

3rd Annual i2b2 AUG Meeting & NLP Workshop
Boston, MA
June 18, 2013

NLM Resources for Mining Biomedical Text



Olivier Bodenreider

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



U.S. National Library of Medicine



Overview

- ◆ An example
- ◆ Types of resources for mining biomedical text
- ◆ Three types of resources
 - Lexical resources
 - Terminological resources
 - Ontological resources



An example

Neurofibromatosis 2

Neurofibromatosis 2

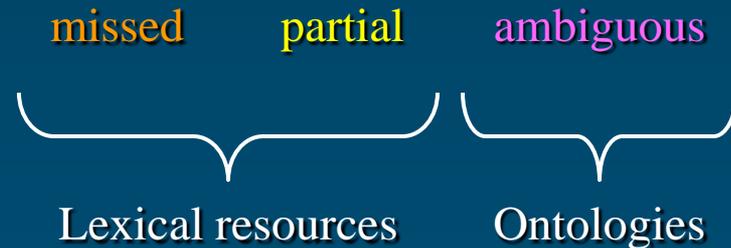
Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

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



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.



Relation extraction

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.

- vestibular schwannomas *manifestation of* neurofibromatosis 2
- neurofibromatosis 2 *associated with* mutation of NF2 gene
- NF2 gene *located on* chromosome 22

Types of resources for mining biomedical text

Types of resources

◆ Lexical resources

- Collections of lexical items
- Additional information
 - Part of speech
 - Spelling variants
- Useful for entity recognition
- UMLS SPECIALIST Lexicon, WordNet

◆ Ontological resources

- Collections of
 - kinds of entities (substances, qualities, processes)
 - relations among them
- Useful for **relation extraction**
- UMLS Semantic Network, BioTop



◆ Terminological resources

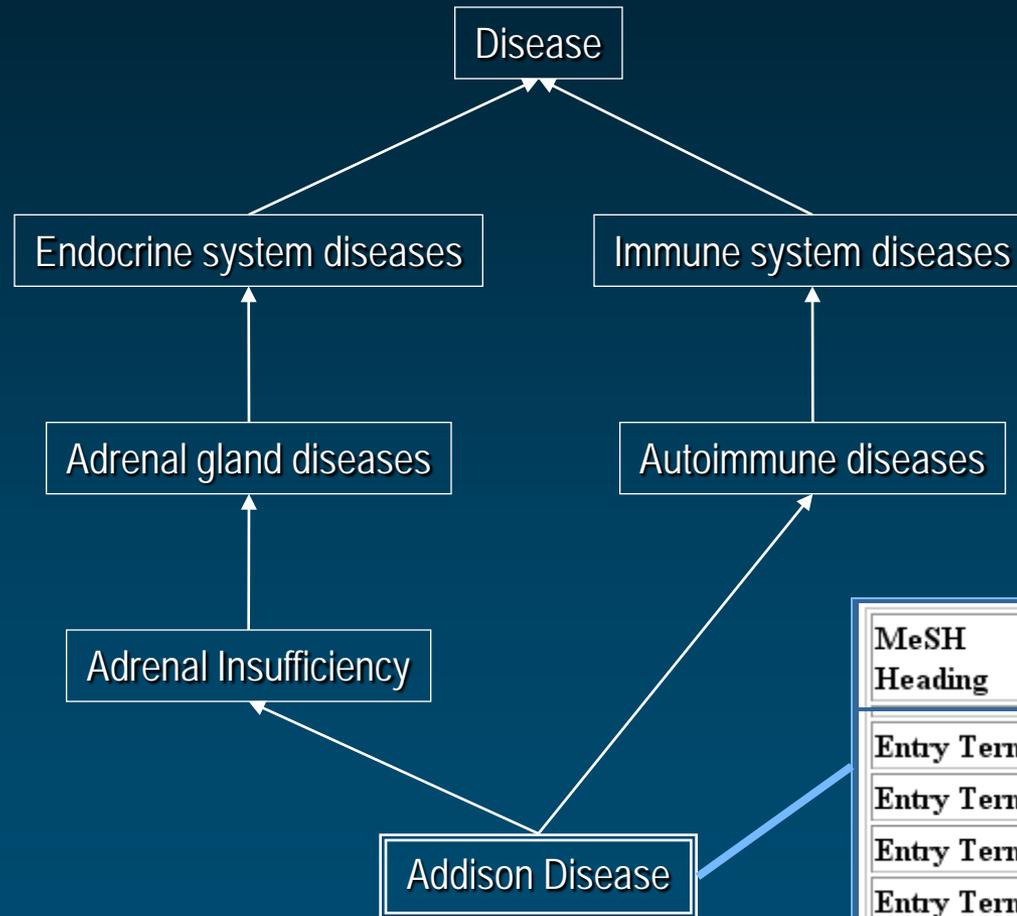
- Collections lexical items + identifiers
- Useful for **entity resolution**
- UMLS Metathesaurus

Types of resources (revisited)

- ◆ Lexical and terminological resources
 - Mostly collections of names for biomedical entities
 - Often have some kind of hierarchical organization (e.g., relations)
- ◆ Ontological resources
 - Mostly collections of relations among biomedical entities
 - Sometimes also collect names

Lexical / Ontological MeSH

<http://www.nlm.nih.gov/mesh/MBrowser.html>



MeSH Heading	Addison Disease
Entry Term	Addison's Disease
Entry Term	Primary Adrenal Insufficiency
Entry Term	Primary Adrenocortical Insufficiency
Entry Term	Primary Hypoadrenalism



Lexical / Ontological FMA

<http://fme.biostr.washington.edu/index.html>



Foundational Model Explorer

Options

Help

Search

Select navigation tree type:

subclass

- Anatomical entity
 - Physical anatomical entity
 - Material anatomical entity
 - Anatomical structure
 - + Body
 - + Cardinal body part
 - + Organ system
 - + Subdivision of cardinal body part
 - + Organ system subdivision
 - Organ
 - + Solid organ
 - Cavitated organ
 - + Organ with organ cavity
 - Organ with cavitated organ parts
 - **Heart**
 - + Cavernous organ
 - + Bone organ
 - Cardinal organ part
 - Organ component
 - + Wall of organ
 - + Organ component layer

PREFERRED NAME: ◆

Heart

NON-ENGLISH EQUIVALENT: ◆

name	language
Cor	Latin
Corazon	Spanish
Coeur	French
Herz	German
Cuore	Italian
Puso	Filipino

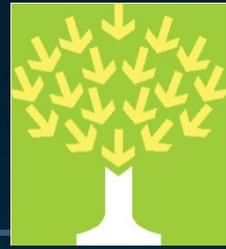
FMAID: ◆

7088

DEFINITION: ◆

Organ with cavitated organ parts, which is continuous with the systemic and pulmonary arterial and venous trees. Examples: There is only one heart.

Unified Medical Language System



◆ SPECIALIST Lexicon

- 475,000 lexical items
- Part of speech and variant information

◆ Metathesaurus

- 8.5M normalized names
- 2.9M concepts
- >10M relations

◆ Semantic Network

- 133 high-level categories
- 7000 relations among them

Lexical
resources

LVG / Norm

Terminological
resources

MetaMap

Ontological
resources

SemRep



Lexical resources

SPECIALIST Lexicon and lexical tools



Lexical Systems Group

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

SPECIALIST Lexicon

- ◆ Content
 - English lexicon
 - Many words from the biomedical domain
- ◆ 475,000 lexical items
- ◆ Word properties
 - morphology
 - orthography
 - syntax
- ◆ Used by the lexical tools



Morphology

◆ Inflection

- noun nucleus, nuclei
- verb cauterize, cauterizes, cauterized, cauterizing
- adjective red, redder, reddest

◆ Derivation

- verb ↔ noun cauterize -- cauterization
- adjective ↔ noun red -- redness



Orthography

◆ Spelling variants

- oe/e oesophagus - esophagus
- ae/e anaemia - anemia
- ise/ize cauterise - cauterize
- genitive mark Addison's disease
Addison disease
Addisons disease

Syntax

◆ Complementation

- verbs

- intransitive I'll treat.
- transitive He treated the patient.
- ditransitive He treated the patient with a drug.

- nouns

- prepositional phrase

Valve of coronary sinus

◆ Position for adjectives



SPECIALIST Lexicon record

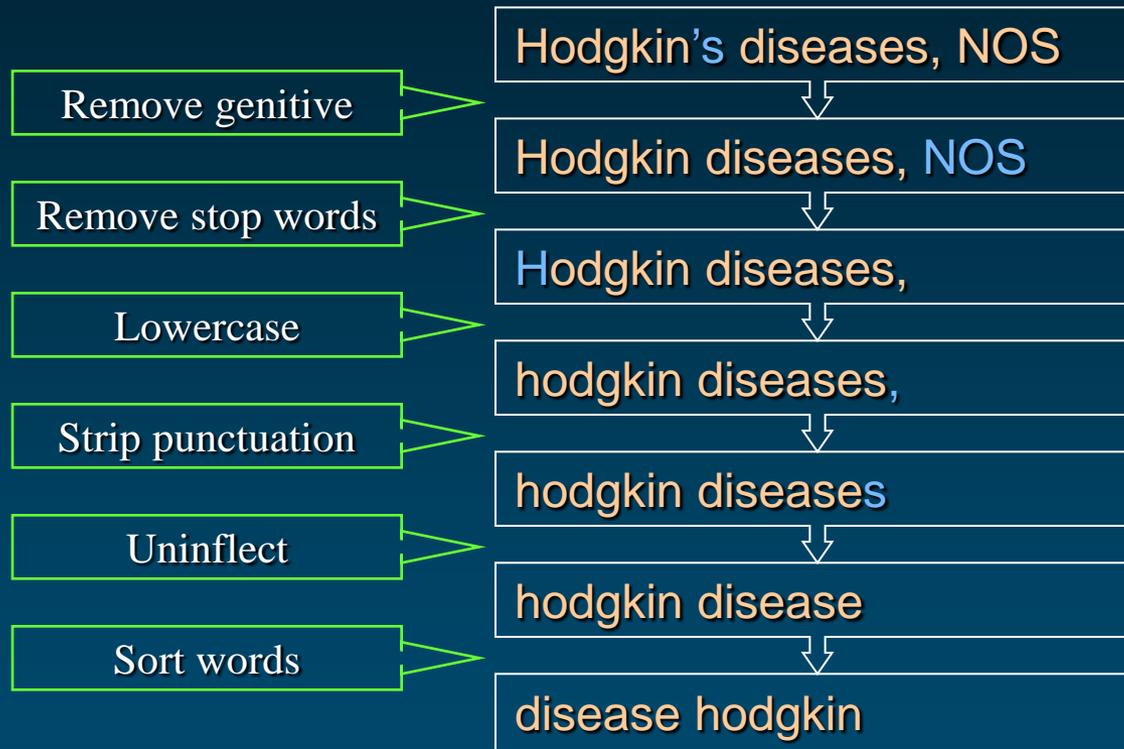
```
{  
  base=hemoglobin      (base form)  
  spelling_variant=haemoglobin  
  entry=E0031208      (identifier)  
  cat=noun            (part of speech)  
  variants=uncount    (no plural)  
  variants=reg        (plural: hemoglobins, hemoglobins)  
}
```



Lexical tools

- ◆ To manage lexical variation in biomedical terminologies
- ◆ Major tools
 - Normalization
 - Indexes
 - Lexical Variant Generation program (lvg)
- ◆ Based on the SPECIALIST Lexicon
- ◆ Used by noun phrase extractors, search engines

Normalization



Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin

normalize

disease hodgkin



Normalization Applications

- ◆ Model for lexical resemblance
- ◆ Help find lexical variants for a term
 - Terms that normalize the same usually share the same LUI
- ◆ Help find candidates to synonymy among terms
- ◆ Help map input terms to UMLS concepts



Indexes

- ◆ Word index
 - word to Metathesaurus strings
 - one word index per language
- ◆ Normalized word index
 - normalized word to Metathesaurus strings
 - English only
- ◆ Normalized string index
 - normalized term to Metathesaurus strings
 - English only



Lexical Variant Generation program

- ◆ Tool for specialists (linguists)
- ◆ Performs atomic lexical transformations
 - generating inflectional variants
 - lowercase
 - ...
- ◆ Performs sequences of atomic transformations
 - a specialized sequence of transformations provides the normalized form of a term (the *norm* program)

Related NLM tools



Lexical Systems Group

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

Public Projects

- ◆ [SPECIALIST Lexicon](#)
- ◆ [LexAccess](#)
- ◆ [Lexical Tools](#)
- ◆ [Text Tools](#)
- ◆ [Text Categorization](#)
- ◆ [GSpell](#)
- ◆ [dTagger](#)

The SPECIALIST Text Tools includes tokenizers that analyze text into word, term, phrase, sentence and section pieces. The tools also include a variant lookup module that retrieves variant ways of expressing the phrases found in the text. The tools are intended to analyze documents into instances of document objects.

The tools are written in Java. These tools include the following:

- ◆ [a word/Sentence/section Tokenizer](#)
- ◆ [a term tokenizer](#)
- ◆ [a phrase tokenizer](#)
- ◆ [a term variant lookup](#)
- ◆ [a part-of-speech tagger \(client\)](#)
- ◆ [a document index maker](#)
- ◆ [a tool to create the textTool indexes](#)

The SPECIALIST spelling resources include two programs GSpell a spelling suggestion tool and BagOWordsPlus a phrase retrieval tool.

GSpell uses several word similarity algorithms to suggest correct spellings for misspelled words. Unlike other spelling suggestion programs GSpell treats space as it would any other letter so that GSpell can correct errors in word compounding. GSpell also be used in word similarity tasks that do not involve misspelling.

BagOWordsPlus uses the word similarity algorithms of GSpell to perform word similarity based phrase level information retrieval.

The dTagger is a Part of Speech (POS) tagger. A POS tagger assigns part of speech tags such as noun, adjective, adverb to sentences. Such tag assignments are a needed component to determining phrase boundaries and head assignment. The dTagger includes the following features: It can tokenize text into single or multi-word terms. It is built specifically for use with the SPECIALIST Lexicon. A default trained model is included, trained on a set of annotated MEDLINE abstracts in the genomics field, (the MedPost corpus). The trainer and updater programs are included to allow the creation of new trained models. Models can be updated with lots of untagged text. Can be trained with just untagged text, if need be. The dTagger is an open source resource and is freely available subject to these [terms and conditions](#).



Lexical resources

Other resources

Similar resources

◆ BioLexicon

- Developed as part of the EU project BOOTStrep
- Focus on biological entities (genes, proteins, chemical entities, organisms) in support of the extraction of gene regulation events

◆ BioThesaurus

- Developed as part of the Protein Information Resource
- Focus on proteins



Need for additional resources

- ◆ More generic
 - WordNet
- ◆ More specific
 - Lexical items specific to specialized subdomains
 - Not listed in biolexicons
 - Not amenable to normalization
 - Examples
 - Genes, proteins
 - MAPK3 / Mapk3 / mapk3
 - Chemicals
 - 5'-3' exonuclease / 3'-5' exonuclease
 - Drugs
 - Acronyms



Gene and protein names

◆ Additional resources

Genew	http://www.gene.ucl.ac.uk/nomenclature/
Entrez Gene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
UniProt	http://www.ebi.uniprot.org/index.shtml

◆ Additional identification methods

- e.g., ABGene (Tanabe & Wilbur, NCBI)
- BioCreAtIvE
 - Gene mention identification
 - Gene normalization



Chemical names

◆ Additional resources

PubChem	http://pubchem.ncbi.nlm.nih.gov/
ChemIDplus	http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp
ChEBI	http://www.ebi.ac.uk/chebi/

Drug names

- ◆ Covered by UMLS
- ◆ Specialized resource: RxNorm
 - Branded names / generic names
 - Various levels of aggregation
 - Ingredient
 - Ingredient + dose
 - Ingredient + form
 - Ingredient + dose + form
 - Codes in various reference systems
- ◆ Mostly US drugs, few “over-the-counter” drugs



Acronyms

◆ Many resources available

- AcroMine

<http://www.nactem.ac.uk/software/acromine/>

- ARGH: Biomedical Acronym Resolver

<http://lethargy.swmed.edu/ARGH/argh.asp>

- Stanford Biomedical Abbreviation Server

<http://bionlp.stanford.edu/abbreviation/>

- AcroMed

<http://medstract.med.tufts.edu/acro1.1/index.htm>

- SaRAD

<http://www.hpl.hp.com/research/idl/projects/abbrev.html>



Terminological resources

UMLS Metathesaurus



<http://www.nlm.nih.gov/research/umls/>



Source Vocabularies

(2013AA)

- ◆ 168 source vocabularies
- ◆ 21 languages
- ◆ Broad coverage of biomedicine
 - 8.5M normalized names
 - 2.9M concepts
 - >10M relations
- ◆ Common presentation



Organize terms

- ◆ 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

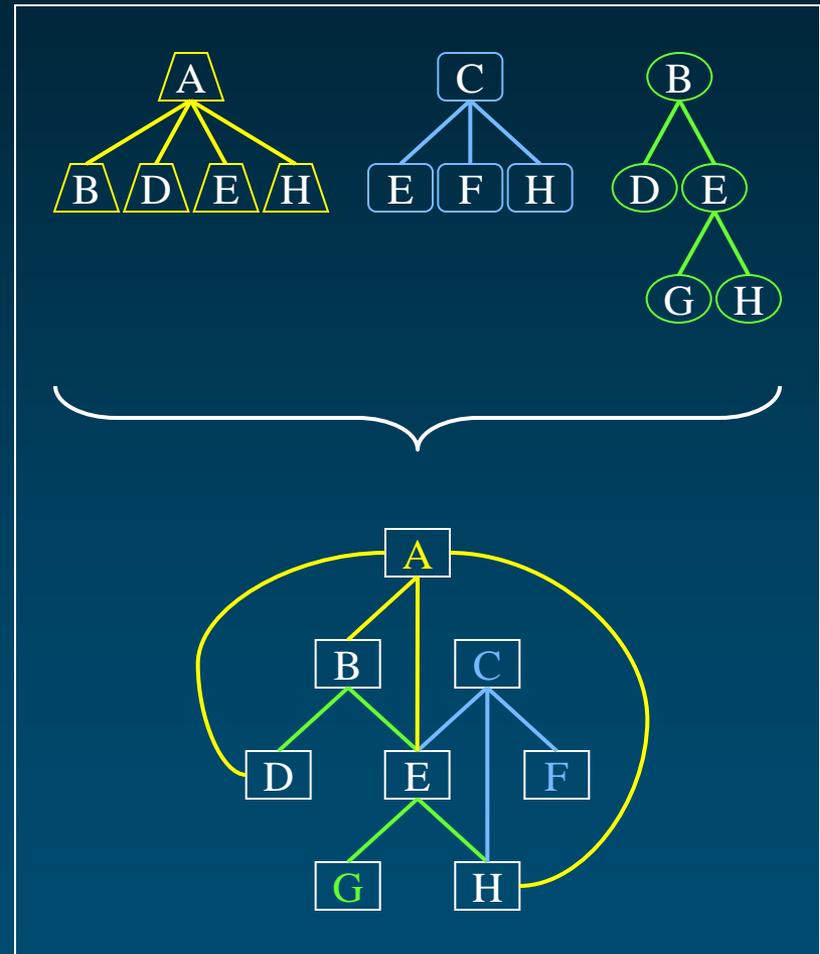
C0001403

Addison's disease

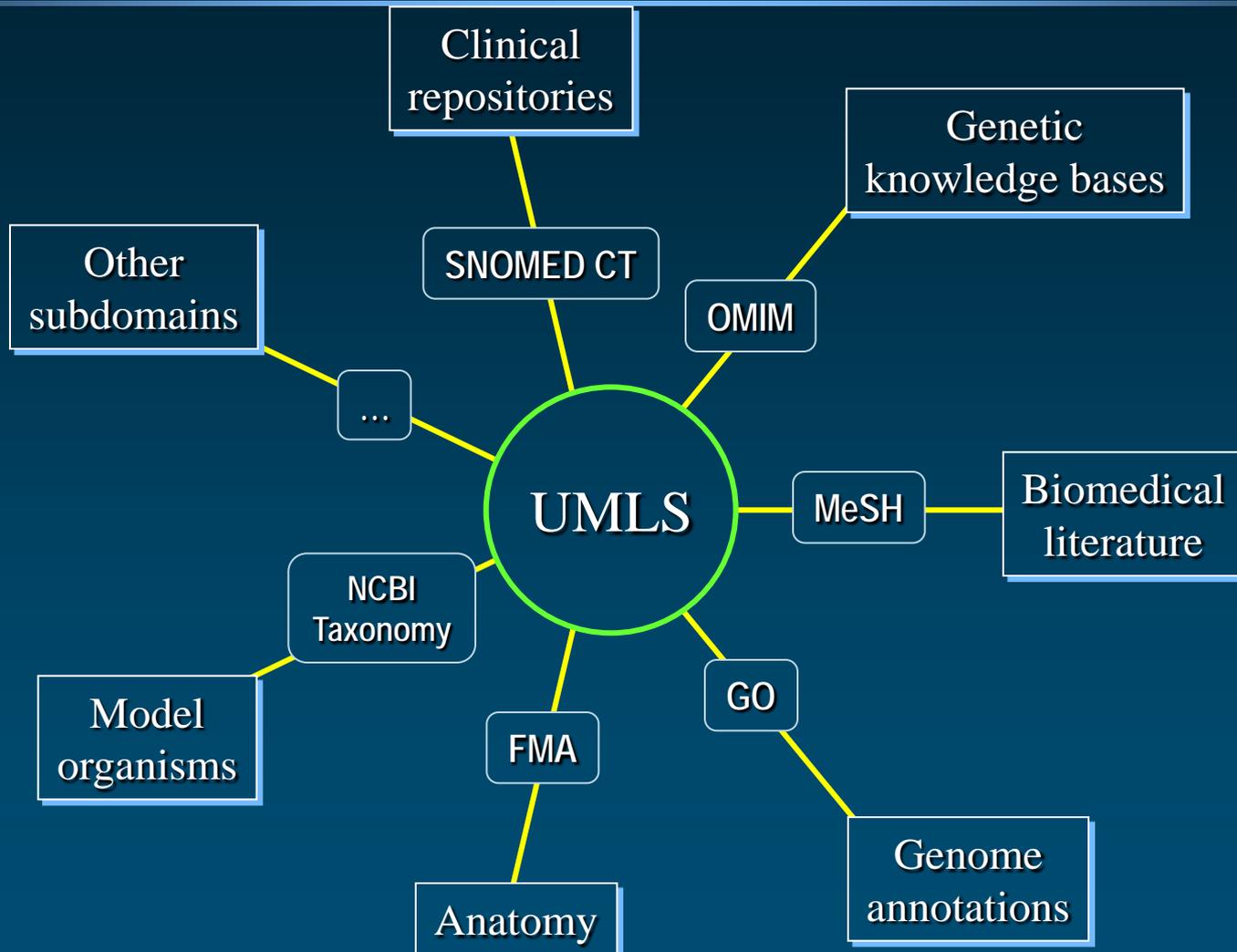


Organize concepts

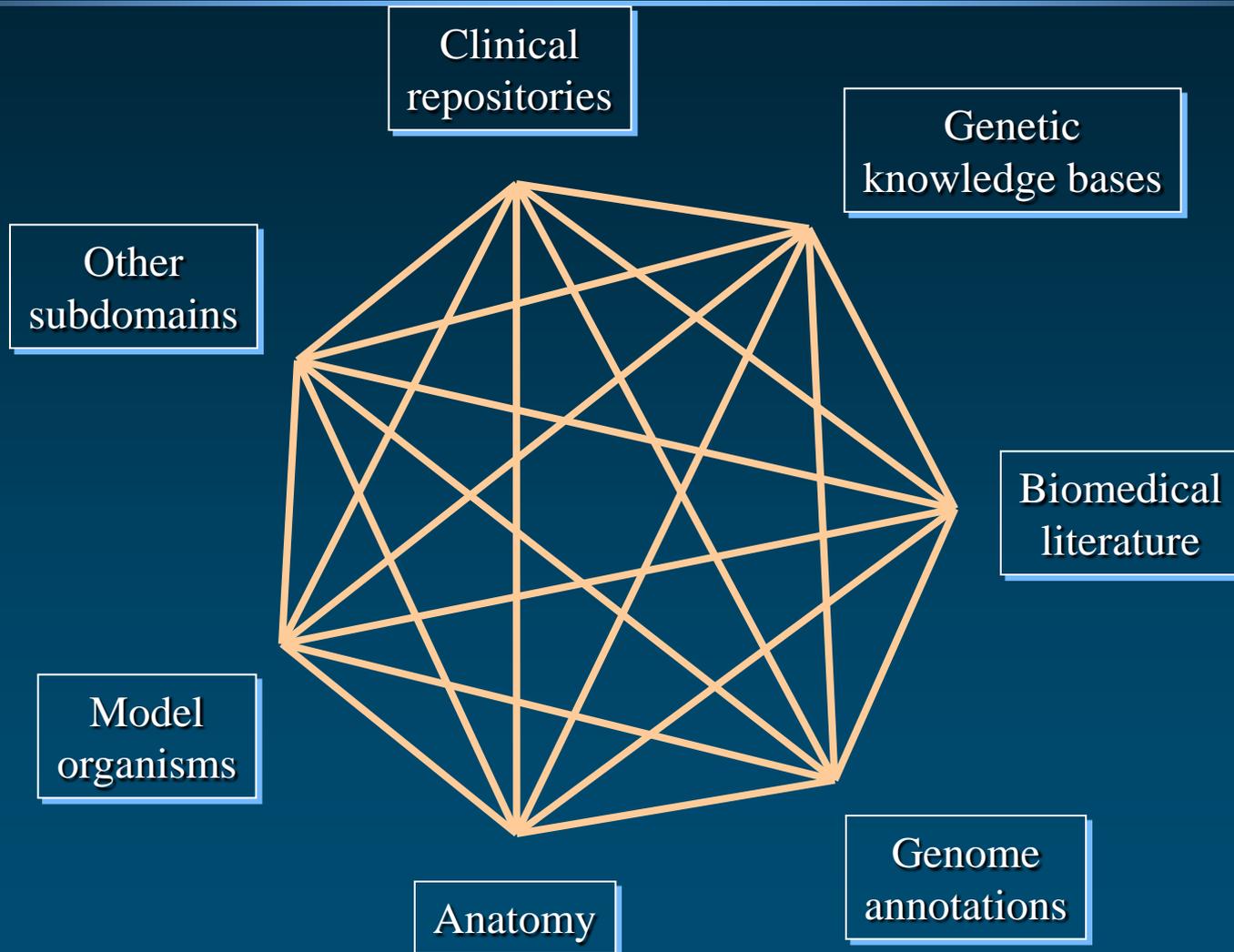
- ◆ Inter-concept relationships: hierarchies from the source vocabularies
- ◆ Redundancy: multiple paths
- ◆ One graph instead of multiple trees (multiple inheritance)



Integrating subdomains



Integrating subdomains



Entity mention vs. resolution

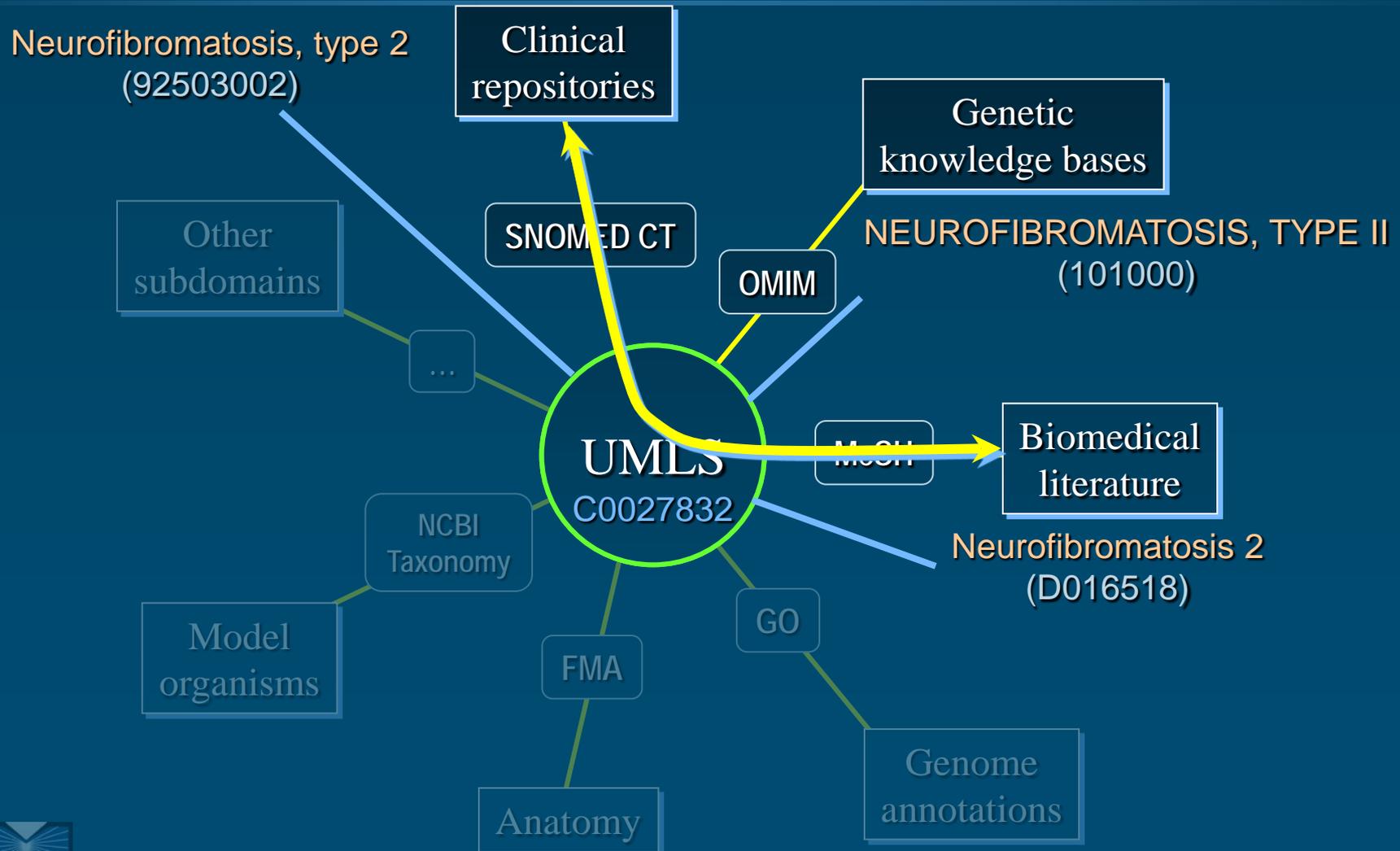
Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis. NF2 is a predominantly intracranial condition whose hallmark is bilateral vestibular schwannomas. NF2 results from a mutation in the gene named merlin, located on chromosome 22.

UMLS:C0027832
MeSH:D016518
SNOMEDCT:92503002
OMIM:101000

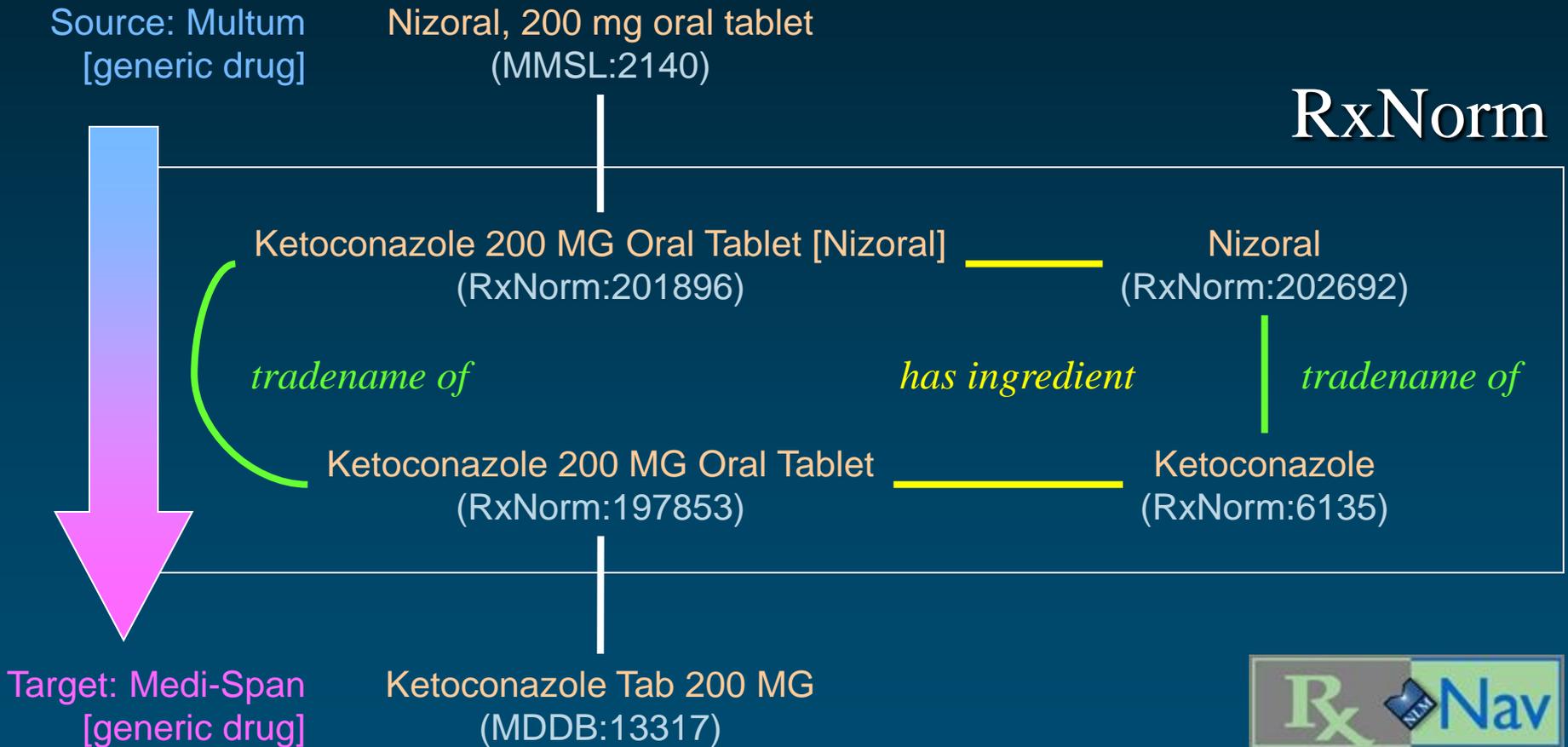
UMLS:C0254123
EG:4771
HGNC:7773
UniProt:P35240



Trans-namespace resolution (1)



Trans-namespace resolution (2)



<http://mor.nlm.nih.gov/download/rxnav/>



Terminological resources

MetaMap

INDEXING INITIATIVE

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

MetaMap

- ◆ UMLS-based entity recognition system
 - Linguistically motivated
 - Exploits both the SPECIALIST lexicon and Metathesaurus
- ◆ In practice, used to identify UMLS concepts in biomedical text
- ◆ Freely available (UMLS license)
- ◆ Two versions
 - Web-based
 - Standalone (MMTx)



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.

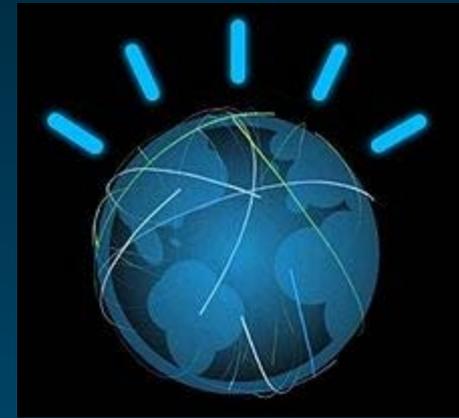
C0254123

Neurofibromin 2	MeSH
Merlin	SNOMED CT
Schwannomin	MeSH
Schwannomerlin	NCI Thesaurus



MetaMap Recent developments

- ◆ Negation processing (NegEx)
- ◆ Word sense disambiguation
- ◆ Integration in NLP pipelines
 - UIMA wrapper
- ◆ Performance improvements
- ◆ User-defined acronyms
- ◆ Applications
 - MetaMap is a key component of the Medical Text Indexer (MEDLINE indexing)
 - From indexing support to “first-line indexing”



Terminological resources

Other NER systems

BioPortal Annotator

Annotator

Annotate text with terms from ontologies [?](#)

insert sample text

Neurofibromatosis type 2 (NF2) is often not recognised as a distinct entity from peripheral neurofibromatosis.

Select Ontologies

Type here to select ontologies

Select UMLS Semantic Types

Type here to select UMLS semantic types

Include Mappings: Automatic

Include Ancestors Up To Level

Get Annotations

Annotations

Results are filtered by: Ontology

Results filtered to SNOMED CT and MeSH

total results 134 (direct 134 / ancestor 0 / mo...

TERM	filter	ONTOLOGY	filter	TYPE	filter	CONTEXT	MATCHED TERM	filter	MATCHED ONTOLOGY	filter
Neurofibromatosis		SNOMED Clinical Terms		direct		distinct entity from peripheral neurofibromatosis	Neurofibromatosis		SNOMED Clinical Terms	
Neurofibromatosis		SNOMED Clinical Terms		direct		distinct entity from peripheral neurofibromatosis	Neurofibromatosis		SNOMED Clinical Terms	
Neurofibromatosis		SNOMED Clinical Terms		direct		Neurofibromatosis type 2 (NF2) is	Neurofibromatosis		SNOMED Clinical Terms	
Neurofibromatosis		SNOMED Clinical Terms		direct		Neurofibromatosis type 2 (NF2) is	Neurofibromatosis		SNOMED Clinical Terms	
Neurofibromatoses		Medical Subject Headings		direct		Neurofibromatosis type 2 (NF2) is	Neurofibromatoses		Medical Subject Headings	
Neurofibromatoses		Medical Subject Headings		direct		distinct entity from peripheral neurofibromatosis	Neurofibromatoses		Medical Subject Headings	
Type		SNOMED Clinical Terms		direct		Neurofibromatosis type 2 (NF2) is often	Type		SNOMED Clinical Terms	
Peripheral		SNOMED Clinical Terms		direct		a distinct entity from peripheral neurofibromatosis.	Peripheral		SNOMED Clinical Terms	
Type		SNOMED Clinical Terms		direct		Neurofibromatosis type 2 (NF2) is often	Type		SNOMED Clinical Terms	
Type 2		SNOMED Clinical Terms		direct		Neurofibromatosis type 2 (NF2) is often not	Type 2		SNOMED Clinical Terms	
Neurofibromatosis, type 2		SNOMED Clinical Terms		direct		Neurofibromatosis type 2 (NF2) is often not	Neurofibromatosis, type 2		SNOMED Clinical Terms	
Type - attribute		SNOMED Clinical Terms		direct		Neurofibromatosis type 2 (NF2) is often	Type - attribute		SNOMED Clinical Terms	
Neurofibromatosis 2		Medical Subject Headings		direct		Neurofibromatosis type 2 (NF2) is often not	Neurofibromatosis 2		Medical Subject Headings	
Neurofibromatosis 1		Medical Subject Headings		direct		a distinct entity from peripheral neurofibromatosis	Neurofibromatosis 1		Medical Subject Headings	



TerMine (C-value) analysis

[Service
questionnaire](#)

Found 5 terms in 2.2 seconds - all terms ([in table](#)) ([in text](#)) - threshold:

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.

Thank you for using TerMine. Please now complete a [questionnaire](#) to let us know your views about this service.

Whatizit

The Whatizit logo, with the word "Whatizit" in a stylized font where the 'i' is red and the 'z' is blue.

3

Select a pipeline:

whatizitEBIMedDiseaseChemicals

Resulting tagged text

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 .

Ontological resources

Ontological resources

- ◆ Provide background knowledge
 - For resolving ambiguity in entity recognition
 - Merlin: Protein or Bird?
 - For relation extraction
 - Template relations between high-level concepts
 - Used in combination with clues from linguistic phenomena in text

Ontological resources

- ◆ Various level of formality
 - Formal top-level ontologies (e.g., BioTop)
 - Informal top-level ontologies (e.g., UMLS Semantic Network)
 - Domain-Range constraints for roles in DL-based terminologies (e.g., SNOMED CT, NCI Thesaurus)
 - Relations in terminologies
- ◆ Various level of granularity
 - UMLS Semantic Network: 133 types
 - Foundational Model of Anatomy: 70,000 classes

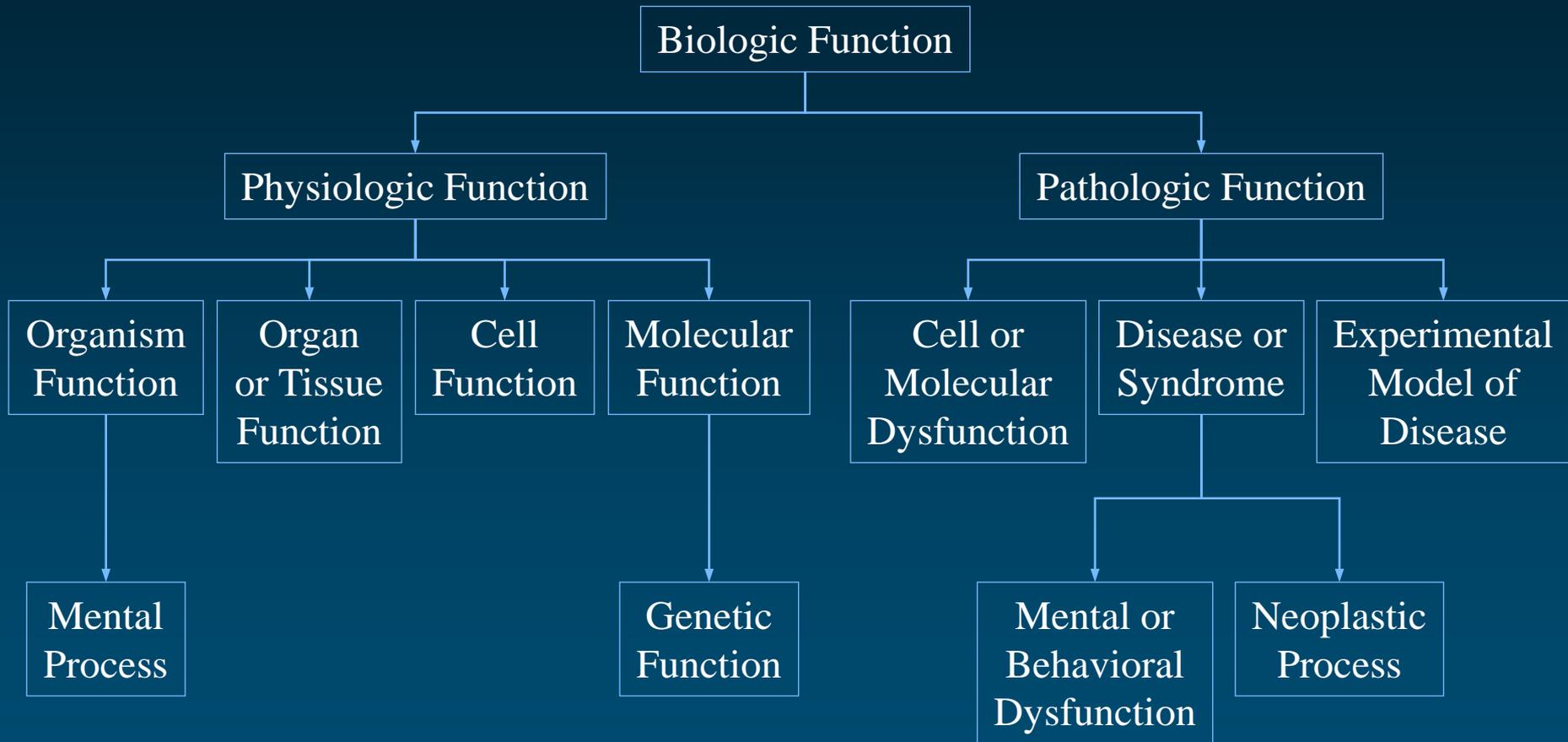


Ontological resources

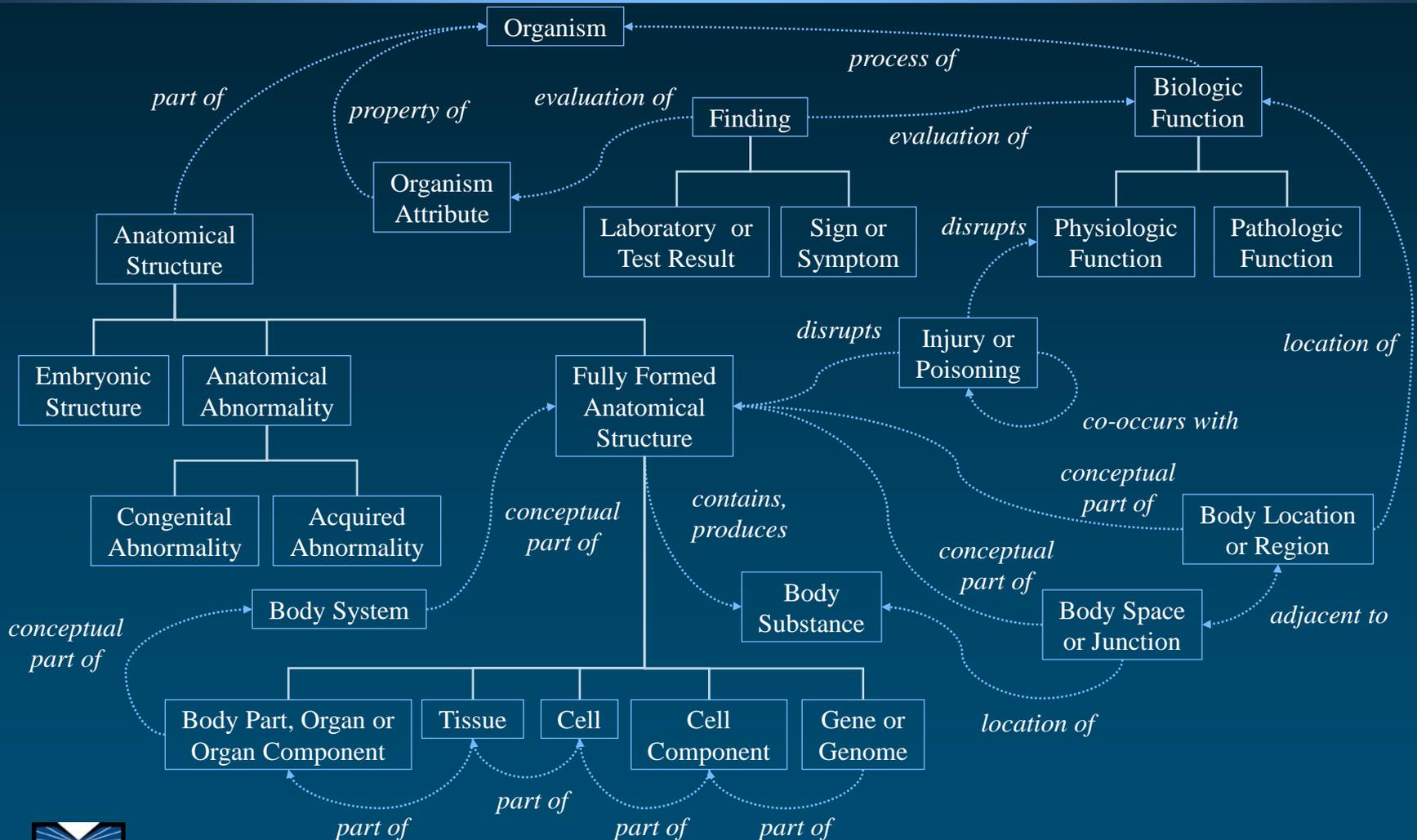
UMLS Semantic Network

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

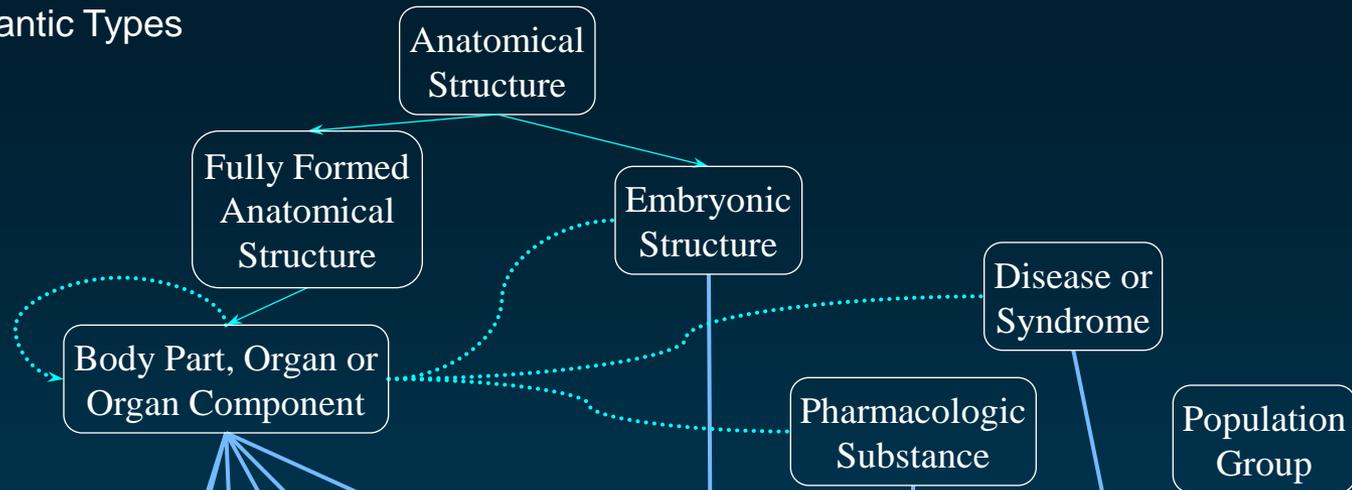
“Biologic Function” hierarchy (isa)



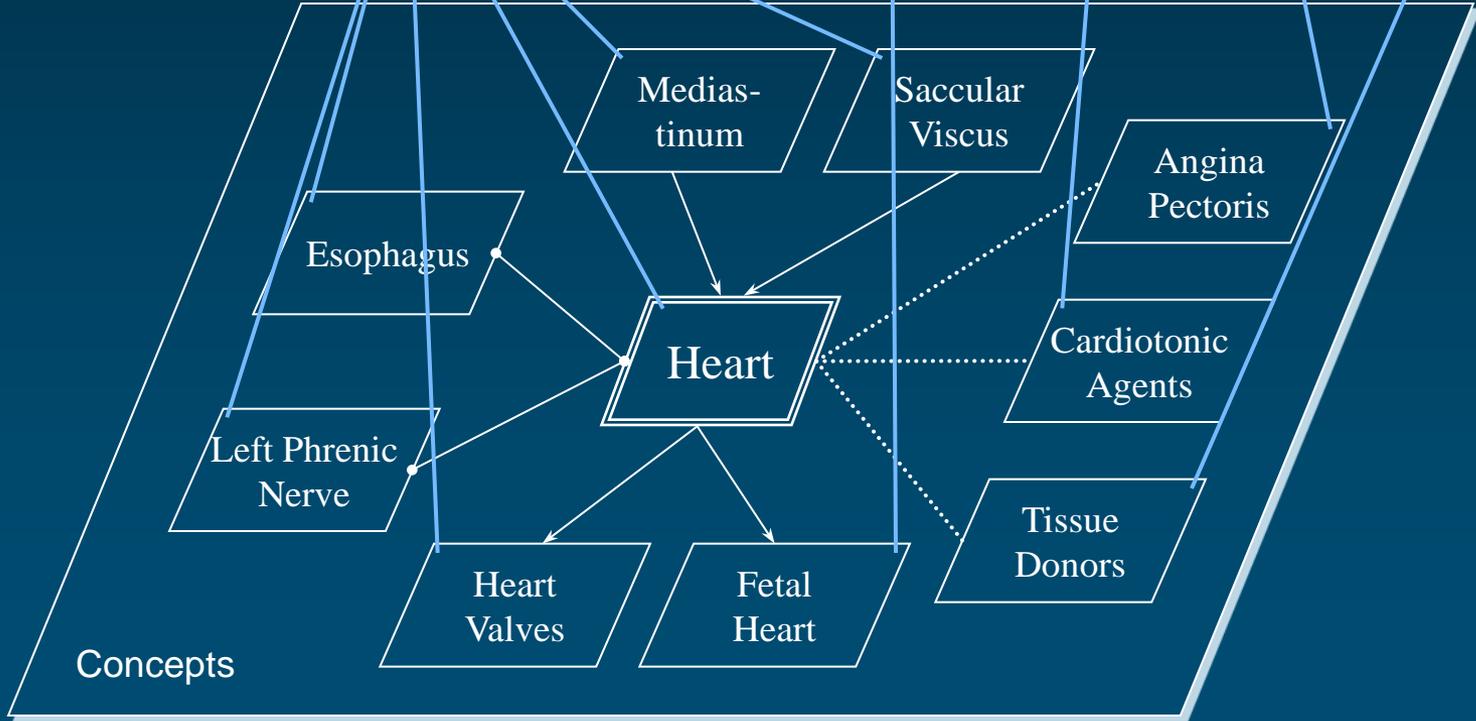
Associative (non-isa) relationships



Semantic Types



Semantic Network



Metathesaurus

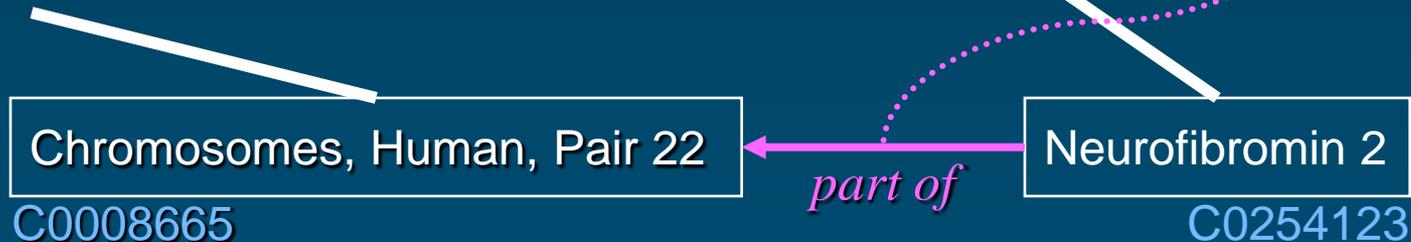
Concepts

Ontological resources

SemRep

SemRep Relation extraction

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.



SemRep Recent developments

- ◆ All of MEDLINE processed
 - 60M predications
- ◆ Use of graph theory principles for selecting salient associations
- ◆ Applications
 - Literature-based discovery
 - Relationship between testosterone, cortisol and sleep quality in aging men
 - Automatic summarization
 - Semantic MEDLINE <http://skr3.nlm.nih.gov/SemMedDemo/>



Ontological resources

Other resources

Other ontological resources

◆ Ontology integration systems

- NCBO BioPortal

<http://www.bioontology.org/BioPortal>

◆ Ontologies

- Top-level ontologies (e.g., BioTop)
- Domain ontologies (e.g., FMA, SNOMED CT, NCI Thesaurus)

Other relation extraction systems

- ◆ Many relation extraction systems available
 - Specialized
 - Protein-protein interaction (e.g., Info-PubMed, TextPresso, ...)
 - BioCreAtIvE (task 2)
 - More generic (e.g., MedLEE / BioMedLEE)
 - Commercial systems (TeSSI, Linguamatics, ...)

Future directions

- ◆ Evolution of existing resources
 - Modularization
 - Compatibility with NLP pipelines (UIMA, GATES)
 - Services (web services, CTS2)
- ◆ Clinical text
 - Limited availability to non-clinical institutions
 - Opportunity to develop new tools (e.g., de-identification)
 - Integration with knowledge bases
 - Definitional vs. assertional knowledge
- ◆ Collaborative development



ORBIT project

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

The Online Registry of Biomedical Informatics Tools (ORBIT) Project is a community-wide effort to create and maintain a structured, searchable metadata registry for informatics software, knowledge bases, data sets and design resources. For tips on how to use ORBIT visit our [help page](#).

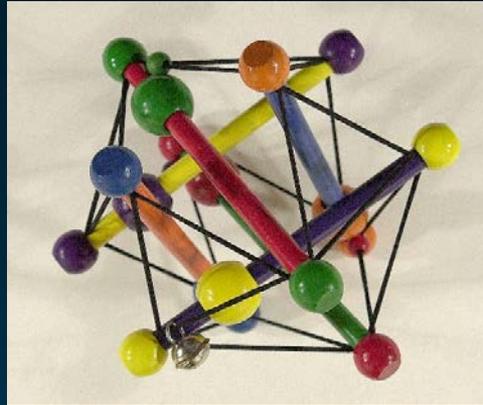
Recently added

Title	Posted
Corpus for Evidence Based Medicine Query-based Summarisation	7 weeks 5 days ago
ABNER	9 weeks 1 hour ago
Wikimeta	9 weeks 10 hours ago
MTI ML (Machine Learning Package)	9 weeks 13 hours ago
AMIA Clinical NLP Tutorial, Parts 1 & 2	12 weeks 5 days ago
NEMO	12 weeks 6 days ago
Structured Abstracts in MEDLINE®	13 weeks 1 day ago
Integrated cTAKES	14 weeks 3 days ago
Clinical NLP Evaluation Workbench	15 weeks 2 days ago
iDASH Virtual Machine of Clinical NLP Tools	18 weeks 6 days ago



References

- ◆ Bodenreider O.
Lexical, terminological and ontological resources for biological text mining.
In: Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine: Artech House; 2006. p. 43-66.



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