



Conference on
Semantics in Healthcare and Life Sciences
(CSHALS 2013)
Boston, MA
February 28, 2013

From biomedical information integration to knowledge discovery through the Semantic Web



Olivier Bodenreider

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

Semantic Web

- ◆ Extract information from structured and unstructured sources
 - From text: text mining
 - From ontologies and knowledge bases
- ◆ Integrate information
 - From structured and unstructured sources
- ◆ Aggregate information
 - Subsumption reasoning
- ◆ Use the extracted information for a meaningful purpose
 - Hypothesis generation / knowledge discovery
 - Better information retrieval
 - Question answering

Outline

- ◆ Knowledge, integration and aggregation
- ◆ Biomedical Knowledge Repository
- ◆ Towards a biomedical Semantic Web



KNOWLEDGE, INTEGRATION AND AGGREGATION

Definitional knowledge

◆ Definitional knowledge

- Universally true
- Examples
 - Lung cancer *has_location* Lung
 - Myocardial infarction *isa* Cardiovascular disease
 - Liver *part_of* Abdomen (canonical anatomy, in a given species)
- Typically found in ontologies
- Useful as background knowledge

Assertional knowledge

◆ Assertional knowledge

- True in a given context
- Examples
 - Aspirin *treats* headache
 - IL-13 *inhibits* COX2
 - Chest pain *manifestation_of* Myocardial infarction
 - Ciprofloxacin *causes* Tendon rupture
- Typically found in knowledge bases (and in text)
- Useful for knowledge discovery, question answering, biocuration support, etc.



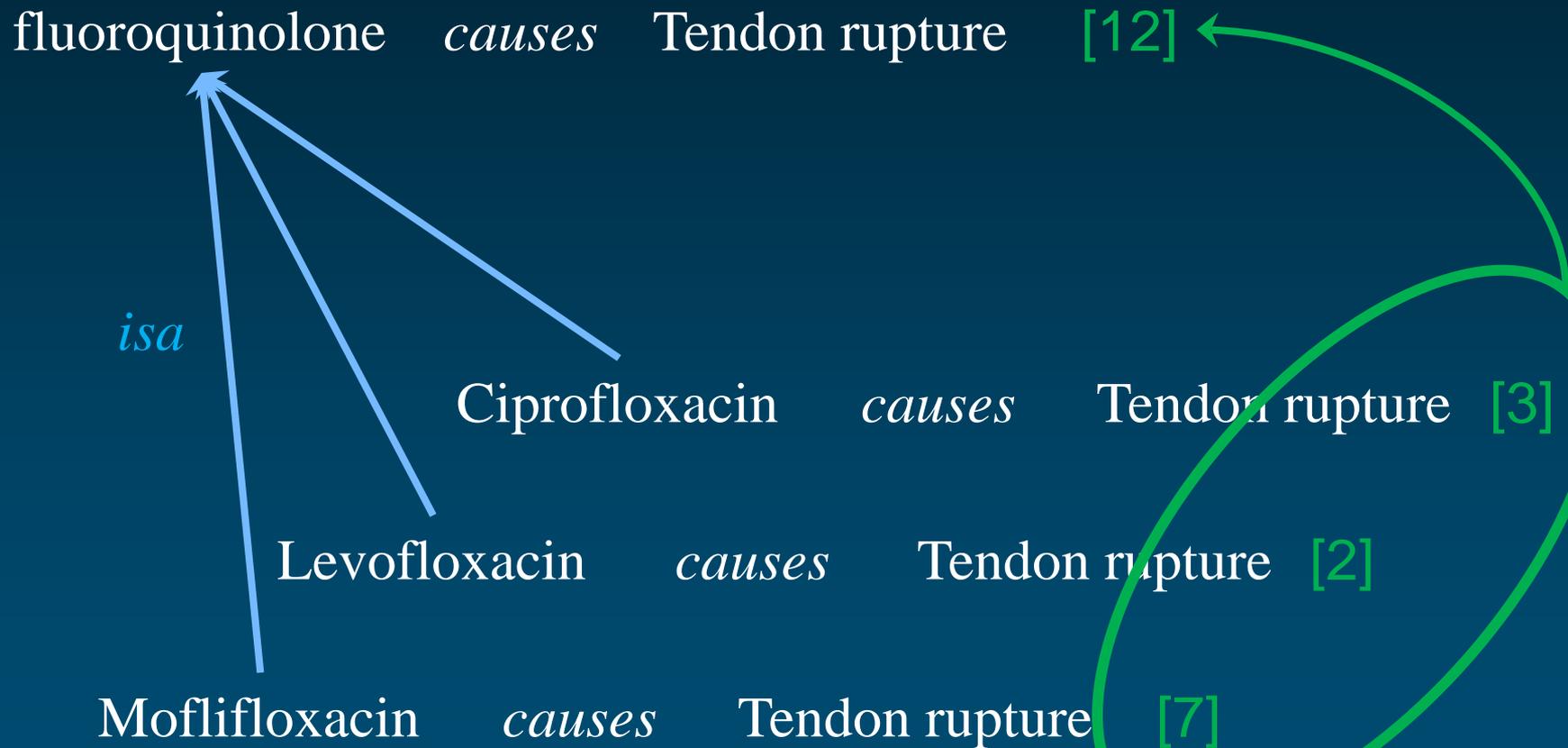
Definitional vs. assertional knowledge

- ◆ Definitional knowledge
 - Universally true
 - Typically found in ontologies
 - Useful as background knowledge
- ◆ Assertional knowledge
 - True in a given context
 - Typically found in knowledge bases (and in text)
 - Useful for knowledge discovery, question answering, biocuration support, etc.

Why integrate assertional and definitional knowledge?

- ◆ To increase statistical power
 - Low frequency for individual, fine-grained assertions
 - Higher frequency when frequencies are aggregated at a coarser level
- ◆ To bridge the granularity mismatch
 - Differences in granularity between
 - What is expressed in in text (or structured sources)
 - What is needed in “semantic mining” applications

Aggregating frequencies



Bridging the granularity mismatch

- ◆ A researcher is interested in glycosylation and its implications for one disorder: congenital muscular dystrophy.

Link between glycosyltransferase activity and congenital muscular dystrophy?



All Databases PubMed Nucleotide Protein Genome Structure PMC

Search Gene for 9215[uid] [Save Search](#)

Limits Preview/Index History Clipboard Details

Display Full Report Show 20 Send to

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

1: LARGE like-glycosyltransferase [*Homo sapiens*]
 GeneID: 9215 updated 02-Jul-2007

LARGE
(GeneID: 9215)

Phenotypes

has_associated_disease

Muscular dystrophy, congenital, type 1D
[MIM: 608840](#)

Congenital muscular dystrophy, type 1D Produced by GOA

GeneOntology

Function	Evidence
acetylglucosaminyltransferase activity	TAS PubMed

Process	Evidence
N-acetylglucosamine metabolic process	TAS PubMed
carbohydrate biosynthetic process	IEA
glycosphingolipid biosynthetic process	TAS PubMed
muscle maintenance	ISS
protein amino acid glycosylation	TAS PubMed

Component	Evidence
integral to Golgi membrane	TAS PubMed
integral to membrane	IEA
membrane	IEA



All Databases PubMed Nucleotide Protein Genome Structure PMC

Search Gene for 9215[uid] [Save Search](#)

Display Full Report Show 20 Send to

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

LARGE
(GeneID: 9215)

1: LARGE like-glycosyltransferase [*Homo sapiens*]

GeneID: 9215

updated 02-Jul-2007

Phenotypes

Muscular dystrophy, congenital, type 1D
[MIM: 608840](#)

GeneOntology

has_molecular_function

Provided by [GOA](#)

Function	Evidence
acetylglucosaminyltransferase activity	TAS PubMed

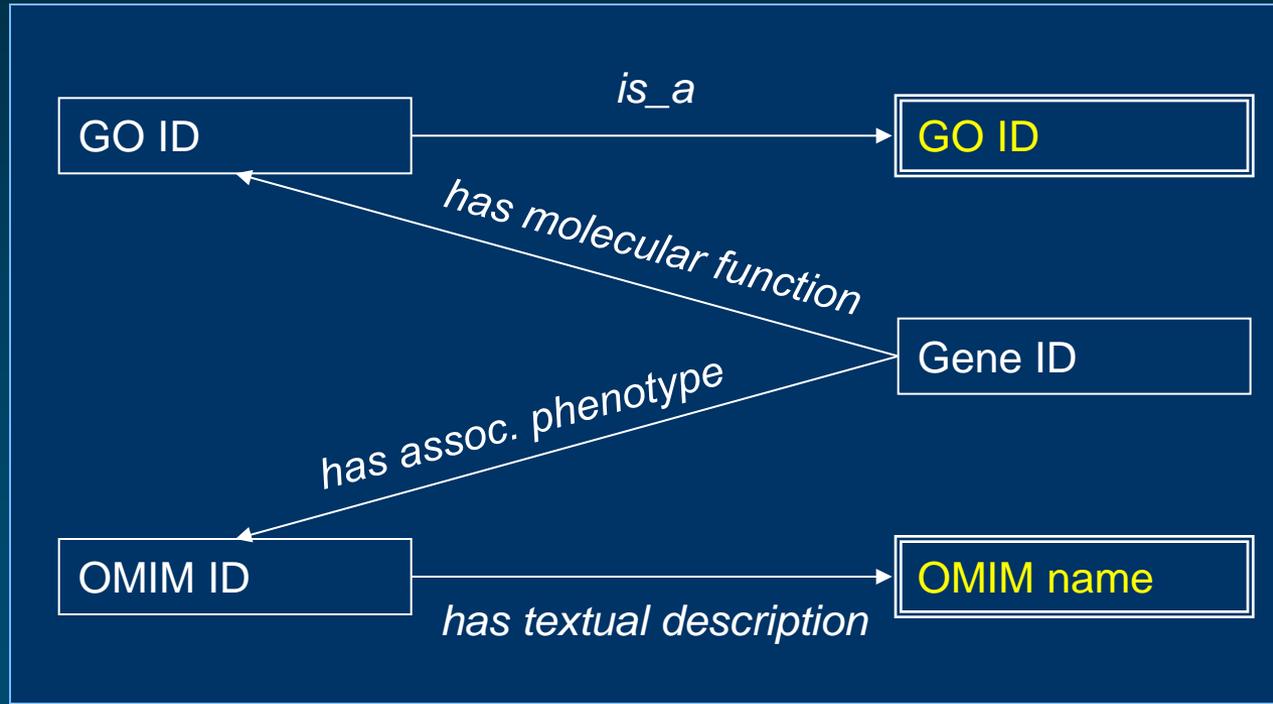
acetylglucosaminyltransferase activity

Process	Evidence
N-acetylglucosamine metabolic process	TAS PubMed
carbohydrate biosynthetic process	IEA
glycosphingolipid biosynthetic process	TAS PubMed
muscle maintenance	ISS
protein amino acid glycosylation	TAS PubMed

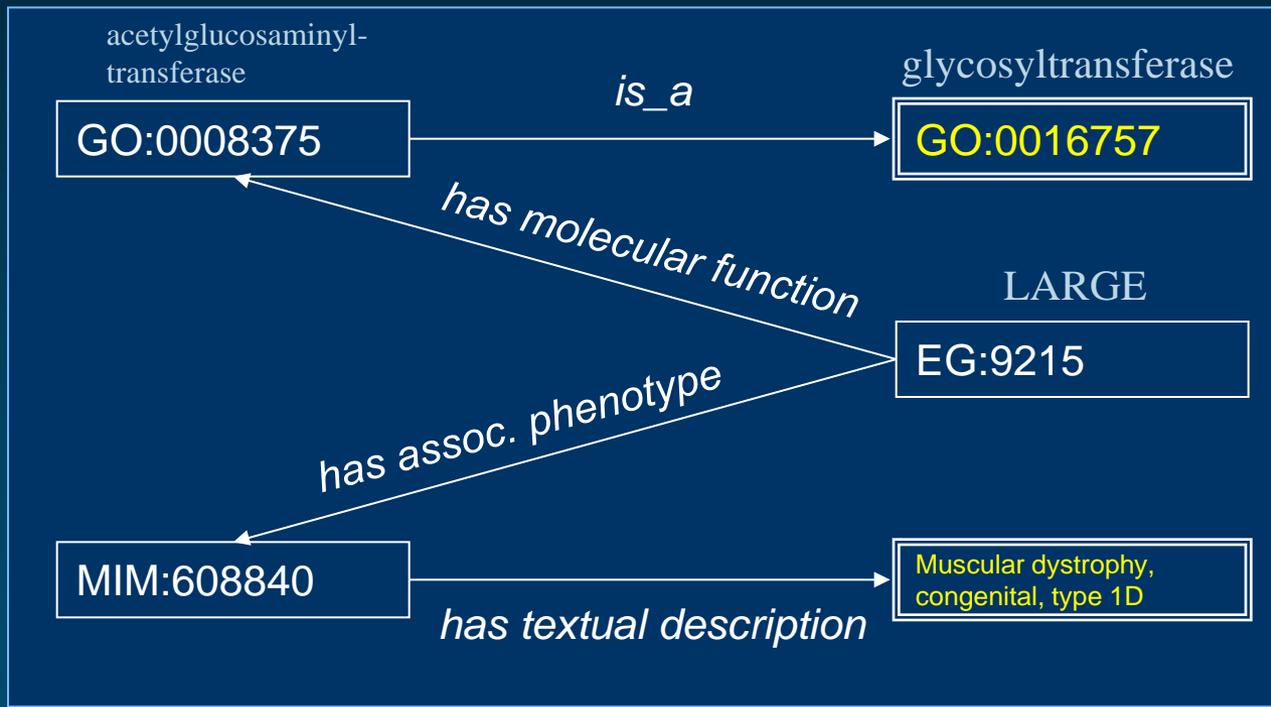
Component	Evidence
integral to Golgi membrane	TAS PubMed
integral to membrane	IEA
membrane	IEA

Using SPARQL to test a hypothesis

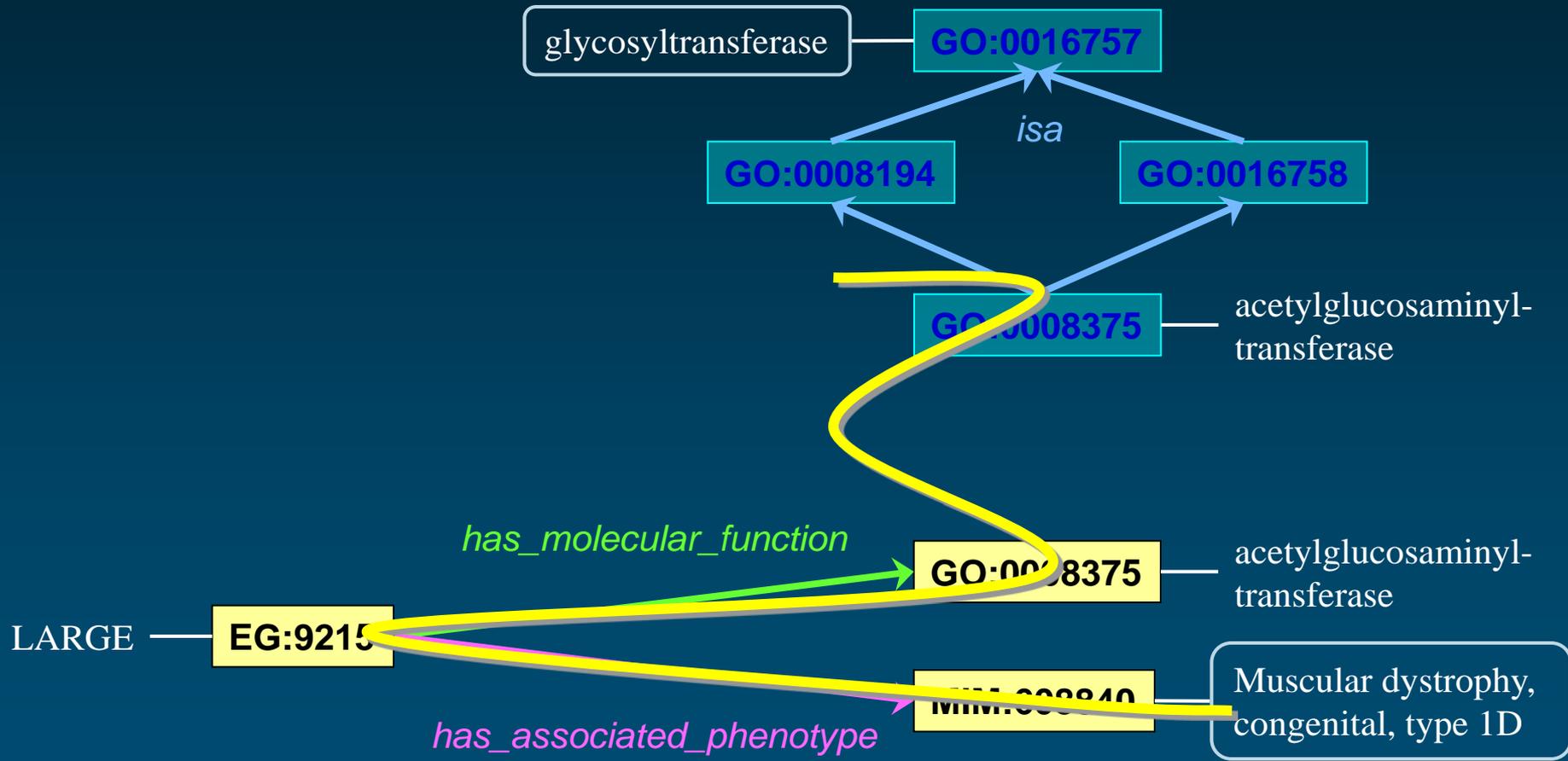
Find all the genes annotated with the GO molecular function glycosyltransferase or any of its descendants and associated with any form of congenital muscular dystrophy



Results Instantiated graph



From *glycosyltransferase* to *congenital muscular dystrophy*



NLM BIOMEDICAL KNOWLEDGE REPOSITORY

Biomedical Knowledge Repository

- ◆ Experimental resource
- ◆ Integrated set of relations
 - From the UMLS Metathesaurus
 - Extracted from MEDLINE by SemRep
- ◆ Together with metadata
 - Source of the relations (provenance)
- ◆ Semantic Web technologies
 - RDF store (Virtuoso)



Knowledge sources

- ◆ Ontologies – **definitional knowledge (mostly)**
 - Terminology integration systems
 - Unified Medical Language System (NLM)
 - BioPortal (NCBO)
- ◆ Relations extracted from text – **assertional knowledge (mostly)**
 - Text corpus
 - MEDLINE
 - Relation extraction system
 - SemRep (NLM), MedLEE (Columbia)
 - Commercial systems, specialized systems



Unified Medical Language System



◆ SPECIALIST Lexicon

- 460,000 lexical items
- Part of speech and variant information

◆ Metathesaurus

- 8.3M names from over 160 terminologies
- 2.9M concepts
- 16M relations

◆ Semantic Network

- 133 high-level categories
- 7000 relations among them

Lexical
resources

Terminological
resources

Ontological
resources



UMLS Metathesaurus

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

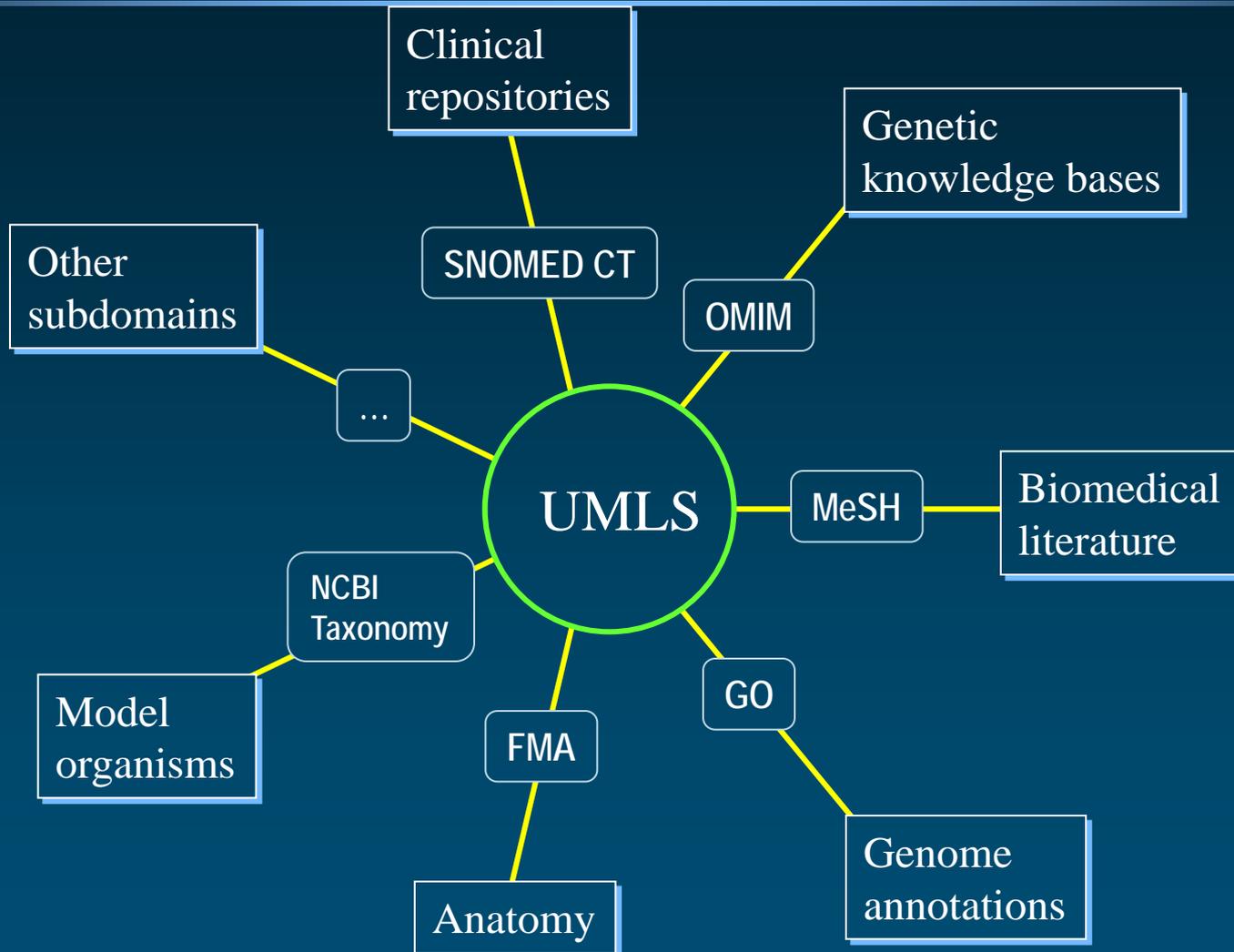
Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

C0001403

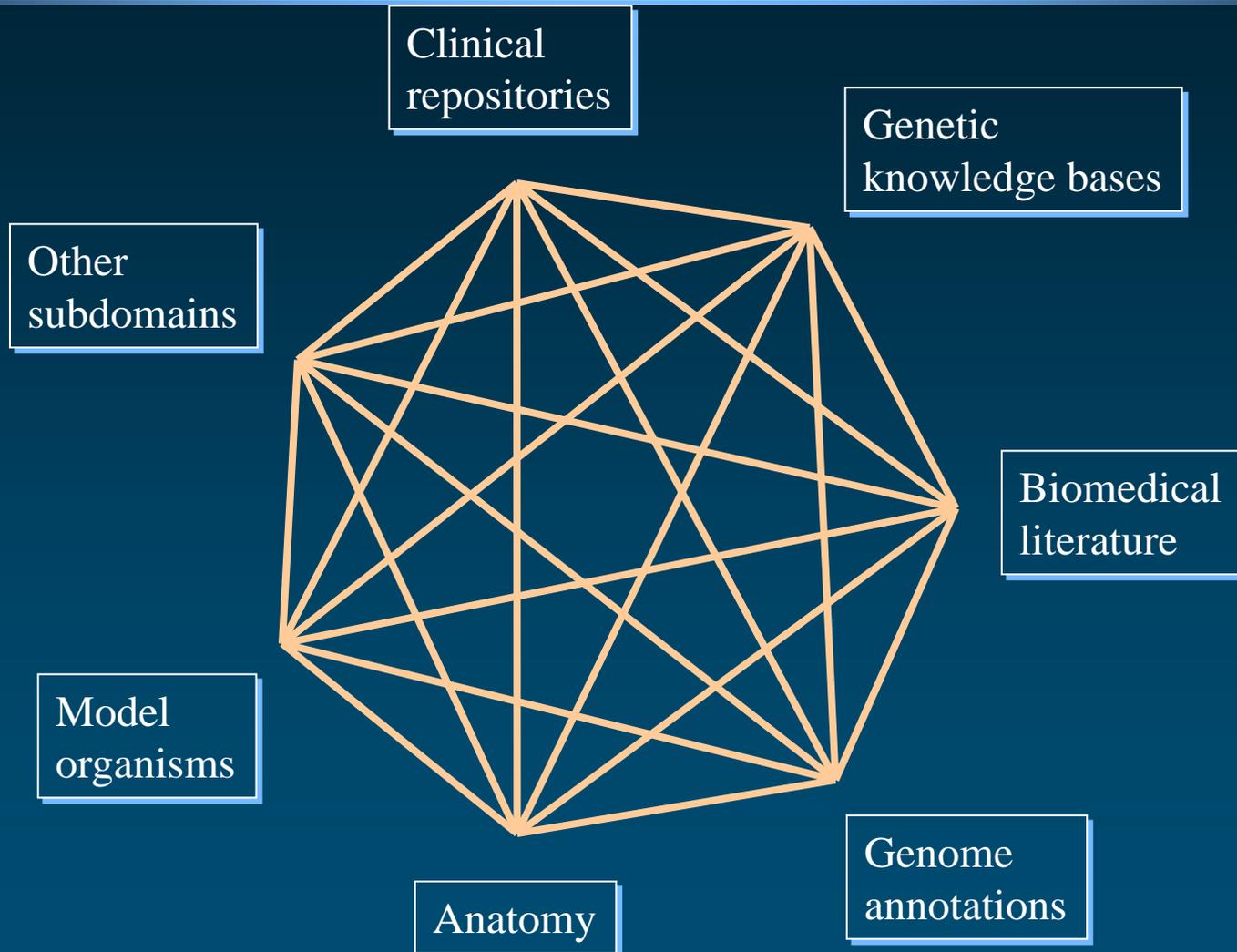
Addison's disease



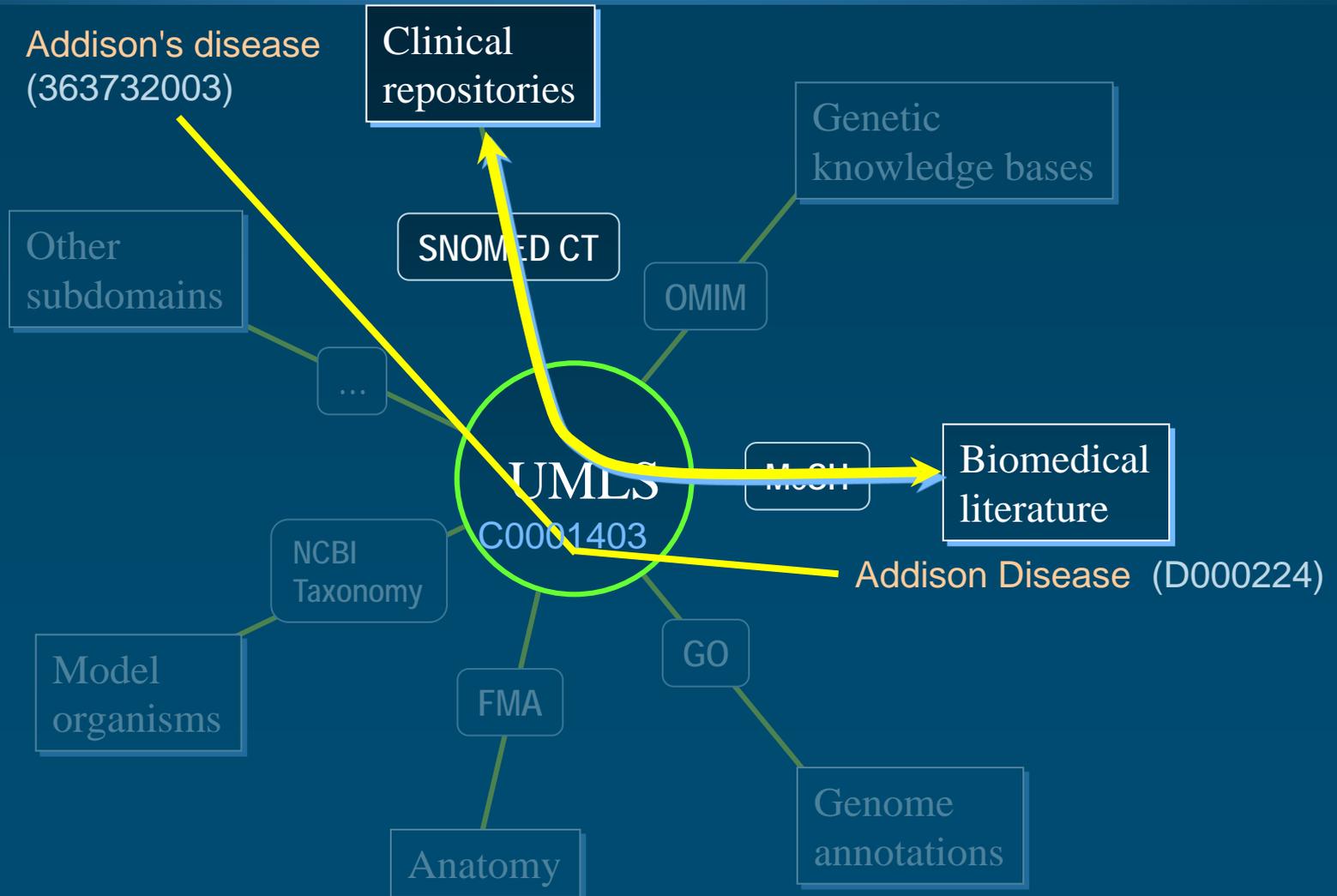
Integrating subdomains



Integrating subdomains



Trans-namespace integration



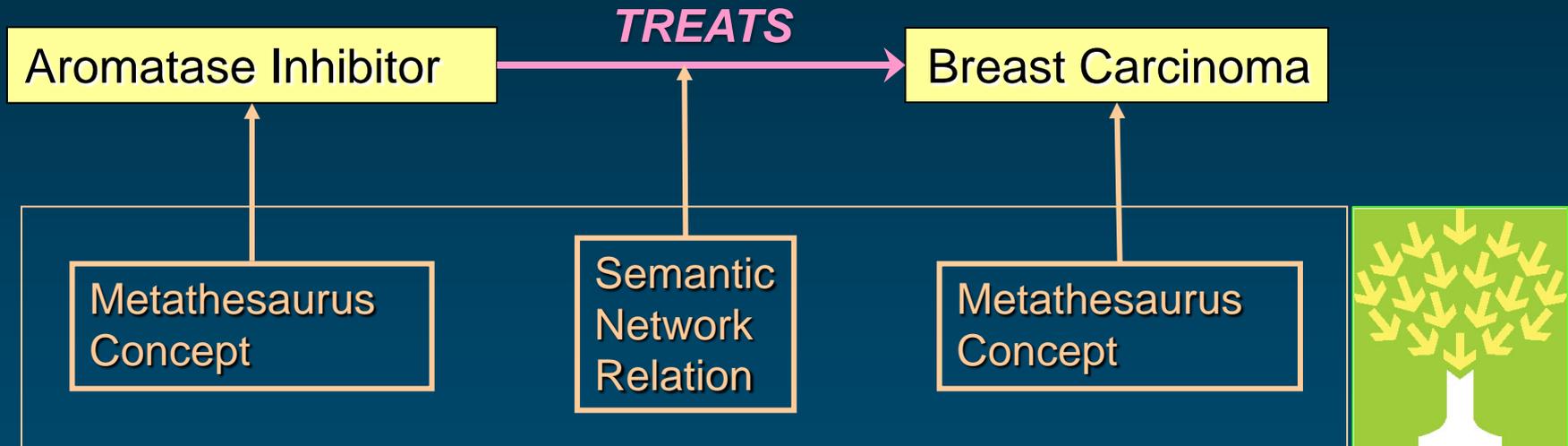
SemRep

- ◆ Part of the Semantic Knowledge Representation project at NLM
 - Tom Rindflesch & Marcelo Fiszman
- ◆ Knowledge extraction system for the automatic summarization system SemanticMEDLINE
 - <http://skr3.nlm.nih.gov/SemMedDemo/>
- ◆ Extract semantic predications from biomedical research literature (MEDLINE citations)



SemRep: Extract Predication

... Exemestane after non-steroidal aromatase inhibitor **for** post-menopausal women with advanced **breast cancer**

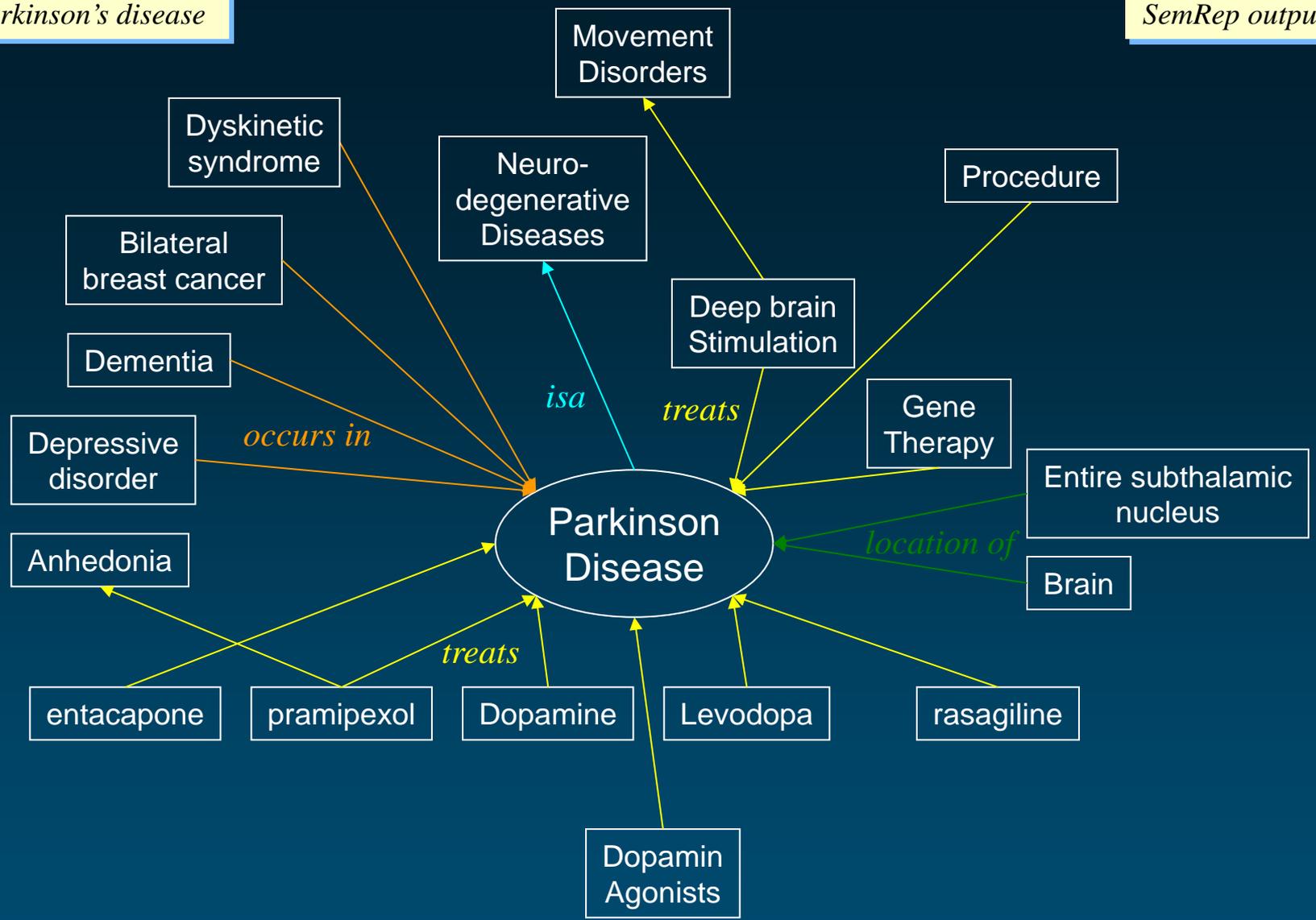


Unified Medical Language System

Predication Database: SemMedDB

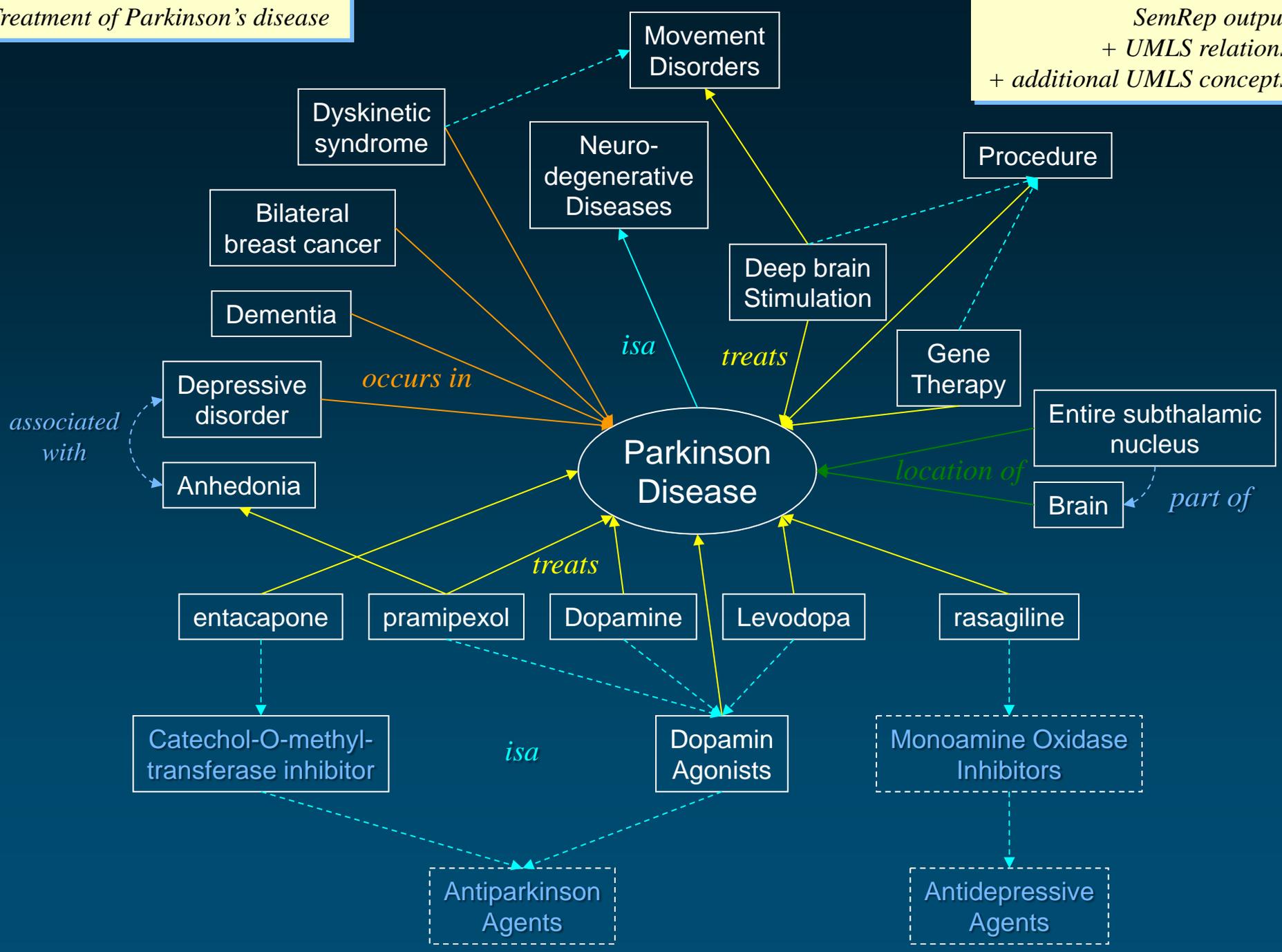
- ◆ Processed all of MEDLINE
 - More than 21 million citations
 - Titles and abstracts
- ◆ SemRep predications extracted
 - 57 million predications (through 06/30/2012)
- ◆ Made available to the research community
 - MySQL database
 - RDF triples





Treatment of Parkinson's disease

SemRep output
+ UMLS relations
+ additional UMLS concepts



Status

- ◆ Experimental
- ◆ Fully populated
 - UMLS 2012AA
 - 50M relations extracted from MEDLINE
- ◆ SemMedDB available for download
- ◆ UMLS in RDF not yet available for download
- ◆ Not available as a SPARQL endpoint
 - Licensing issues
 - Lack of access control in RDF stores



Potential applications

- ◆ Multi-document summarization
 - Semantic MEDLINE “plus”
- ◆ Information retrieval of relations
 - Beyond keywords or concepts
- ◆ Simple question answering
 - Which drugs treat congestive heart failure?
- ◆ Knowledge discovery
 - Swanson’s paradigm (e.g., finding “B”s)
 - Patterns of relations



TOWARDS A BIOMEDICAL SEMANTIC WEB

Challenges

- ◆ Linked data vs. Linked OPEN data
 - Intellectual property restrictions on some of the data sources
 - “UMLS license”
 - Privacy issues with clinical data
- ◆ Lack of Semantic Web awareness/interest from some data source / ontology providers
 - RDF versions produced by third parties
 - Inconsistent URIs
 - Inconsistent updates

Things are changing

- ◆ Data exposed through APIs
 - E.g., <http://www.nlm.nih.gov/api/>
- ◆ Linked Data Service
 - Library of Congress
 - Access to authority data
 - <http://id.loc.gov/>
- ◆ Aggressive “data liberation” initiatives
 - E.g., <http://healthdata.gov/>
- ◆ Common interface to ontologies
 - CTS2





ChenMed Transforming Health Care Delivery for Seniors



Innovative Care Models and Uses of Clinical Practice Data - the Future of Medicine

HHS Chief Technology Officer Bryan Sivak visited ChenMed to see how they have implemented an innovative care model and how clinical data is being used to improve practice. See what he learned from his trip. [Read more »](#)



HDI Starter Kit - Learn about all of the HHS data available to you.

[Get the Kit!](#)



Found a great health-related dataset on another site? Tell us about it!

[Suggest a Dataset](#)

Search the Data

Search for

Sub-Agency

Subject Area

[Search](#)

Recent Datasets

CMS 2008-2010 Data Entrepreneurs'...

CMS Data Navigator

Cross Federal Workgroup on Telehealth (...)

Child Support Enforcement Annual Data Report...

Child Welfare Monitoring Documents Library

[View more »](#)

Recent Blog Entries

CMS Launches Medicare Claims Synthetic...

Find the CMS data you are looking for with...

Dwayne Spradlin joins the Health Data...

Using CMS Data to Set Targets for ACOs

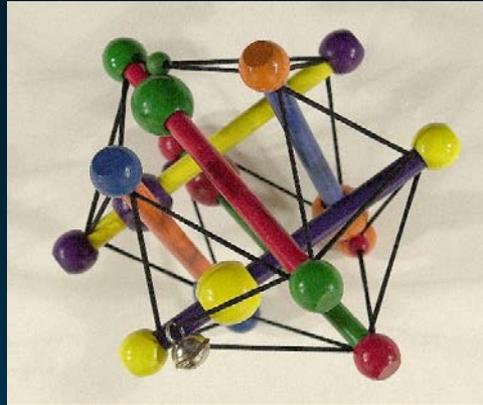
Requesting your help with TXT4Tots Text...

[View more »](#)

Biomedical Semantic Web

- ◆ Infrastructure for data integration
 - Definitional knowledge from ontologies
 - Assertional knowledge
 - From structured knowledge bases
 - Extracted through text mining
- ◆ Often requires semantic glue between datasets
 - UMLS, mappings
- ◆ Enabling technology for
 - Better information retrieval
 - Question answering
 - Hypothesis generation / knowledge discovery





Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: <http://mor.nlm.nih.gov>



Olivier Bodenreider

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