



Integrated R&D Informatics  
& Knowledge Management

February 25, 2011

# Knowledge Representation & Ontology in the Biomedical Domain



*Olivier Bodenreider*

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

# What are ontologies for?

[Bodenreider, YBMI 2008]

- ◆ 3 major categories of use
  - **Knowledge management** (indexing and retrieval of data and information, access to information, mapping among ontologies)
  - **Data integration**, exchange and semantic interoperability
  - **Decision support and reasoning** (data selection and aggregation, decision support, natural language processing applications, knowledge discovery).
- ◆ Barriers to usability of biomedical ontologies



# 1. Knowledge management

# Knowledge management

*Annotating data and resources*

# Terminology in ontology

- ◆ Ontology as a source of vocabulary
  - List of names for the entities in the ontology (ontology vs. terminology)
- ◆ Most ontologies have some sort of terminological component
  - Exceptions: GALEN, LOINC
- ◆ Not all surface forms represented
  - Often insufficient for NLP applications
  - Large variation in number of terms per concept across ontologies

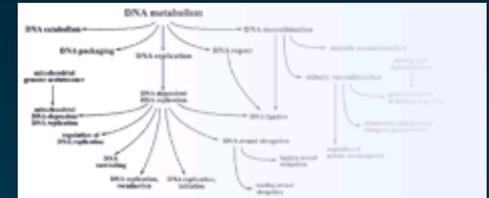


# Annotating gene products

## ◆ Gene Ontology

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

- 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/>



# Indexing the biomedical literature

## ◆ Medical Subject Headings (MeSH)

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

- Used for indexing and retrieval of the biomedical literature (MEDLINE)



## ◆ Indexing

- Performed manually by human indexers
  - With help of semi-automatic systems (suggestions)  
e.g., Indexing Initiative at NLM
- Automatic indexing systems

# MeSH MEDLINE indexing

□ I: [Anesth Analg](#). 2008 Jun;106(6):1813-9.

[Related Articles,](#)  
[Links](#)



## Free cortisol in sepsis and septic shock.

[Bendel S](#), [Karlsson S](#), [Pettilä V](#), [Loisa P](#), [Varpula M](#), [Ruukonen E](#); [Finnsepsis Study Group](#).

► [Collaborators \(26\)](#)

Department of Intensive Care, Kuopio University Hospital, PL 16222 Kuopio, Finland. [Stepani.Bendel@kuh.fi](mailto:Stepani.Bendel@kuh.fi)

**BACKGROUND:** Severe sepsis activates the hypothalamopituitary axis, increasing cortisol production. In some studies, hydrocortisone substitution based on an adrenocorticotropic hormone-stimulation test or baseline cortisol measurement has improved outcome. Because only the free fraction of cortisol is active, measurement of free cortisol may be more important than total cortisol in critically ill patients. We measured total and free cortisol in patients with severe sepsis and related the concentrations to outcome. **METHODS:** In a prospective study, severe sepsis was defined according the American College of Chest Physicians/Society of Critical Care Medicine criteria. Blood samples were drawn within 24 h of study entry. Serum cortisol was analyzed by electrochemiluminescence immunoassay. The Coolens method was used for calculating serum free cortisol concentrations. **RESULTS:** Blood samples were collected from 125 patients, of whom 62 had severe sepsis and 63 septic shock. Hospital mortality was 21%. Calculated free serum cortisol correlated well with serum total cortisol ( $r = 0.90$ ,  $P < 0.001$ ). There was no difference in the total cortisol concentrations in patients with sepsis and septic shock ( $728 \pm 386$  nmol/L vs  $793 \pm 439$  nmol/L,  $P = 0.44$ ). Nonsurvivors had higher calculated serum free ( $209 \pm 151$  nmol/L) and total ( $980 \pm 458$  nmol/L) cortisol concentrations than survivors ( $119 \pm 111$  nmol/L,  $P = 0.002$ , and  $704 \pm 383$  nmol/L,  $P = 0.002$ ). Depending on the definition, the incidence of adrenal insufficiency varied from 8% to 54%. **CONCLUSIONS:** Clinically, calculation of free cortisol does not provide essential information for identification of patients who would benefit from corticoid treatment in severe sepsis and septic shock.

# MeSH MEDLINE indexing

## MeSH Terms:

- ◆ [Adrenal Cortex Function Tests](#)
- ◆ [Adrenal Insufficiency/blood\\*](#)
- ◆ [Adrenal Insufficiency/drug therapy](#)
- ◆ [Adrenal Insufficiency/mortality](#)
- ◆ [Adult](#)
- ◆ [Biological Markers/blood](#)
- ◆ [Female](#)
- ◆ [Finland/epidemiology](#)
- ◆ [Hospital Mortality](#)
- ◆ [Humans](#)
- ◆ [Hydrocortisone/blood\\*](#)
- ◆ [Hydrocortisone/therapeutic use](#)
- ◆ [Kaplan-Meiers Estimate](#)

- ◆ [Male](#)
- ◆ [Predictive Value of Tests](#)
- ◆ [Prospective Studies](#)
- ◆ [Sepsis/blood\\*](#)
- ◆ [Sepsis/drug therapy](#)
- ◆ [Sepsis/mortality](#)
- ◆ [Severity of Illness Index](#)
- ◆ [Shock, Septic/blood\\*](#)
- ◆ [Shock, Septic/drug therapy](#)
- ◆ [Shock, Septic/mortality](#)
- ◆ [Treatment Outcome](#)

## Substances:

- ◆ [Biological Markers](#)
- ◆ [Hydrocortisone](#)



# Coding clinical records

## ◆ SNOMED CT

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

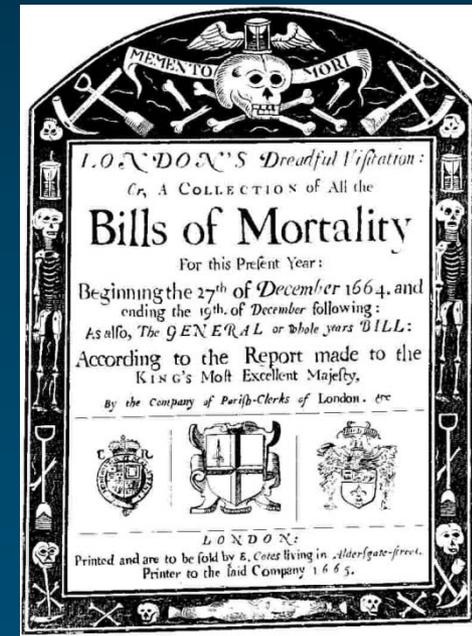
- Clinical documentation
- Problem list

## ◆ International Classification of Diseases

ICD family (ICD 10, ICD 9-CM)

<http://www.who.int/classifications/icd/en/>

- Billing (US)
- Epidemiology (worldwide)
  - Morbidity
  - Mortality statistics

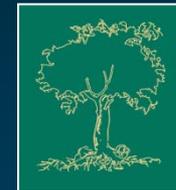


# Knowledge management

*Accessing biomedical information*

# Resources for biomedical search engines

- ◆ Synonyms
- ◆ Hierarchical relations
- ◆ High-level categorization
- ◆ Co-occurrence information
- ◆ Translation



# MeSH “synonyms” MEDLINE retrieval

- ◆ MeSH entry terms
  - Used as equivalent terms for retrieval purposes
  - Not always synonymous
- ◆ Increase recall without hurting precision

<b>MeSH Heading</b>	Addison Disease
<b>Entry Term</b>	Addison's Disease
<b>Entry Term</b>	Primary Adrenal Insufficiency
<b>Entry Term</b>	Primary Adrenocortical Insufficiency

# MeSH “synonyms” MEDLINE retrieval

The screenshot shows the PubMed search interface. At the top, the NCBI logo is on the left, and the PubMed logo with the URL [www.pubmed.gov](http://www.pubmed.gov) is in the center. To the right, it states "A service of the U.S. National Library of Medicine and the National Institutes of Health". Below the logo is a navigation bar with tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", and "OMIM". The search bar contains the text "Search PubMed" and a dropdown menu set to "PubMed". The search term "Primary Hypoadrenalism" is entered in the search box. To the right of the search box are "Go" and "Clear" buttons. Below the search bar is a row of buttons: "Limits", "Preview/Index", "History", "Clipboard", and "Details". The "Query Translation" section is highlighted in blue and contains the following text: 

```
"addison disease"[MeSH Terms] OR ("addison"[All Fields] AND "disease"[All Fields]) OR "addison disease"[All Fields] OR ("primary"[All Fields] AND "hypoadrenalism"[All Fields]) OR "primary hypoadrenalism"[All Fields]
```

 At the bottom of the query translation box are "Search" and "URL" buttons. On the left side of the page, there is a vertical menu with links for "About Entrez", "Text Version", "Entrez PubMed", "Overview", "Help | FAQ", "Tutorials", "New/Noteworthy", "E-Utilities", "PubMed Services", "Journals Database", "MeSH Database", "Single Citation", and "Matcher".



# MeSH hierarchies MEDLINE retrieval

- ◆ MeSH “explosion”
  - Search for a given MeSH term *and all its descendants*
  - A search on Adrenal insufficiency also retrieves articles indexed with Addison disease

▶ [Adrenal Insufficiency \[C19.053.500\]](#)

[Addison Disease \[C19.053.500.263\]](#)

[Adrenoleukodystrophy \[C19.053.500.270\]](#)

[Hypoaldosteronism \[C19.053.500.480\]](#)

[Waterhouse-Friderichsen Syndrome \[C19.053.500.740\]](#)



Search PubMed for "adrenal insufficiency"[MeSH Terms]   [Advanced Search \(beta\)](#)  
[Save Search](#)

Display Summary Show 20 Sort By Send to

All: 8994 Review: 1069

Items 1 - 20 of 8994

Page 1 of 450 Next

1: [Bendel S, Karlsson S, Pettilä V, Loisa P, Varpula M, Ruokonen E; Finnsepsis Study Group.](#) [Related Articles, Links](#)



Free cortisol in sepsis and septic shock.

Anesth Analg. 2008 Jun;106(6):1813-9.  
PMID: 18499615 [PubMed - indexed for MEDLINE]

2: [Luboshitzky R, Qupti G.](#) [Related Articles, Links](#)



Corticosteroids for septic shock.

N Engl J Med. 2008 May 8;358(19):2069; author reply 2070-1. No abstract available.  
PMID: 18467975 [PubMed - indexed for MEDLINE]

12: [Lövås K, Husebye ES.](#) [Related Articles, Links](#)



Replacement therapy for Addison's disease: recent developments.

Expert Opin Investig Drugs. 2008 Apr;17(4):497-509. Review.  
PMID: 18363515 [PubMed - indexed for MEDLINE]

# Co-indexing

**gpubmed**

*Searching is now sorted!*

<http://www.gpubmed.com/>



COX-2



what

## Top categories

- [-] (M) Chemicals and Drugs [992]
  - (M) Cyclooxygenase 2 [517]
  - (M) Cyclooxygenase 2 Inhibitors [289]
  - (M) Prostaglandins [358]
  - (M) Prostaglandin-Endoperoxide Synthases [667]
  - (M) NF-kappa B [138]
  - (M) RNA, Messenger [222]
  - (M) Anti-Inflammatory Agents [414]
  - more
- [-] (G) biological\_process [851]
  - (G) cyclooxygenase pathway [305]
  - more
- [-] (M) Biological Sciences [960]
  - (M) Up-Regulation [166]
  - more
- [-] (M) Diseases [781]
  - (M) Inflammation [192]
  - more
- [+] (M) Organisms [398]
- [+] (M) Techniques and Equipment [809]
- [+] (G) molecular\_function [483]
- [+] (M) Anatomy [778]
- [+] (M) Named Groups [285]
- [+] (G) cellular\_component [307]
- [+] (M) Natural Sciences [661]
- [+] (M) Technology, Industry, Agriculture [147]
- [+] (M) Psychiatry and Psychology [386]



Lister Hill National Center for E

# Resources for text annotation

- ◆ Named entity recognition software
  - MetaMap (UMLS) – NLM  
<http://metamap.nlm.nih.gov/>
  - NCBO Annotator (BioPortal) – NCBO  
<http://bioportal.bioontology.org/annotator#>
  - Many other academic tools (e.g., NaCTeM)
  - Many commercial tools
- ◆ Annotated corpora
  - CALBC – EBI  
<http://www.calbc.eu/>



## 2. Data integration, exchange and semantic interoperability

# “Standards”

- ◆ Ontologies help standardize patients data
  - Facilitate the exchange of data across institutions
  - Help connect “islands of data” (silos)
- ◆ In conjunction with
  - Information models
    - Clinical research (e.g., BRIDG, CDISC)
    - Healthcare (e.g., CDA)
  - Messaging standards (HL7 messaging)  
e.g., exchange orders/results between labs and hospitals

# “Meaningful use” of health IT

## ◆ 3 major standards

- **SNOMED CT** – Clinical documentation

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

- Quality measures, decision support

- **RxNorm** – Medications

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

- Quality measures, medication reconciliation, pharmacovigilance

- **LOINC** – Lab tests and observations

<http://loinc.org/>

- Quality measures, public health (newborn screening)



# Semantic interoperability projects caCORE

## ◆ Cancer Common Ontologic Representation Environment

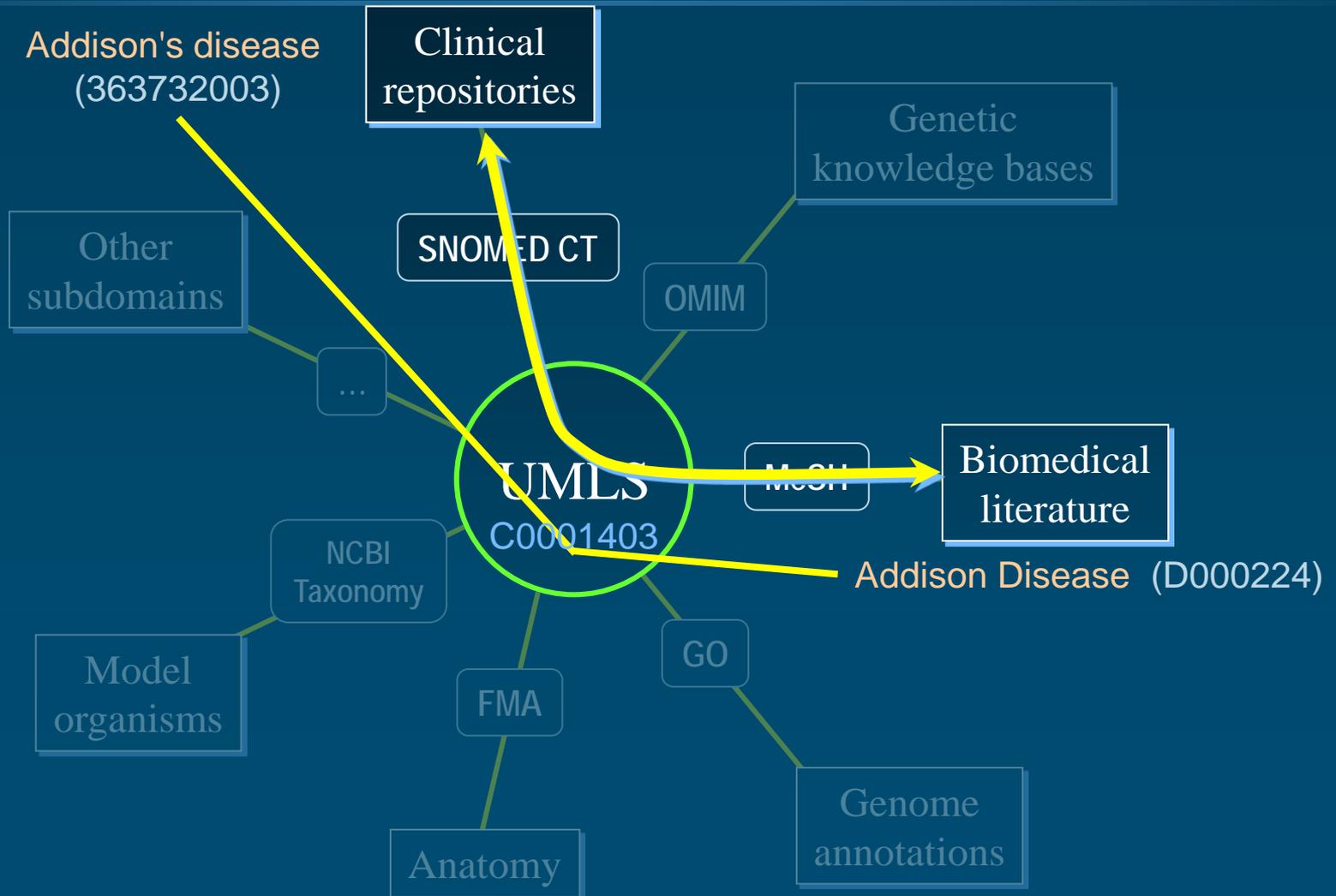
- Infrastructure developed to support an interoperable biomedical information system for cancer research
- Uses the **NCI Thesaurus** as a component  
<http://ncit.nci.nih.gov/>



# Data integration

- ◆ Warehousing approach
  - Ontologies help define the schema and convert data
- ◆ Mediation / Federation
  - Ontologies help convert between local and global schemas
- ◆ Linked data / Mashups
  - Ontologies provide a reference for equivalent entities across datasets

# Trans-namespace integration



### 3. Decision support and reasoning

*Data selection and aggregation, decision support, natural language processing applications, knowledge discovery*

# Decision support

## ◆ Clinical decision support

- Ontologies help normalize the vocabulary and increase the recall of rules
- Ontologies provide some domain knowledge and make it possible to create high-level rules (e.g., for a class of drugs rather than for each drug in the class)

## ◆ Other forms of decision support

- Based on automatic reasoning services for OWL ontologies (e.g., grading gliomas with NCIt)



# Knowledge discovery

- ◆ By standardizing the vocabulary in a given domain, ontologies are enabling resources for knowledge discovery through data mining
- ◆ Less frequently, the structure of the ontology is leveraged by data mining algorithms
- ◆ Example of available datasets
  - ICD-coded clinical data (in conjunction with non-clinical information, e.g., environmental data)
  - Annotation of gene products to the GO (function prediction)

# Barriers to usability of biomedical ontologies

# 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: 156 source vocabularies  
(biased towards healthcare applications)  
<http://www.nlm.nih.gov/research/umls/>
- NCBO BioPortal: ~200 ontologies  
(biased towards biological applications)  
<http://bioportal.bioontology.org/>
- Some overlap between the two repositories

## ◆ Need for discovery services



# 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)

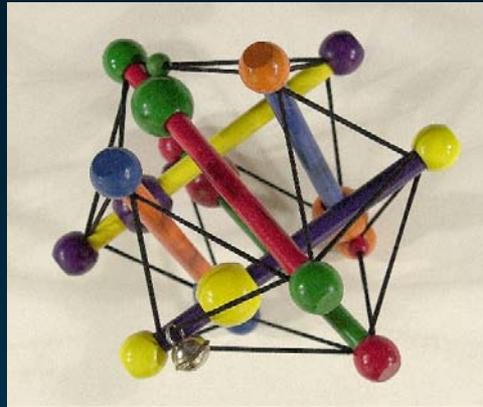


# Ontology integration

- ◆ *Post hoc* integration, form the bottom up
  - UMLS approach
  - Integrates ontologies “as is”, including legacy ontologies
  - Facilitates the integration of the corresponding datasets
  - **Current harmonization efforts (e.g., IHTSDO)**
- ◆ 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



# 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