



Wellcome Trust Genome Campus, Cambridge, Hinxton, UK  
February 20, 2006

Mini-Symposium  
“Semantic Enrichment of Scientific Literature”

## NLM Resources for Semantic Enrichment



*Olivier Bodenreider*

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

# Q. What is required for semantic enrichment?

## ◆ Tools

- Entity recognition
- Relation extraction

## ◆ Enabling resources

- Lexical
- Terminological
- Ontological



# NLM tools and resources

## ◆ Tools

- Entity recognition
- Relation extraction

MetaMap (MMTx)

SemRep/SemGen

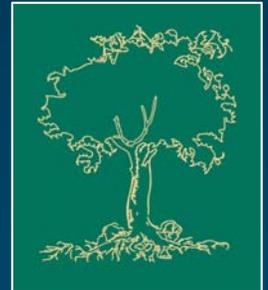
## ◆ Enabling resources

- Lexical
- Terminological
- Ontological

SPECIALIST Lexicon

Metathesaurus

Semantic Network



Unified Medical Language System



# Unified Medical Language System



## ◆ SPECIALIST Lexicon

- 200,000 lexical items
- Part of speech and variant information

## ◆ Metathesaurus

- 5M names from over 100 terminologies
- 1M concepts
- 16M relations

## ◆ Semantic Network

- 135 high-level categories
- 7000 relations among them

Lexical  
resources

Terminological  
resources

Ontological  
resources



# UMLS 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



# Entity recognition MetaMap (MMTx)

- ◆ Input: Biomedical text of variable length (e.g., MEDLINE abstract)
- ◆ Output: Ranked list of Metathesaurus concepts associated with each piece of text
- ◆ Features
  - Based on UMLS (lexicon and Metathesaurus)
  - Complemented with other sources (genes, acronyms)
  - Term variation (synonymy, inflection, derivation)
  - Approximate matching

Developed by Lan Aronson  
<http://mmtx.nlm.nih.gov/>



# 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 output

- ◆ [NF2] [is] [a predominantly **intracranial** condition] [whose hallmark] [is] [**bilateral** vestibular schwannomas].
  - NF2
    - NF2 (NF2 gene) [Gene or Genome]
    - NF2 (Neurofibromin 2) [ ... Protein, ...]
  - a predominantly intracranial condition
    - Intracranial [Body Location or Region]
    - Condition [Qualitative Concept]
  - bilateral vestibular schwannomas
    - Schwannomas, Vestibular (Neuroma, Acoustic) [Neoplastic P.]
    - Bilateral [Spatial Concept]



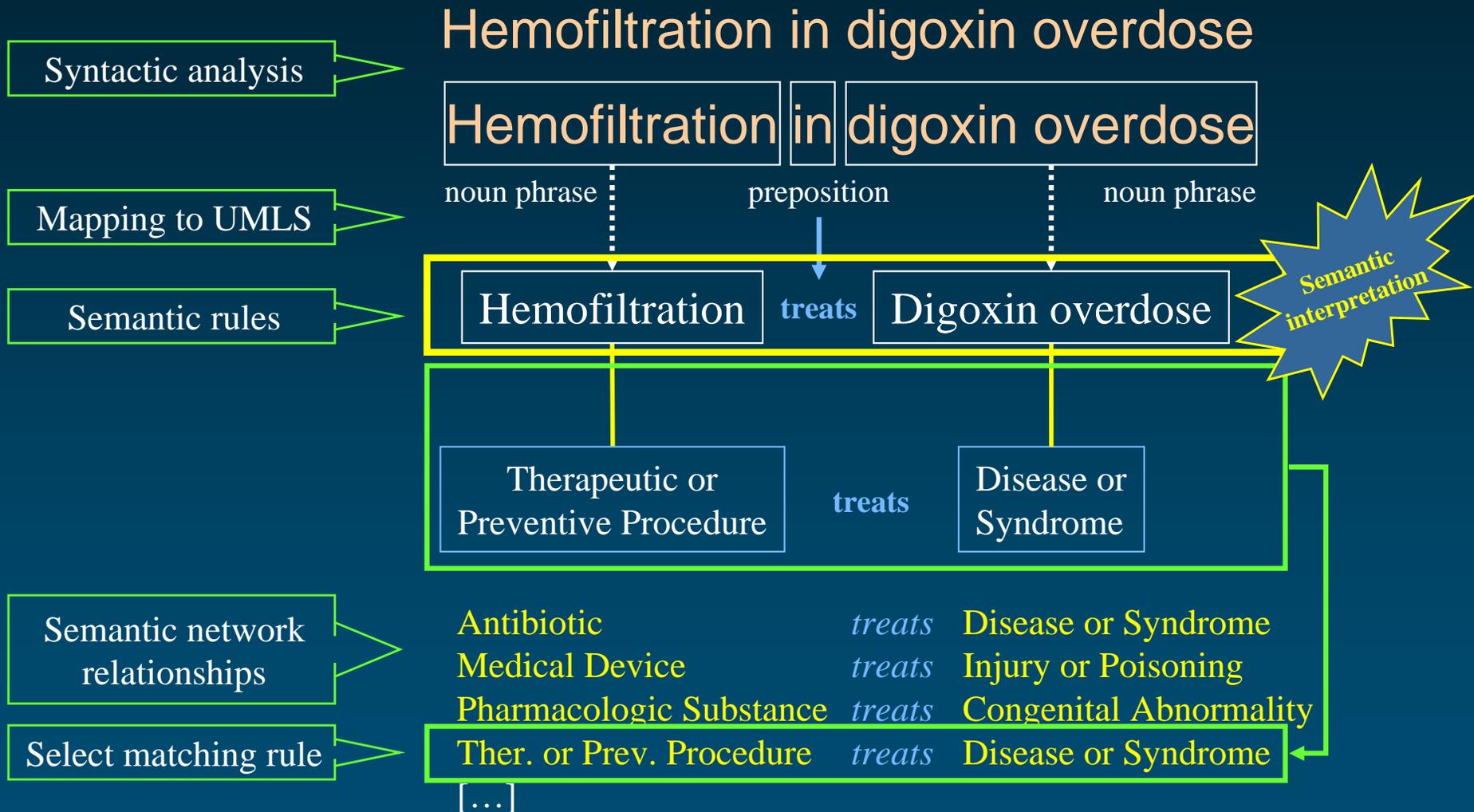
# Relation extraction SemRep/SemGen

- ◆ Correspondence between
  - Linguistic phenomena
  - Semantic relations (predications)
- ◆ Input: Biomedical text of variable length
- ◆ Output: List of semantic predications
- ◆ Features
  - Based on MetaMap
  - Semantic constraints provided by ontologies (UMLS Semantic Network)

Developed by Tom Rindflesch  
<http://skr.nlm.nih.gov/>



# Semantic interpretation with SemRep



# Built-in resources

<http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi>

The screenshot displays the NCBI Entrez search engine interface. At the top left is the NCBI logo. The main header features the Entrez logo and the text "Entrez, The Life Sciences Search Engine". Below this is a navigation bar with links for HOME, SEARCH, SITE MAP, PubMed, All Databases, Human Genome, GenBank, Map Viewer, and BLAST. A search bar contains the text "NF2" with "GO" and "CLEAR" buttons and a "Help" link. The search results are presented in a grid of 20 items, each with a count, an icon, a title, a description, and a help icon.

692		<b>PubMed:</b> biomedical literature citations and abstracts	?	73		<b>Books:</b> online books	?
166		<b>PubMed Central:</b> free, full text journal articles	?	27		<b>OMIM:</b> online Mendelian Inheritance in Man	?
1		<b>Site Search:</b> NCBI web and FTP sites	?	none		<b>OMIA:</b> Online Mendelian Inheritance in Animals	?
278		<b>Nucleotide:</b> sequence database (GenBank)	?	17		<b>UniGene:</b> gene-oriented clusters of transcript sequences	?
160		<b>Protein:</b> sequence database	?	none		<b>CDD:</b> conserved protein domain database	?
1		<b>Genome:</b> whole genome sequences	?	8		<b>3D Domains:</b> domains from Entrez Structure	?
1		<b>Structure:</b> three-dimensional macromolecular structures	?	45		<b>UniSTS:</b> markers and mapping data	?
none		<b>Taxonomy:</b> organisms in GenBank	?	5		<b>PopSet:</b> population study data sets	?
790		<b>SNP:</b> single nucleotide polymorphism	?	1680		<b>GEO Profiles:</b> expression and molecular abundance profiles	?
35		<b>Gene:</b> gene-centered information	?	1		<b>GEO DataSets:</b> experimental sets of GEO data	?
19		<b>HomoloGene:</b> eukaryotic homology groups	?	none		<b>Cancer Chromosomes:</b> cytogenetic databases	?

# Q. Benefit of semantic annotation?

- ◆ More accurate information retrieval
  - Based on concepts, not words
  - Direct link between text and concepts (unlike MeSH indexing)
- ◆ Enables cross-references across heterogeneous resources, including the literature (e.g., NCBI resources)
- ◆ Supports relation extraction



# Q. Benefit of using ontologies?

## ◆ Information extraction

- Source of **domain knowledge**
  - Known relations among entities
  - Constraints for relations among entities

## ◆ Entity recognition

- Often source of **terminology**
  - All names for a given concept
- Support entity normalization (controlled vocabularies)
- Support word sense disambiguation



# Q. Is collaborative annotation possible?

## ◆ Why not?

- See GeneRIF

## ◆ Who should do it?

- Authors?
  - Probably not if it requires knowledge of a controlled vocabulary *and indexing rules*
- Editors / Publishers?
- Text miners?

The steps to submit a GeneRIF include:

1. Query [PubMed](#) to identify the PubMed ID for the reference.
2. Query [Entrez Gene](#) to find the record for the gene.
3. Display the desired gene's Entrez Gene report page by selecting that gene from the query results list.
4. In the grey Bibliography bar, select **Submit GeneRIF**.
5. Enter the PubMed ID in the correct box.
6. Add your description of the function of the gene or its protein product.
7. Enter your e-mail address.
8. Press the **Validate** button to validate your entry. The PubMed ID value you entered will be verified and the author, title, and citation will be displayed for review.
9. When all is correct, press **Submit** to store the GeneRIF.

Each submission is acknowledged using the e-mail address that was entered. The text of the GeneRIF is reviewed for inappropriate content and typographical errors, but is not otherwise edited.

<http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html>



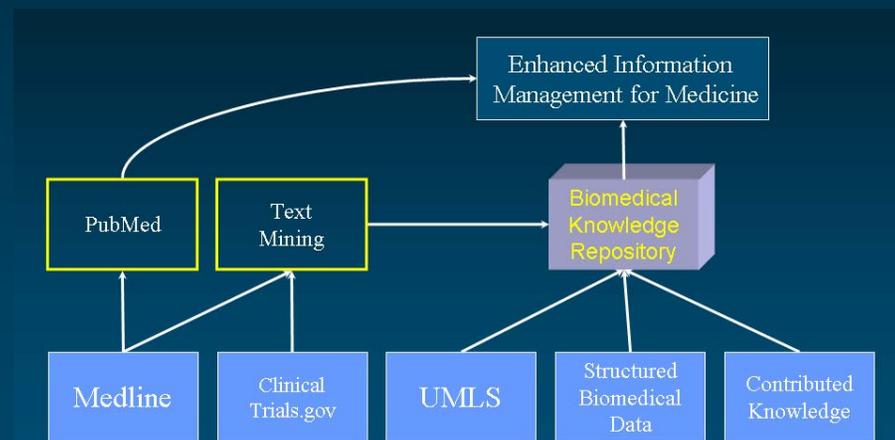
# Q. Towards a fact repository?

## ◆ Integrating knowledge

- Phenotype and genotype together
- Unique repository
- Seamless environment

## ◆ Enabling resource

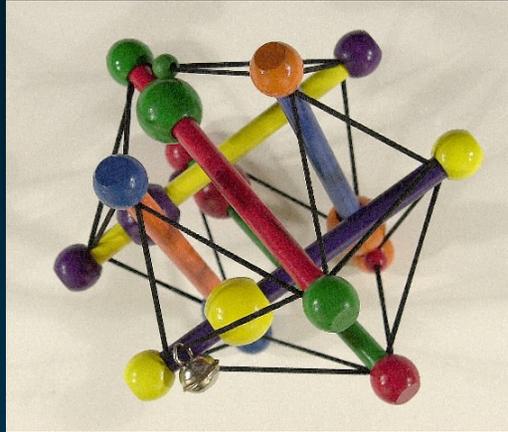
- Summarization
- Question answering
- Knowledge discovery
- Refined information retrieval



# Conclusions

- ◆ Semantic annotation of the biomedical literature will play an important role in biomedical research in the near future
- ◆ Semantic annotation lays the foundations for
  - Advanced library services
  - Information/knowledge management
- ◆ Synergistic with the Semantic Web
  - W3C Health Care and Life Sciences Interest Group





# 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