

Biomedical Linked Annotation Hackathon 3
Tokyo, Japan
January 16, 2017

Tying it all together

From terminology integration to annotation integration



Olivier Bodenreider

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



U.S. National Library of Medicine



Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.



Outline

- ◆ Normalizing annotations
 - From a text span to a concept identifier
- ◆ Reconciling annotations
 - Crosswalk between identifiers for equivalent concepts across ontologies
- ◆ Aggregating annotations
 - Bridging across the granularity divide
- ◆ Tying it all together
 - Two examples



Normalizing annotations

Neurofibromatosis 2

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

[Uppal, S., and A. P. Coatesworth. "Neurofibromatosis Type 2." *Int J Clin Pract*, 57, no. 8, 2003, pp. 698-703.]



MetaMap Example

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

C0254123

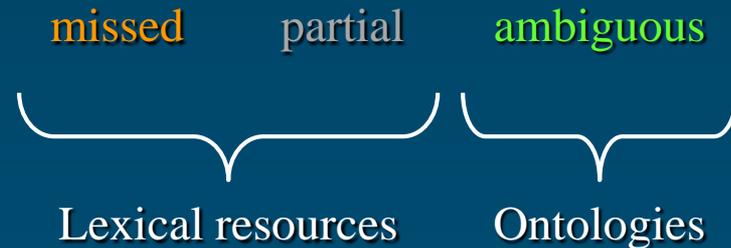


Neurofibromin 2	MeSH
Merlin	SNOMED CT
Schwannomin	MeSH
Schwannomerlin	NCI Thesaurus



Named entity recognition

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.



Annotation with MeSH

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

D016518 - Neurofibromatosis 2

D025581 - Neurofibromin 2



Annotation with OMIM

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

#101000 - NEUROFIBROMATOSIS, TYPE II; NF2

*607379 - NEUROFIBROMIN 2; NF2



Issues in normalizing annotations

- ◆ Which reference to normalize against?
 - Which individual ontology to use as the reference?
 - Sometimes dictated by the task
 - NCBO Ontology Recommender (coverage evaluation)
 - UMLS vs. individual ontologies
- ◆ Fine-grained vs. coarse annotations
 - Fine-grained annotations
 - Lossless
 - But will likely require aggregation
 - Coarse annotations
 - Lossy (i.e., not useful outside a specific project)
 - Unlikely to support linking to knowledge bases



Issues in normalizing annotations

◆ Missing annotations

- Controlled/reference terminologies vs. interface terminologies
 - “Missing” variants – most terminologies ignore variants on purpose
 - Lexical normalization / variant generation
- Leverage ontology integration systems (e.g., UMLS)
 - Benefit from synonyms (and variants) from all terminologies
 - Later restrict to a specific terminology



Issues in normalizing annotations

◆ Exact vs. partial mappings

- Spans may be more specific than ontology terms
 - bilateral vestibular schwannomas → vestibular schwannomas
- Controlled demodification supports partial mappings

◆ Ambiguous terms

- Ontologies provide context for disambiguation (hierarchies, semantic categorization)



Reconciling annotations

Annotation with MeSH

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

D016518 - Neurofibromatosis 2

D025581 - Neurofibromin 2



Annotation with OMIM

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

#101000 - NEUROFIBROMATOSIS, TYPE II; NF2

*607379 - NEUROFIBROMIN 2; NF2



Reconciling annotations

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

D016518 - Neurofibromatosis 2

#101000 - NEUROFIBROMATOSIS, TYPE II; NF2

D025581 - Neurofibromin 2

*607379 - NEUROFIBROMIN 2; NF2



Reconciling annotations UMLS



- ◆ *C0027832* - Neurofibromatosis 2
 - *D016518* - Neurofibromatosis 2
 - *#101000* - NEUROFIBROMATOSIS, TYPE II; NF2
 - (Type II neurofibromatosis, Bilateral acoustic neurofibromatosis, ...)
- ◆ *C0254123* - Merlin
 - *D025581* - Neurofibromin 2
 - **607379* - NEUROFIBROMIN 2; NF2
 - (Schwannomin, Neurofibromin 2, ...)



UMLS Metathesaurus

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

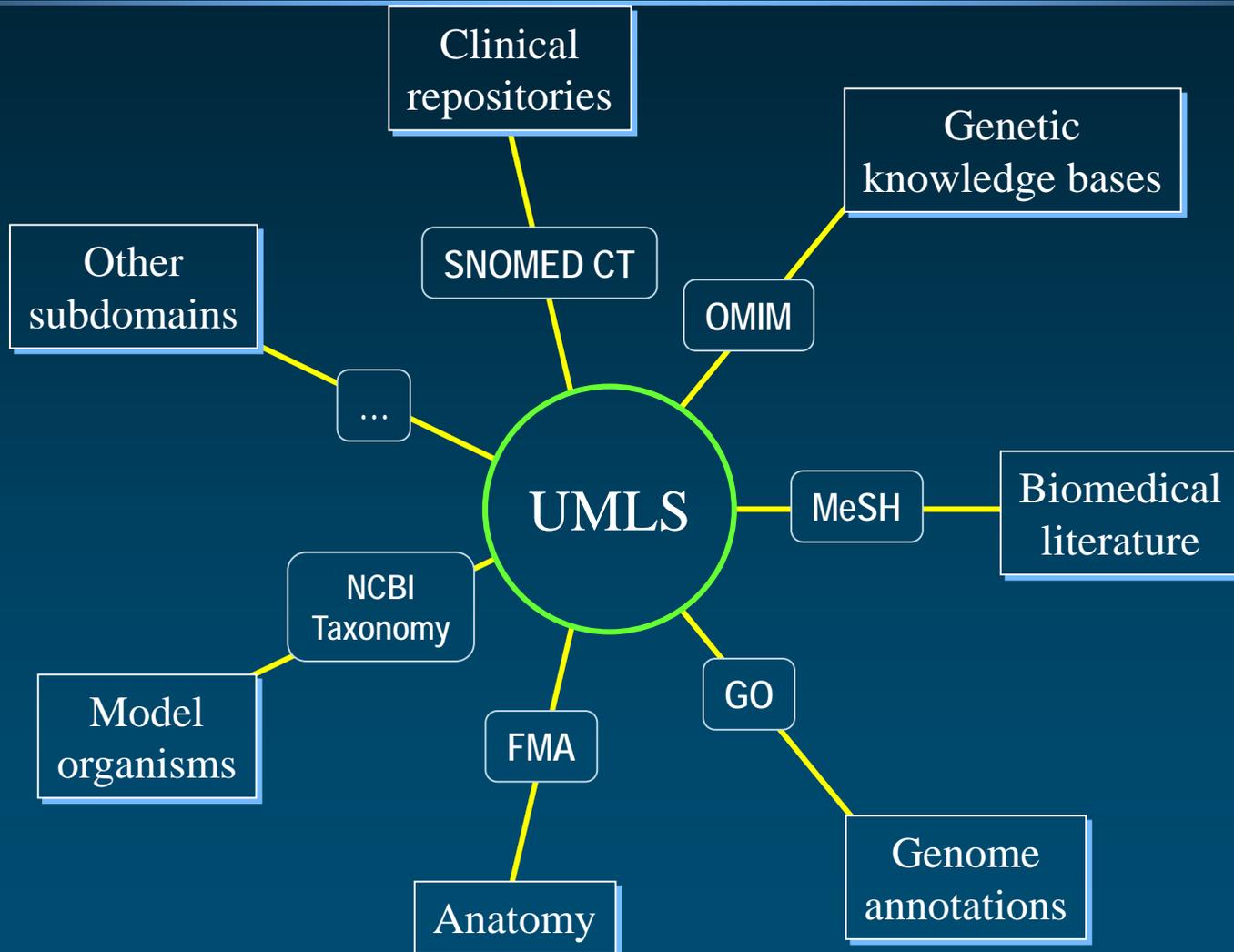
Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

C0001403

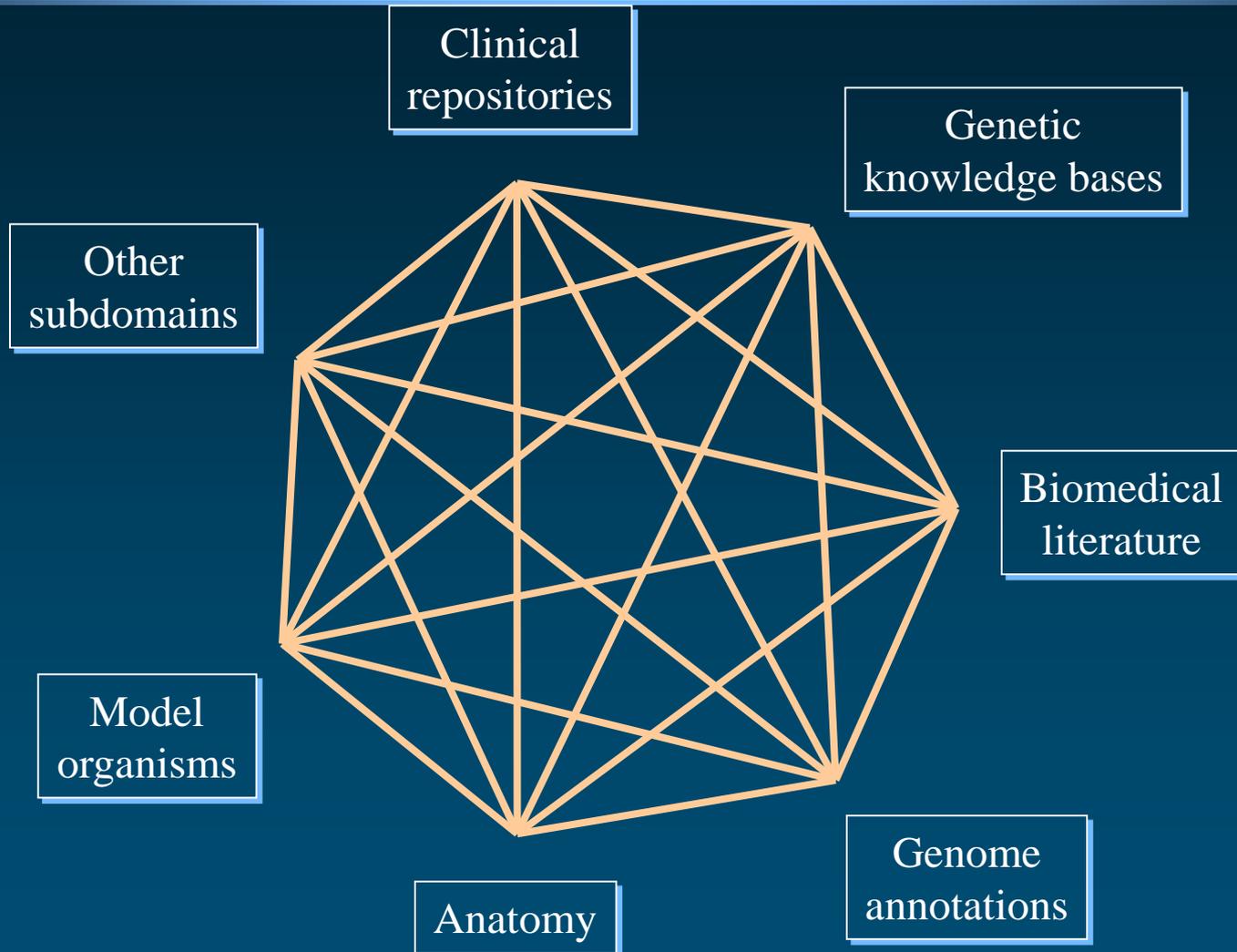
Addison's disease



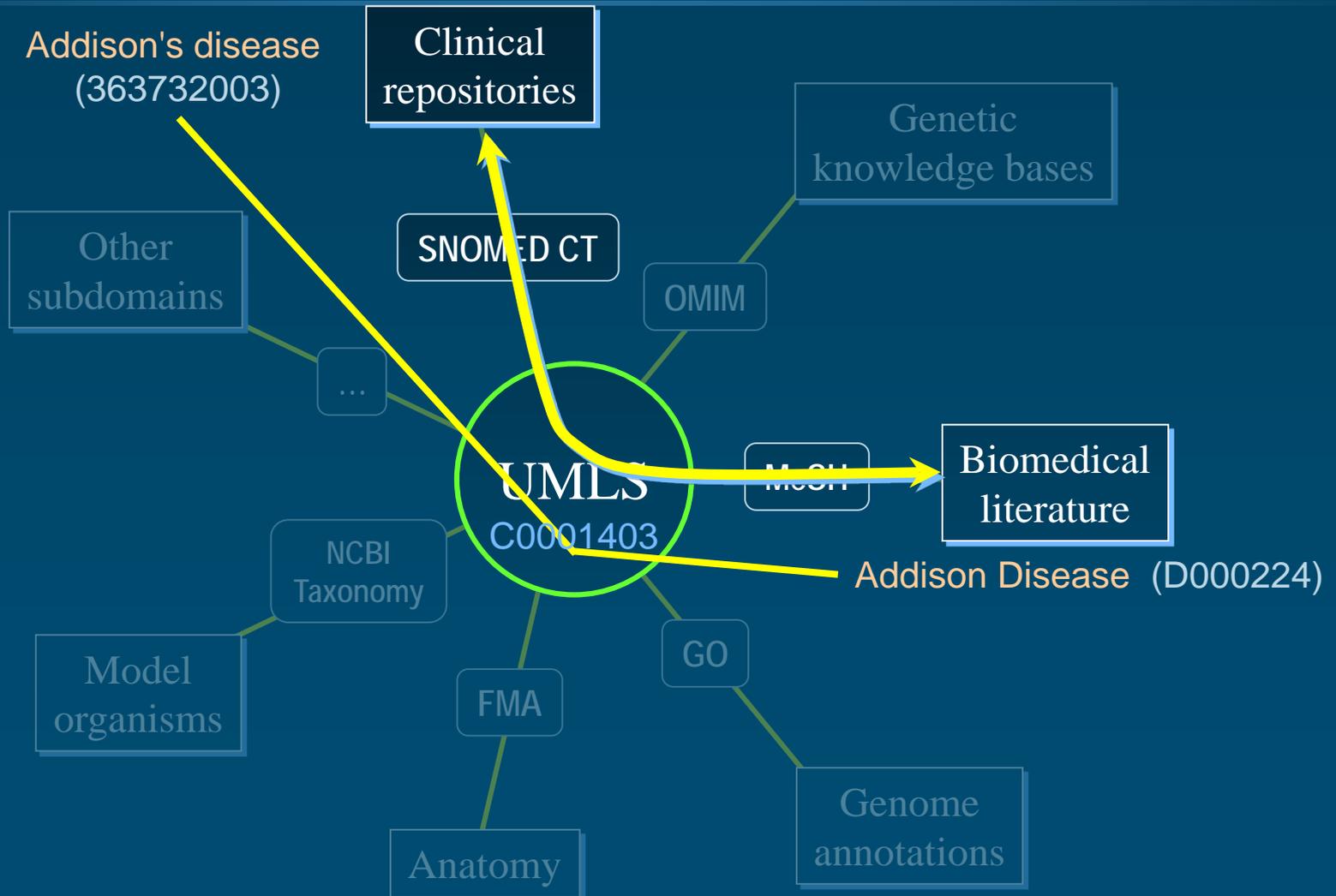
Integrating subdomains



Integrating subdomains



Trans-namespace integration



Source Vocabularies

(2016AA)

- ◆ 150 families of source vocabularies
 - Not counting translations
- ◆ Broad coverage of biomedicine
 - 9.9M names (normalized)
 - ~3.2M concepts
 - > 10M relations
- ◆ Mappings are curated by the Metathesaurus editors



Other ontology integration systems

◆ General

- NCBO BioPortal
 - Over 500 ontologies integrated
 - Mappings across ontologies are not curated
- EBI Ontology Lookup Service
 - OBO ontologies

◆ Domain-specific

- Entrez Gene
 - Integrates names and identifiers for genes
- [...]

Official Symbol	NF2 provided by HGNC
Official Full Name	neurofibromin 2 provided by HGNC
Primary source	HGNC:HGNC:7773
Also known as	ACN; SCH; BANF



Synonymy vs. mapping

◆ Synonymy

- Terms grouped under the same concept
- Best option possible for integrating annotations

◆ Point-to-point mappings across sources

- Developed for a given purpose
- Generally meant to be used in one direction
- May reflect synonymy or small semantic distance between terms
- Might be useful for integrating annotations



Issues in reconciling annotations

- ◆ What is synonymy in terminologies?
 - Splitting vs. lumping
 - “Concept orientation” is in the eye of the beholder
 - MeSH descriptors vs. UMLS concepts
- ◆ Provenance of the annotation (metadata)
 - Distinguish between normalizing and reconciling annotations
 - Keep the original identifier the annotation was made to
 - Including specific version of the ontology and NER system
 - To enable alternative reconciliation if needed

Aggregating annotations

Reconciliation vs. Aggregation

◆ Reconciliation

- Recognize when different term identifiers refer to the same entity (or concept)
- Leveraging synonymy and mapping relations

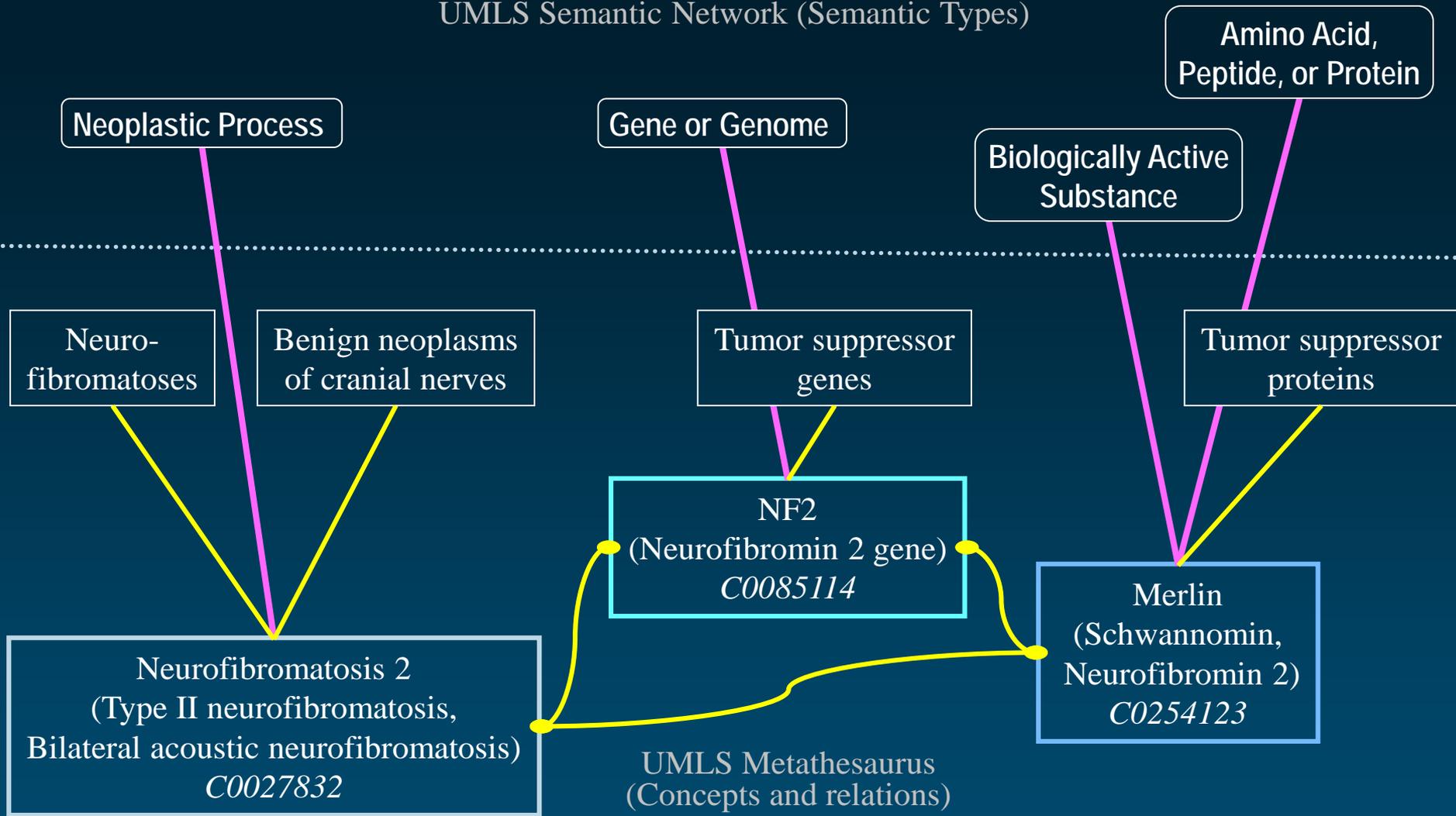
[D016518 - Neurofibromatosis 2
#101000 - NEUROFIBROMATOSIS, TYPE II; NF2

◆ Aggregation

- Bridge across various levels of granularity
- Leveraging hierarchical relations
- From fine-grained annotations to higher-level annotations
 - Neurofibromatosis 1, Neurofibromatosis 2 → Neurofibromatoses



UMLS Semantic Network (Semantic Types)



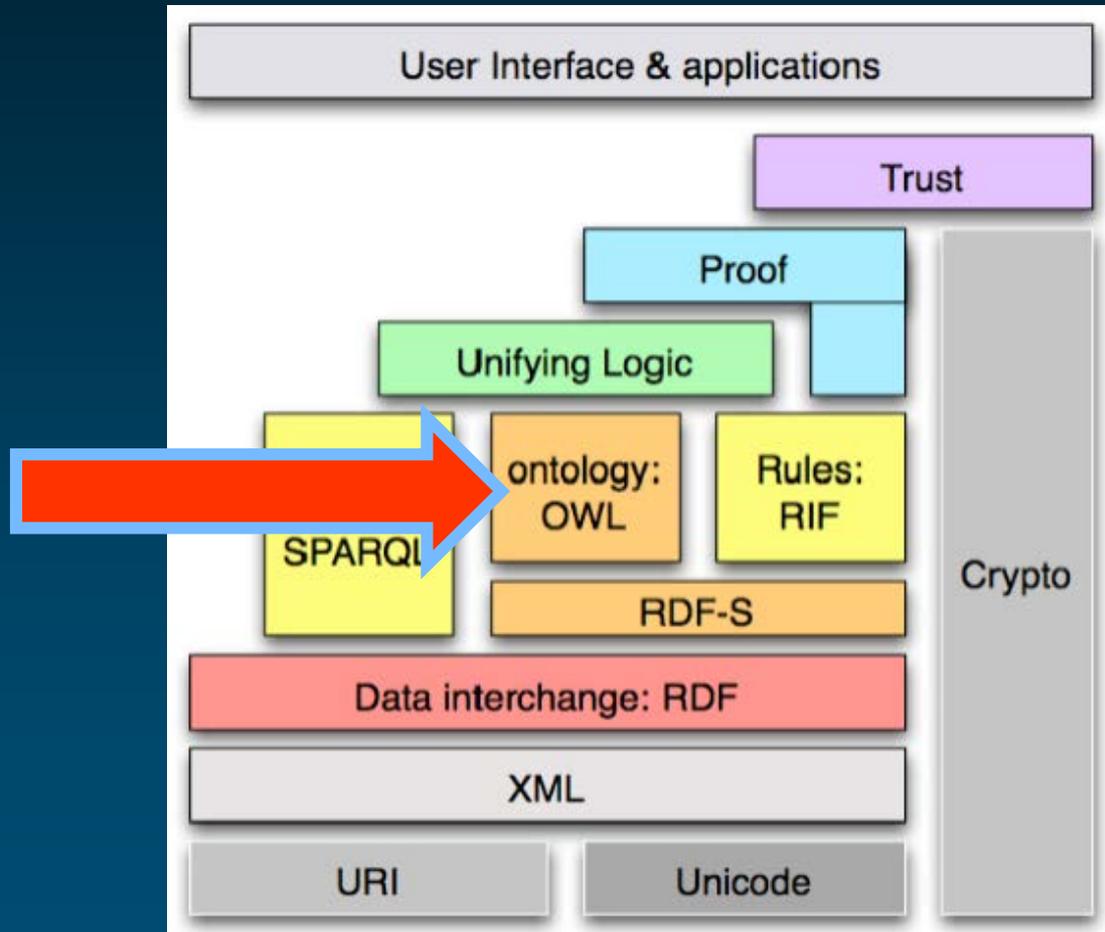
Aggregation

- ◆ Mostly supported by hierarchical relations in ontologies
 - Definitional knowledge (universally true)
- ◆ Required for bridging the granularity mismatch when integrating annotations
 - Between annotations made at different levels of granularity
 - Applicable to various sources
 - Between literature annotations
 - Between literature annotations and knowledge bases

Aggregation and linked data

- ◆ Aggregation is key to linking annotations
 - Hierarchical links among annotations
- ◆ Annotations made in reference to ontologies and these ontologies must be integrated together to support knowledge discovery
 - Linked Data provides a platform
 - Ontologies have been a core component of the Semantic Web historically

Semantic Web “layer cake”



Issues in aggregating annotations

- ◆ Hierarchical relations vs. relations used to organize concepts in trees
 - Hierarchical relations
 - Partial order relations
 - Support subsumption inference (subClassOf)
 - Relations used to organize concepts in trees
 - Not always subClassOf relations
 - LLT to PT in MedDRA (synonymy, lexical variation, subclass)
 - MeSH hierarchy (“aboutness” for retrieval purposes)
 - [...]
- ◆ Hierarchical relations in the UMLS Metathesaurus
 - Not curated, possibly conflicting



Issues in aggregating annotations

◆ Semantic distance

- Can be used to assess when annotations can be aggregated
- Edge counting is a poor surrogate for semantic distance
 - Wide variation in hierarchical depth among ontologies
- Information content-based approaches require frequencies of occurrence
- Large body of literature on the subject

Tying it all together

Example #1

Reasoning with annotations from Entrez Gene

Bridging the granularity mismatch

Example from Entrez Gene annotations

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

Link between glycosyltransferase activity and congenital muscular dystrophy?

[Sahoo, Medinfo 2007]





All Databases PubMed Nucleotide Protein Genome Structure PMC

Search Gene for 9215[uid] [Go](#) [Clear](#) [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

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](#)

Limits Preview/Index History Clipboard Details

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

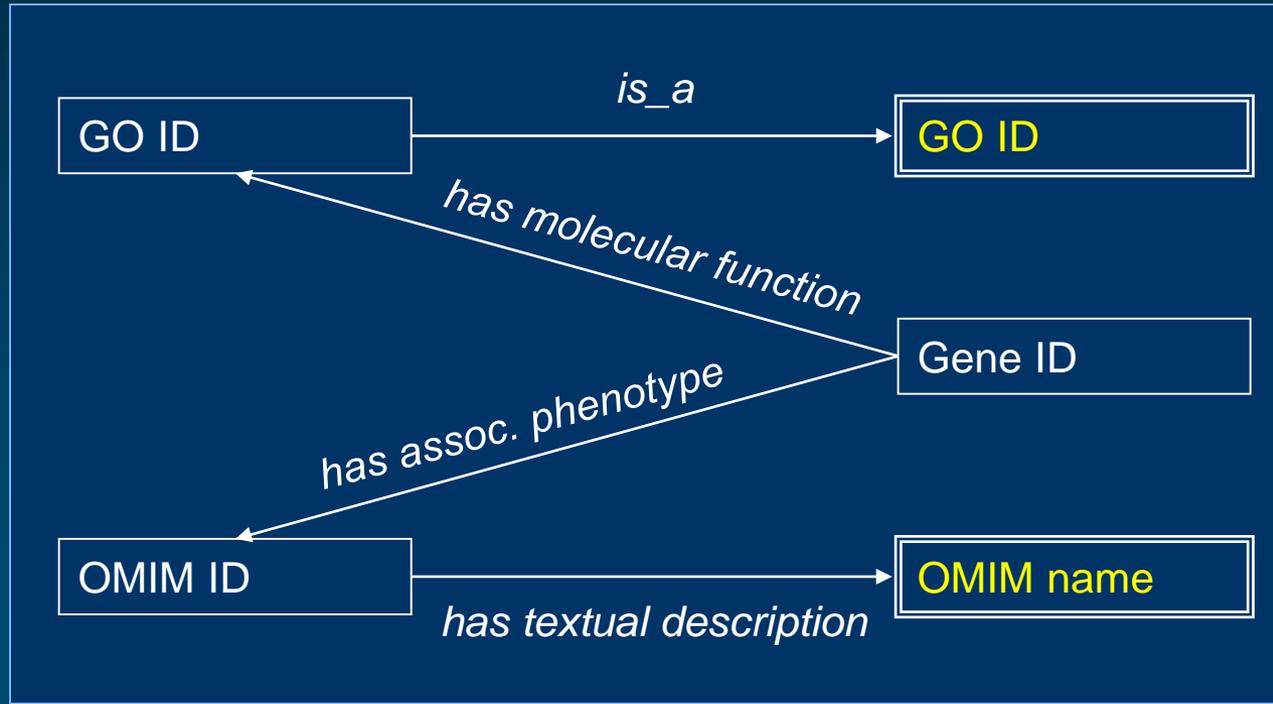
acetylglucosaminyltransferase activity

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

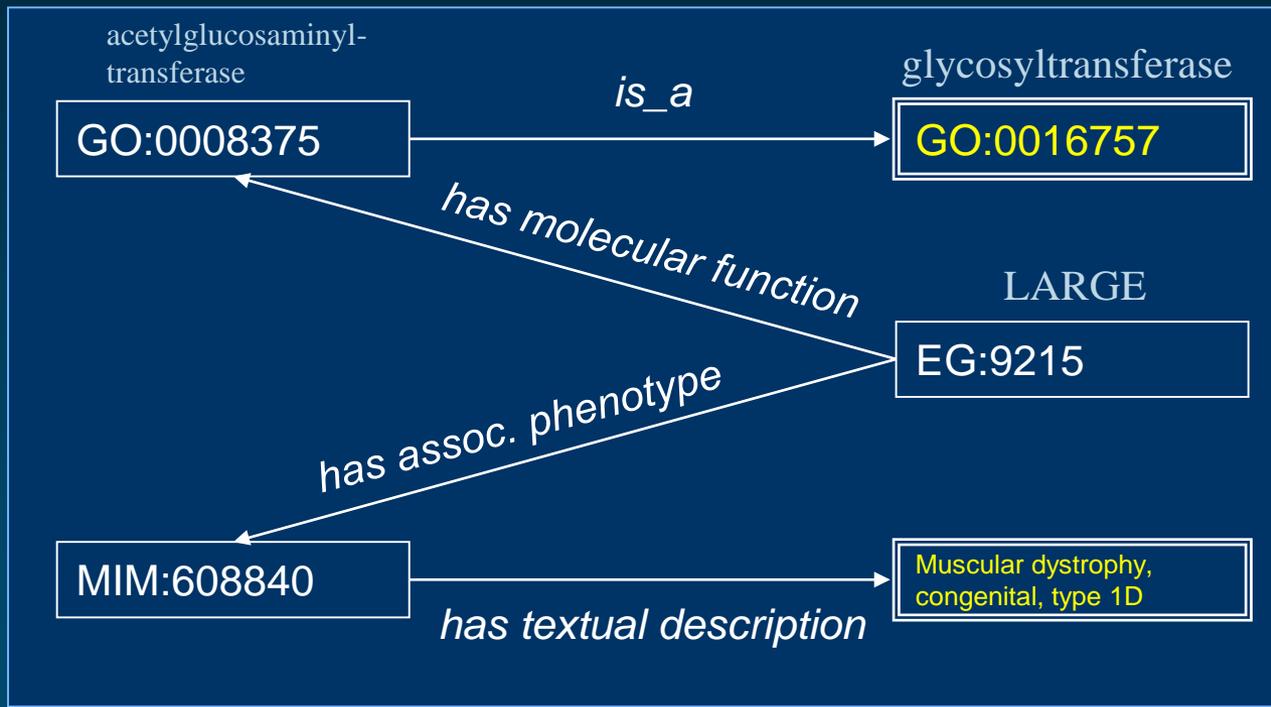
Component	Evidence
integral to Golgi membrane	TAS PubMed
integral to membrane	IEA
membrane	IEA

Using SPARQL to test a hypothesis

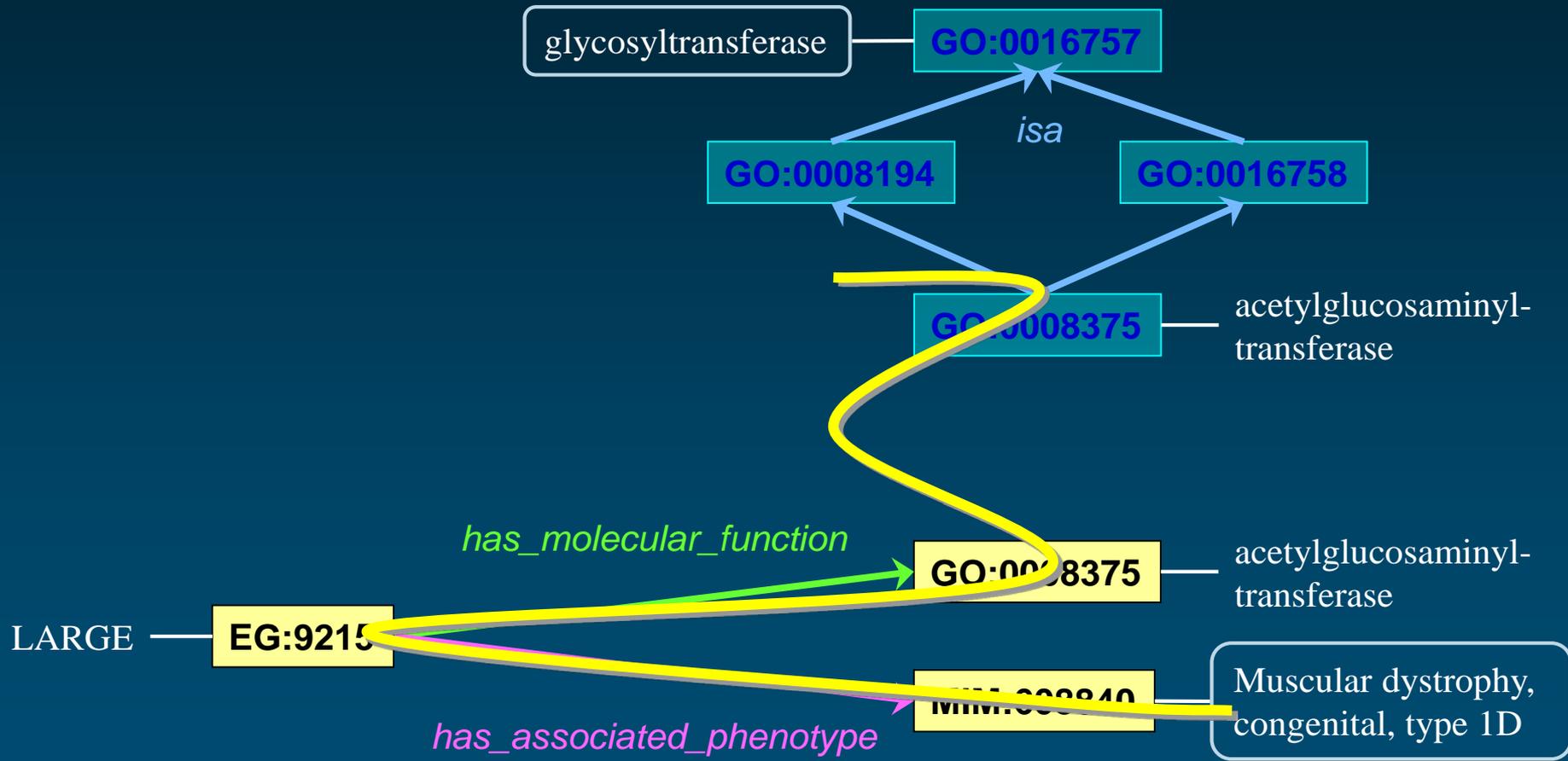
Find all the genes annotated with the GO molecular function glycosyltransferase or any of its descendants and associated with any form of congenital muscular dystrophy



Results Instantiated graph



From *glycosyltransferase* to *congenital muscular dystrophy*



Tying it all together

Example #2

Organizing annotations from SemMedDB

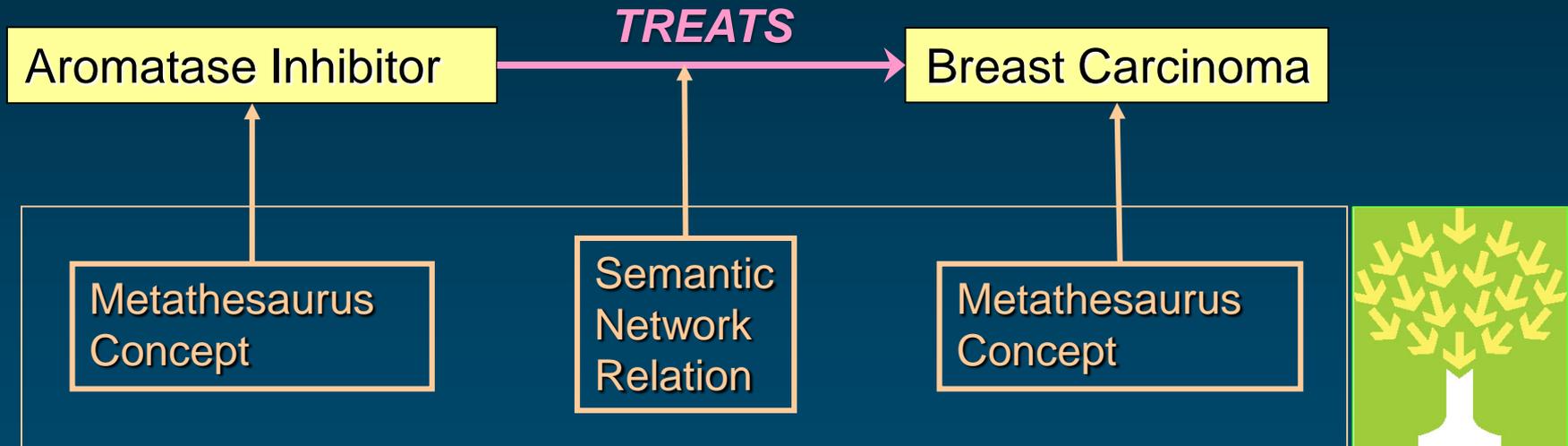
SemRep

- ◆ Relation extraction system (semantic predications)
- ◆ Part of the Semantic Knowledge Representation project at NLM
 - Tom Rindflesch
- ◆ Applied to the biomedical literature (MEDLINE citations)
- ◆ Supports the automatic summarization system, Semantic Medline



SemRep: Extract Predication

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



Unified Medical Language System

Predication Database: SemMedDB

- ◆ SemRep predications extracted
 - From titles and abstracts in MEDLINE
 - 80 million predications
 - Normalized to UMLS
- ◆ Made available to the research community
 - MySQL database
 - RDF triples



Status

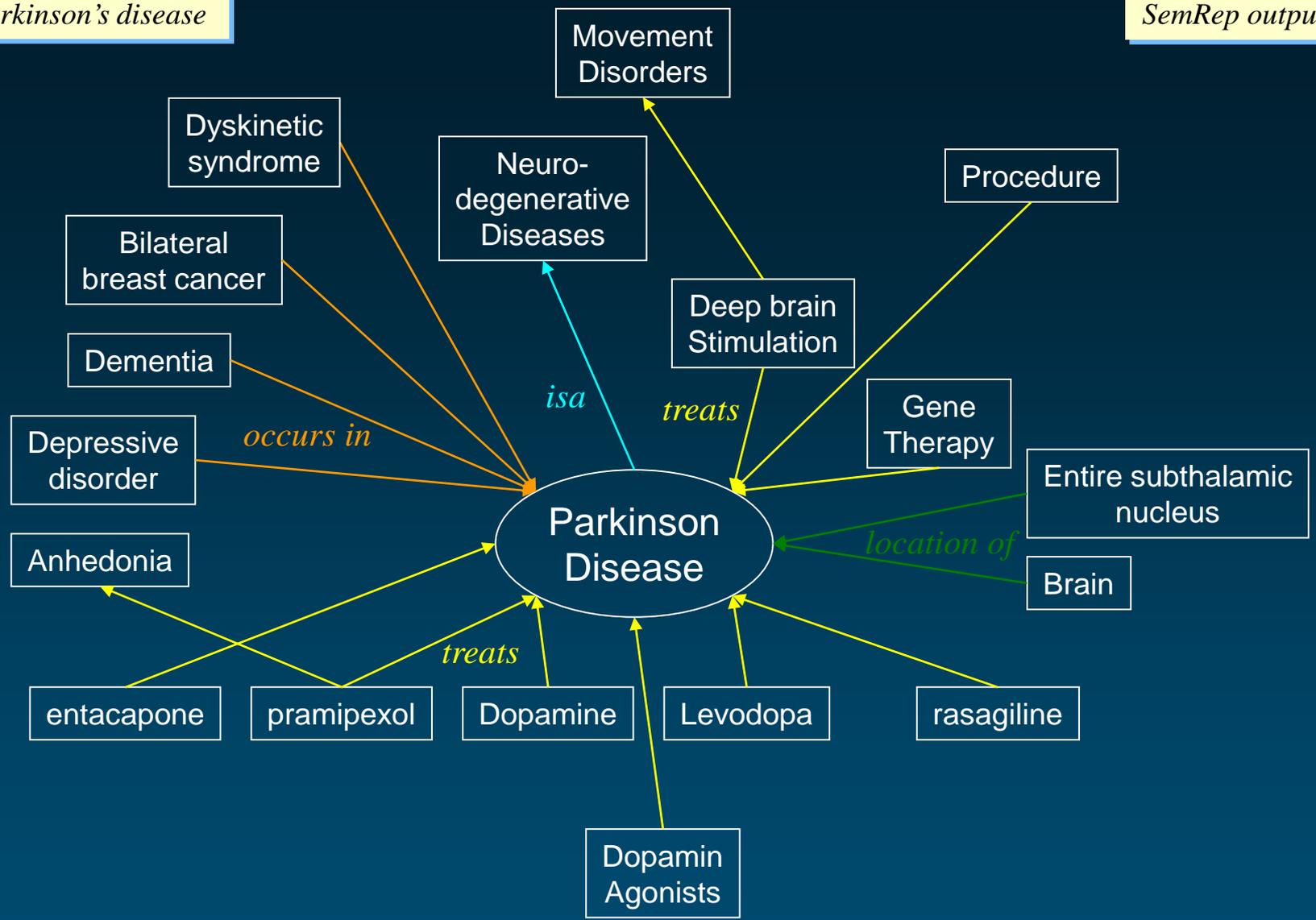
◆ Publicly available

- Semantic Medline graphical interface
 - <https://skr3.nlm.nih.gov/SemMed/>
- SemMedDB predication database (download)
 - <https://skr3.nlm.nih.gov/SemMedDB/>

◆ Experimental integration with UMLS relations

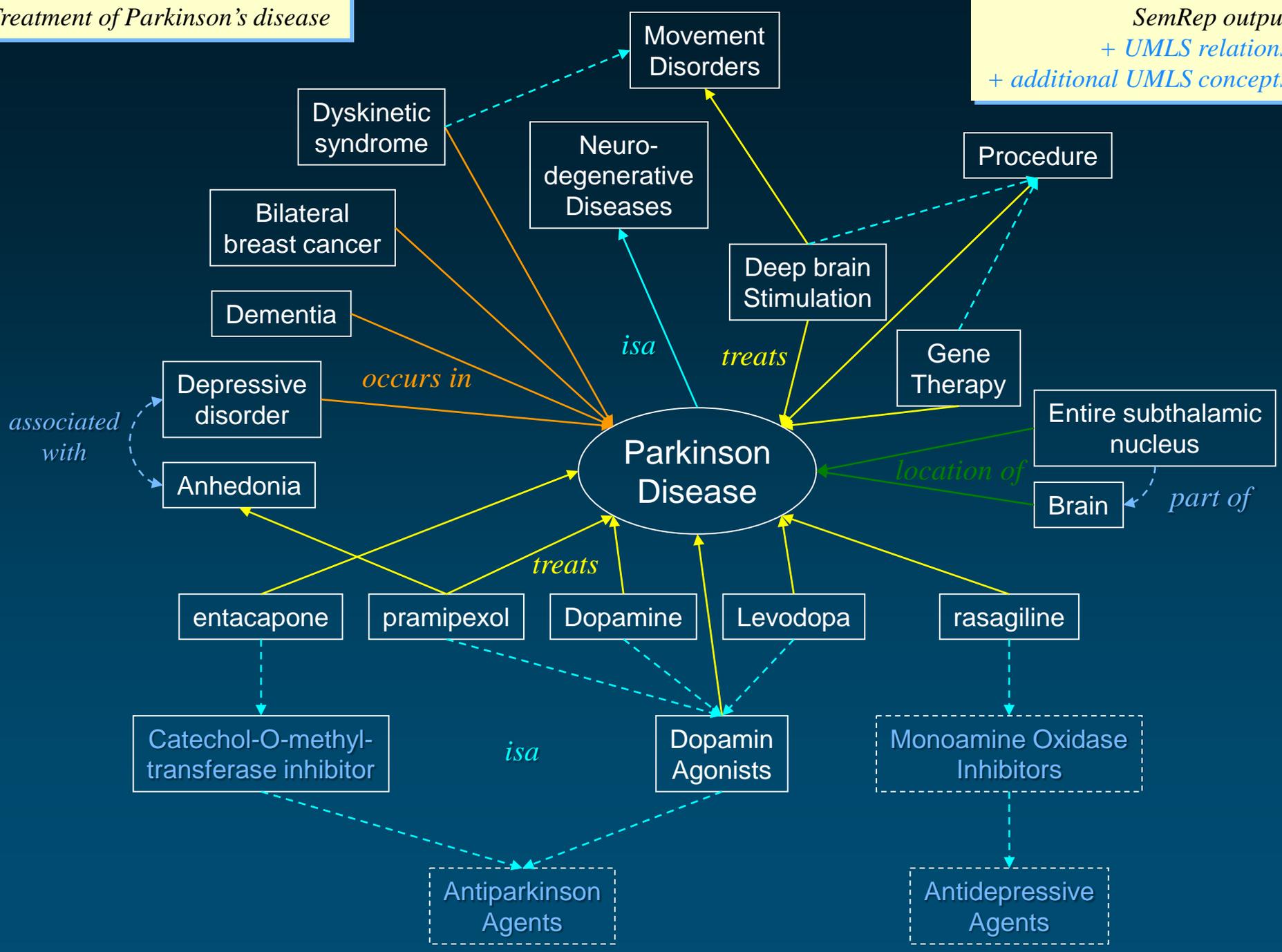
- UMLS in RDF not yet available for download
- Not available as a SPARQL endpoint
 - Licensing issues
 - Lack of access control in RDF stores





Treatment of Parkinson's disease

SemRep output
+ UMLS relations
+ additional UMLS concepts



Summary

◆ Ontologies

- Are key to annotation normalization, reconciliation and aggregation
- Are a core component of the Semantic Web, including Linked Data

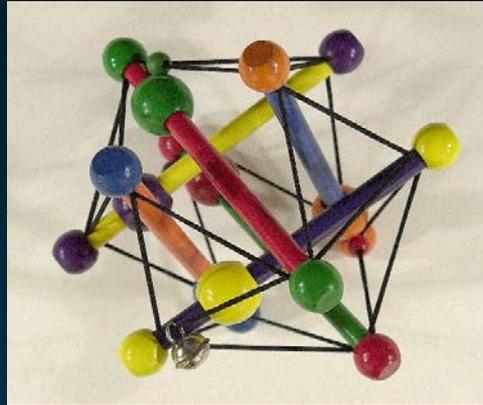
◆ Ontology integration systems

- Can be leveraged to support annotation integration

◆ Annotation integration in action

- Support for hypothesis generation and knowledge discovery





Medical Ontology Research

Contact: olivier@nlm.nih.gov

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



Olivier Bodenreider

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