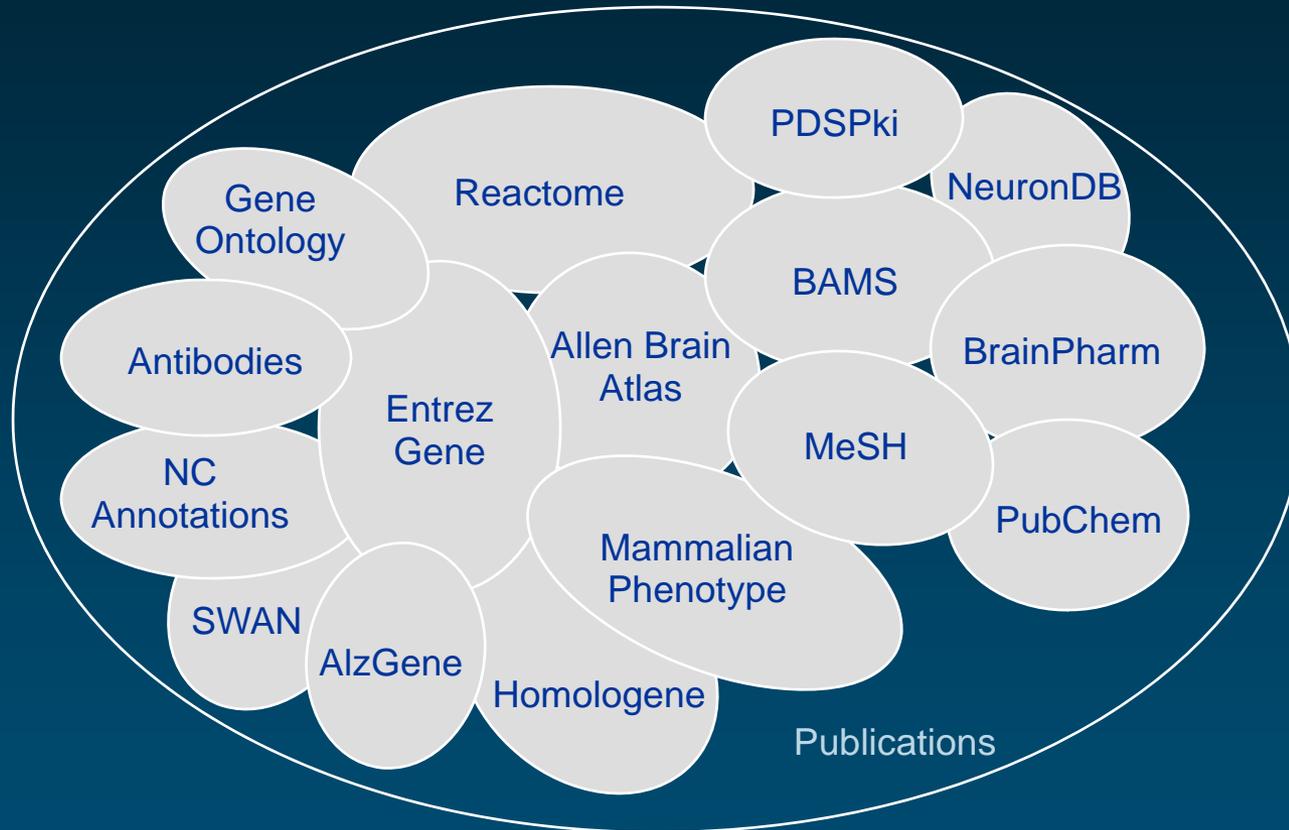
# Biomedical Semantic Web

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

[Ruttenberg, BMC Bioinf. 2007]



# HCLS mashup of biomedical sources

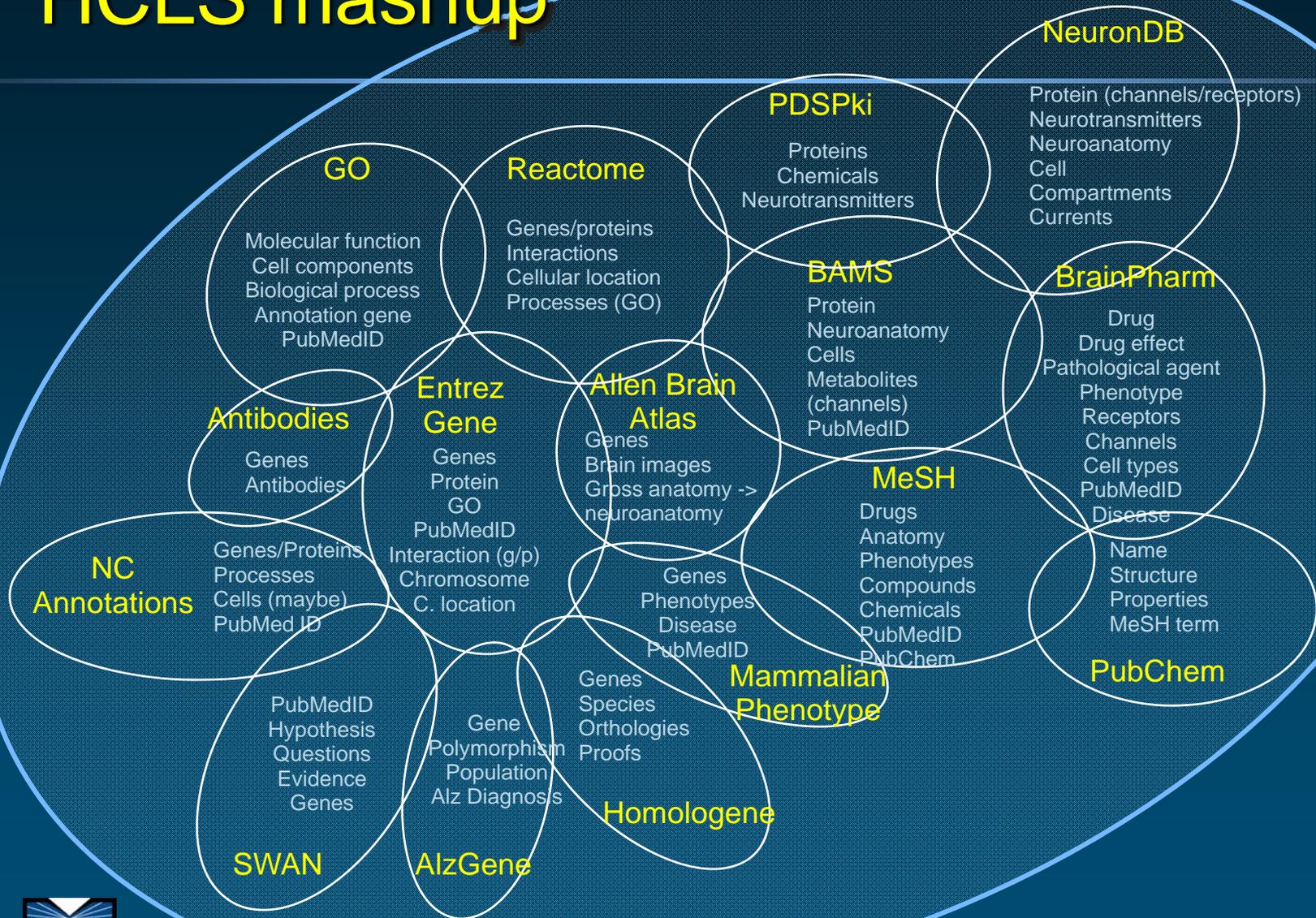


[http://esw.w3.org/topic/HCLS/HCLSIG\\_DemoHomePage\\_HCLSIG\\_Demo](http://esw.w3.org/topic/HCLS/HCLSIG_DemoHomePage_HCLSIG_Demo)

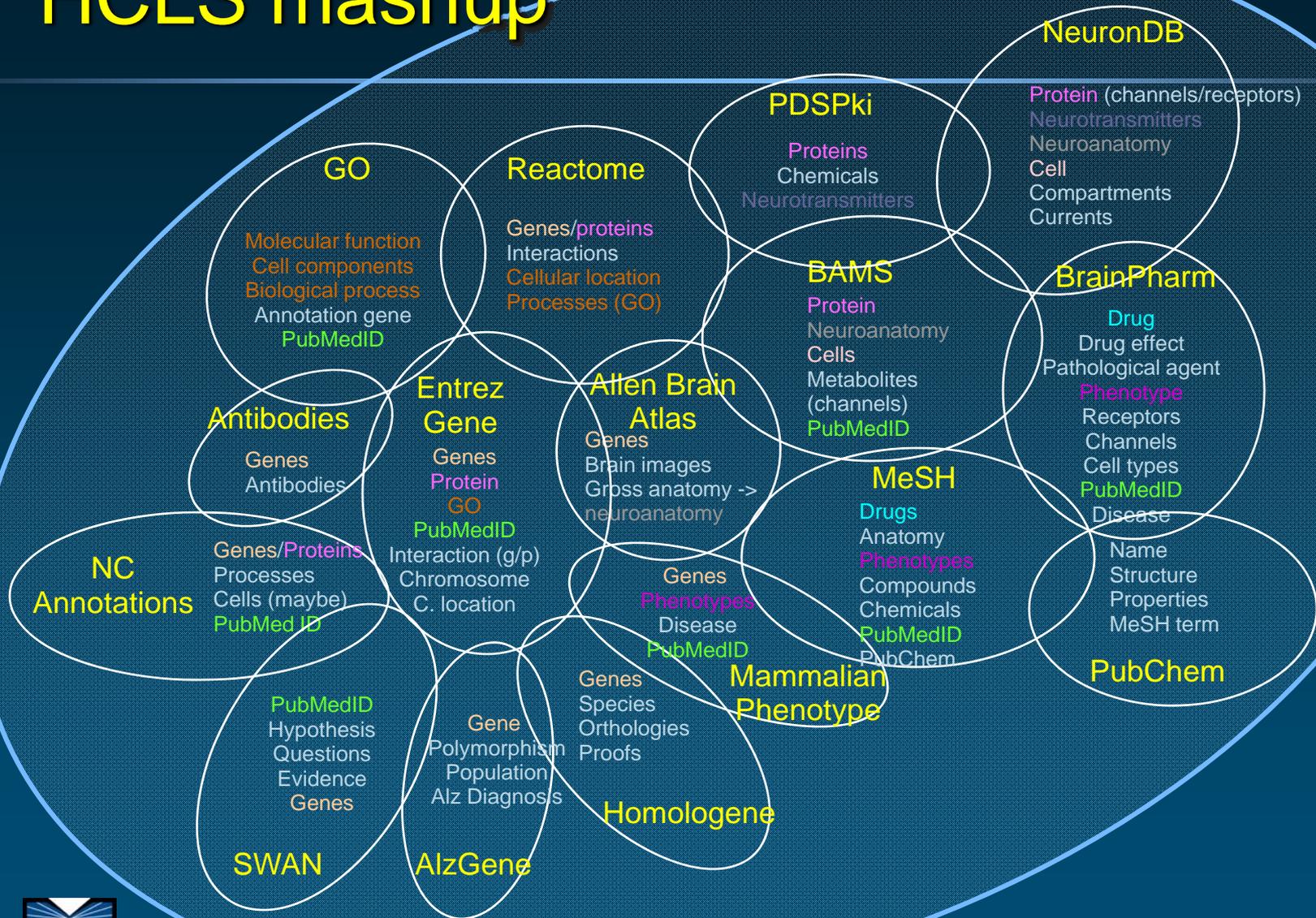




# HCLS mashup



# 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]



# Challenging issues

# Challenging issues

---

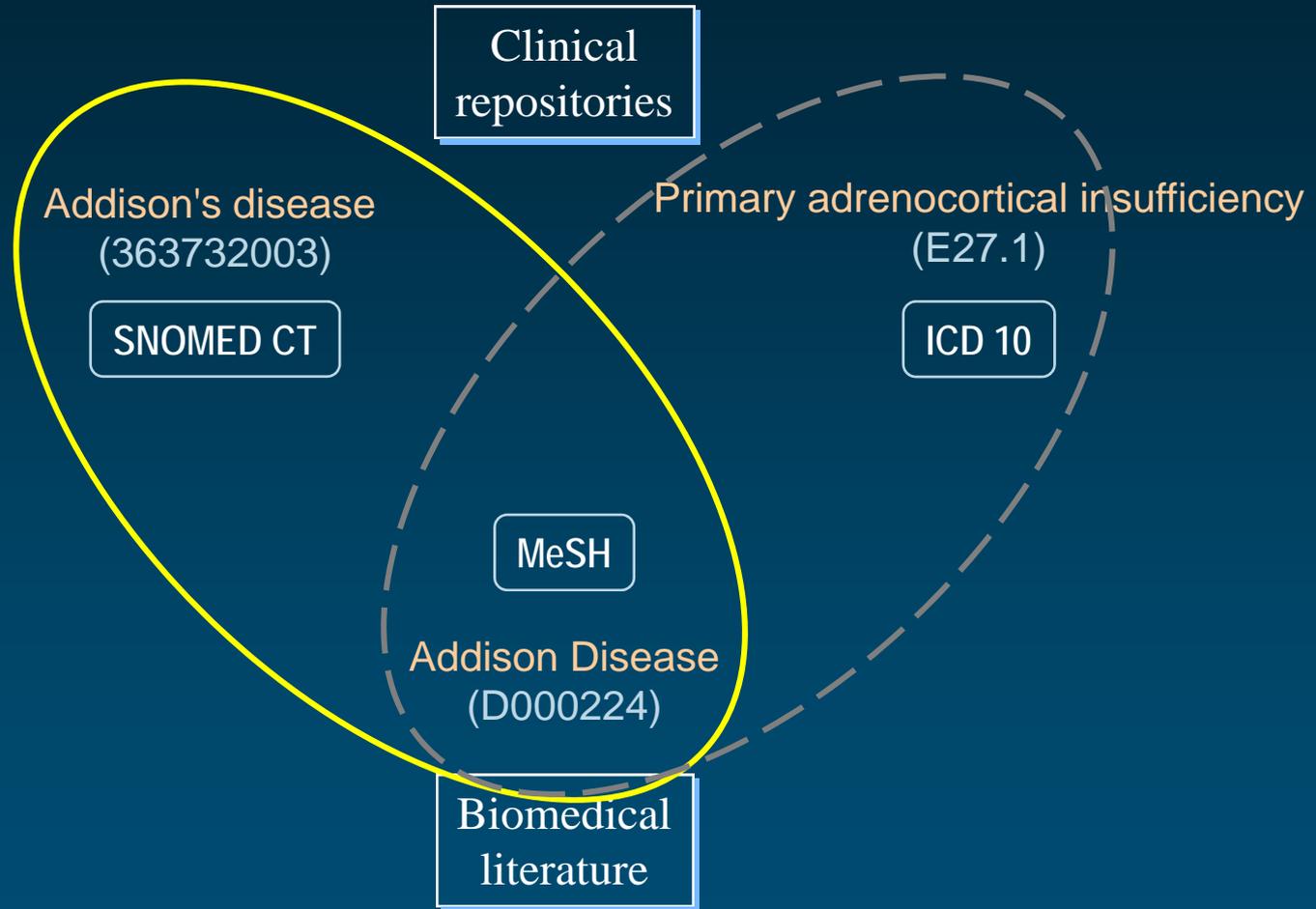
- ◆ Bridges across ontologies
- ◆ Permanent identifiers for biomedical entities
- ◆ Other issues



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

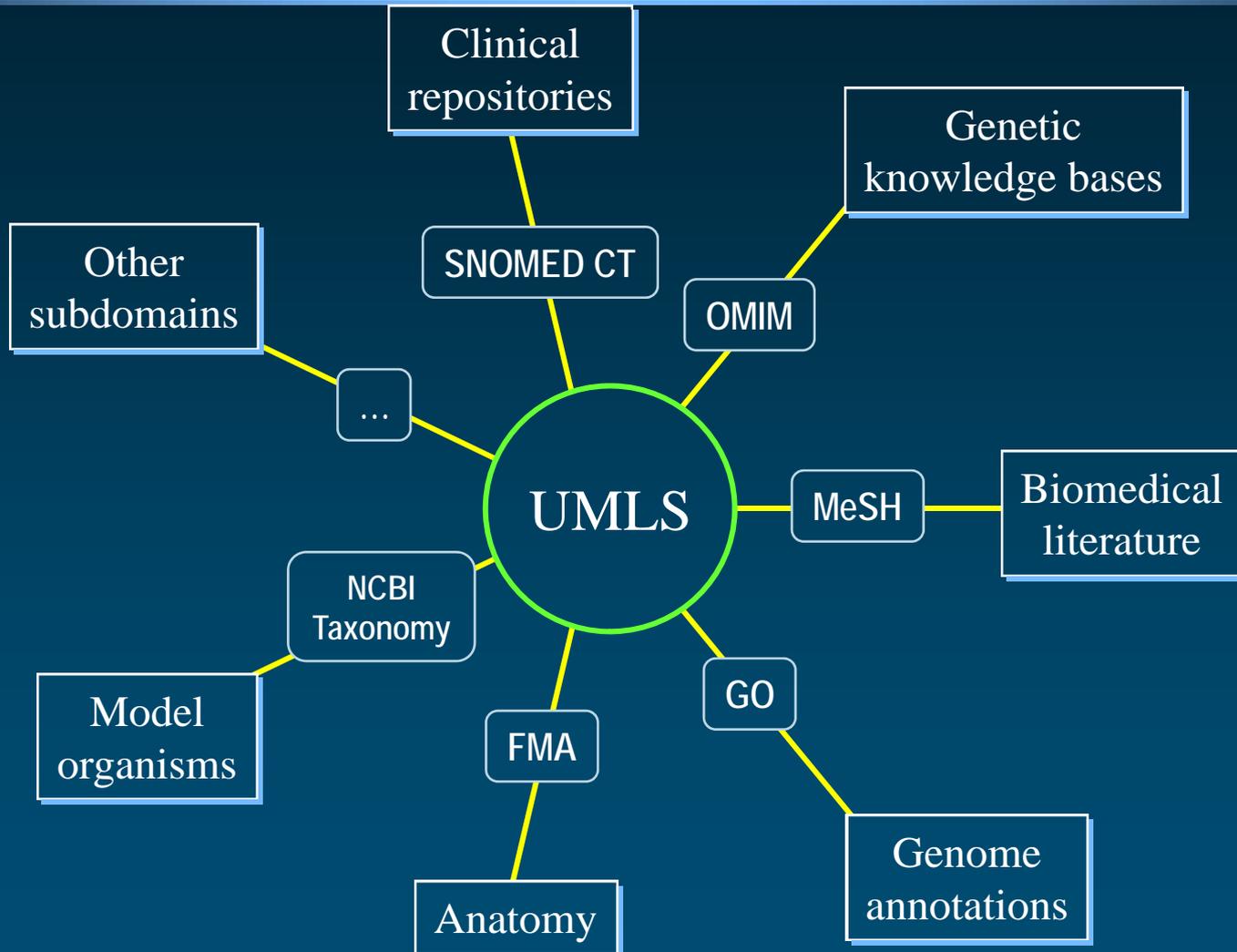
[http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore\\_overview/cadsr](http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr)

- ◆ Open Biomedical Ontologies (OBO)

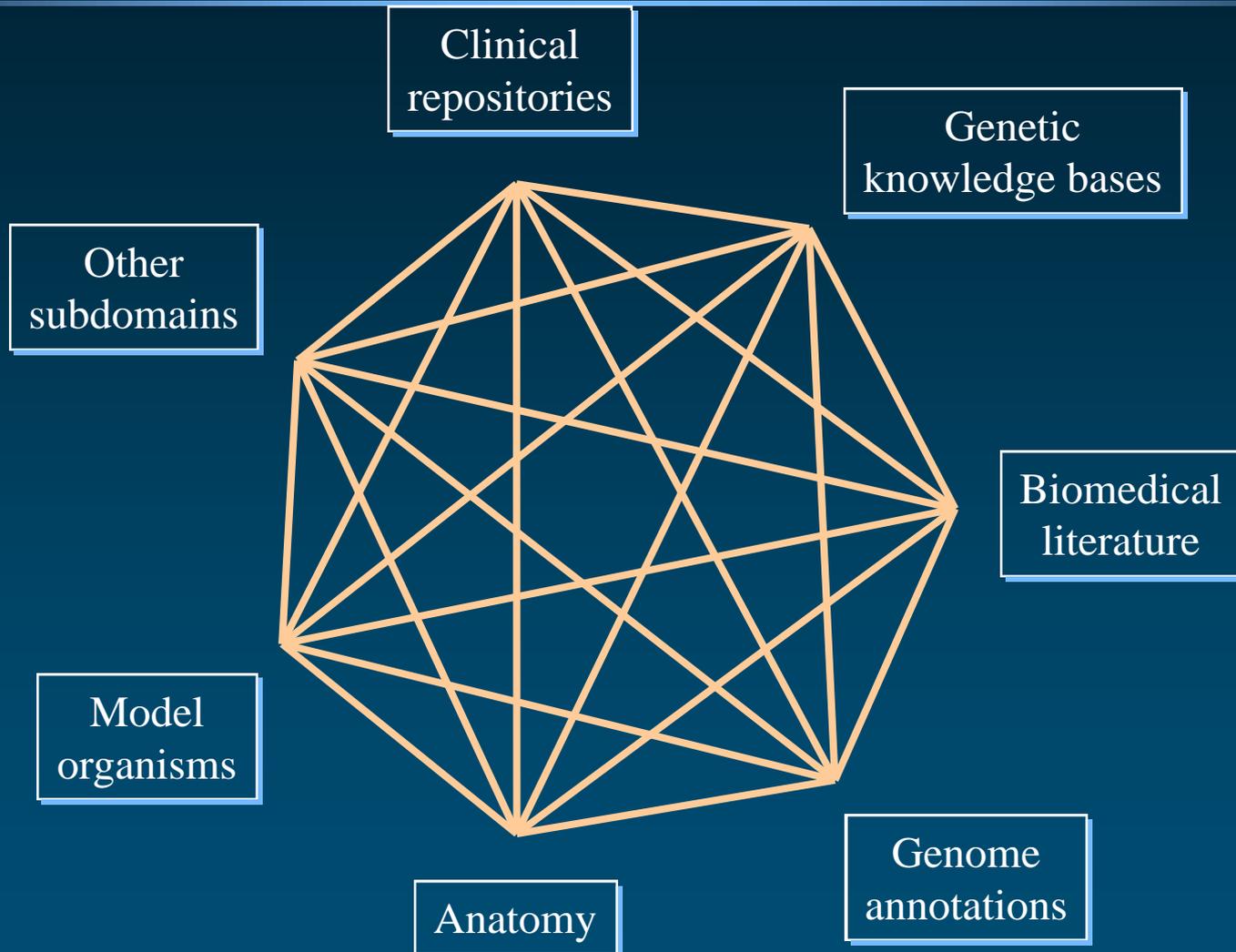
<http://obofoundry.org/>



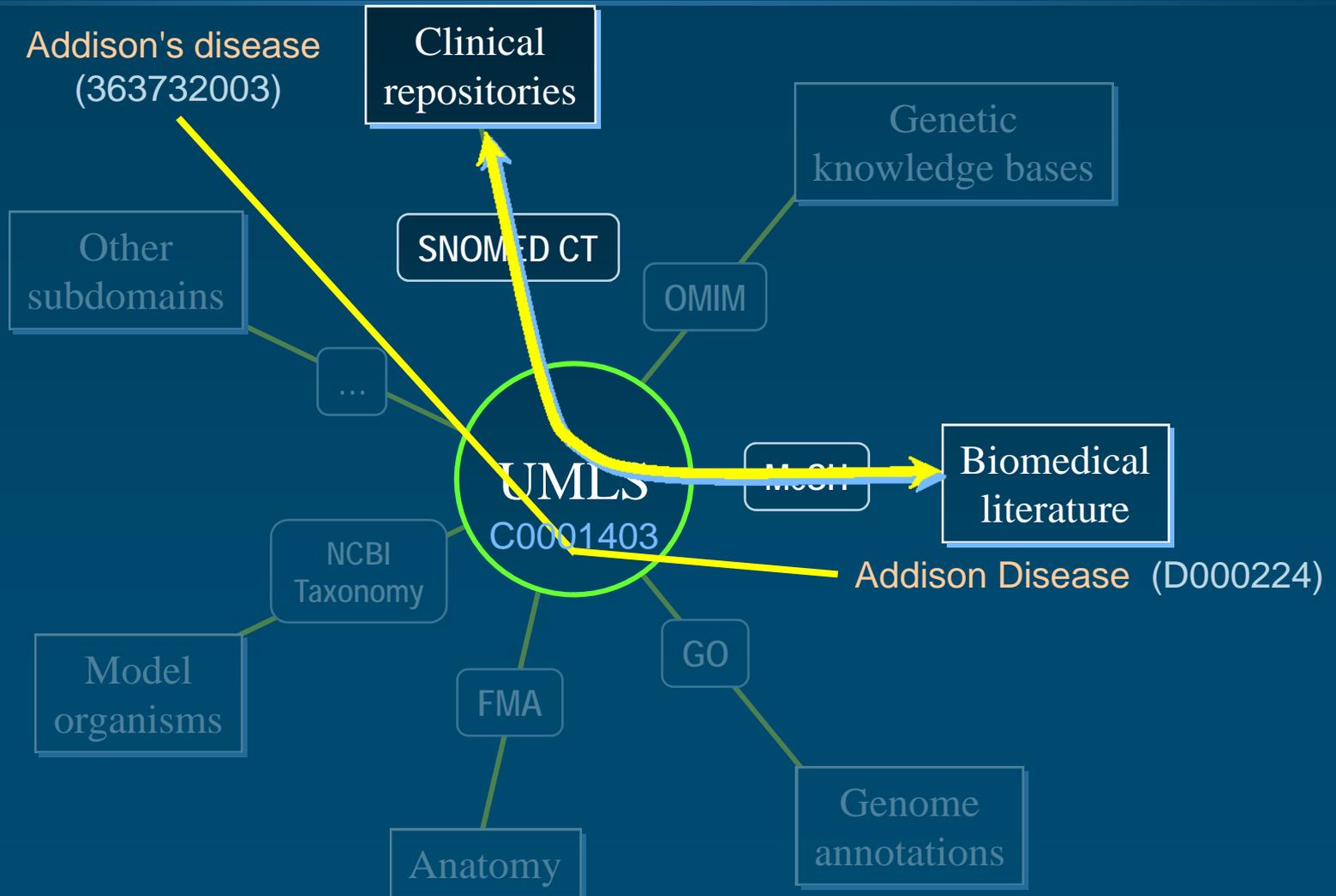
# Integrating subdomains



# 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

*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

*Other issues*

# Availability

- ◆ Many ontologies are freely available
- ◆ The UMLS is freely available for research purposes
  - Cost-free license required
- ◆ Licensing issues can be tricky
  - SNOMED CT is freely available in member countries of the IHTSDO
- ◆ Being freely available
  - Is a requirement for the Open Biomedical Ontologies (OBO)
  - Is a *de facto* prerequisite for Semantic Web applications



# Discoverability

- ◆ 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

# Formalism

## ◆ Several major formalism

- Web Ontology Language (OWL) – NCI Thesaurus
- OBO format – most OBO ontologies
- UMLS Rich Release Format (RRF) – UMLS, RxNorm

## ◆ Conversion mechanisms

- OBO to OWL
- LexGrid (import/export to LexGrid internal format)



# Ontology integration

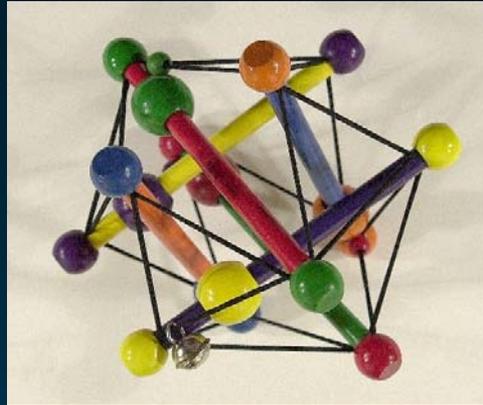
- ◆ *Post hoc* integration , form the bottom up
  - UMLS approach
  - Integrates ontologies “as is”, including legacy ontologies
  - Facilitates the integration of the corresponding datasets
- ◆ Coordinated development of ontologies
  - OBO Foundry approach
  - Ensures consistency *ab initio*
  - Excludes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

# Quality

- ◆ Quality assurance in ontologies is still imperfectly defined
  - Difficult to define outside a use case or application
- ◆ Several approaches to evaluating quality
  - Collaboratively, by users (Web 2.0 approach)
    - Marginal notes enabled by BioPortal
  - Centrally, by experts
    - OBO Foundry approach
- ◆ Important factors besides quality
  - Governance
  - Installed base / Community of practice

# Conclusions

- ◆ Ontologies are enabling resources for data integration
- ◆ Standardization works
  - Grass roots effort (GO)
  - Regulatory context (ICD 9-CM)
- ◆ Bridging across resources is crucial
  - Ontology integration resources / strategies (UMLS, BioPortal / OBO Foundry)
- ◆ Massive amounts of imperfect data integrated with rough methods might still be useful



# Medical Ontology Research

Contact: [olivier@nlm.nih.gov](mailto:olivier@nlm.nih.gov)

Web: [mor.nlm.nih.gov](http://mor.nlm.nih.gov)



*Olivier Bodenreider*

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