

Ontologies: What Librarians Need to Know

Barry Smith
Department of Philosophy
University at Buffalo

Some questions

- How to find data?
- How to understand data when you find it?
- How to use data when you find it, for example in hypothesis-checking and reasoning?
- How to integrate with other data?
- How to label the data you are collecting?
- How to build as set of labels for a new domain that will integrate well with labels used in neighboring domains?

Network effects of the Web

- You build a site.
- Others discover the site and they link to it
- The more they link to it, the more important and well known the page becomes (this is what Google exploits)
- Your page becomes important, and others begin to rely on it
- *The same network effect works on the raw data*
 - Many people link to the data, use it
 - Many more (and diverse) applications will be created than the authors would even dream of
- New 'secondary uses' are discovered

The problem: doing it this way, we end up with data in many, many silos because links are formed in overlapping and redundant ways



The roots of Semantic Technology

To avoid silos:

1. The raw data must be available in a standard way on the Web.
2. There must be links among the datasets to create a 'web of data'

Use ontologies to capture common meanings with definitions that are understandable to both humans and computers

Ontologies as controlled vocabularies for the tagging of data

- Hardware changes rapidly
- Organizations rapidly forming and disbanding collaborations
- Data is exploding
- Meanings of common words change slowly
- Use web architecture to annotate exploding data stores using ontologies exploiting these common meanings

Mandates for Data Reuse

- Organizations such as the NIH now require use of common standards in a way that will ensure that the results obtained through funded research are more easily accessible to external groups
- http://grants.nih.gov/grants/policy/data_sharing/
- <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>
- **Data Ontologies for Biomedical Research (R01):**
<http://grants.nih.gov/grants/guide/pa-files/par-07-425.html>


NCBO: National Center for Biomedical Ontology

- Stanford Biomedical Research Informatics
- Mayo Clinic Department of Bioinformatics
- University at Buffalo, Department of Philosophy

<http://bioportal.bioontology.org/>

NCBO Bioportal

[Browse](#)[Search](#)[Mappings](#)[Recommender](#)[Annotator](#)[Resource Index](#)[Projects](#)

Welcome to BioPortal! For help using BioPortal, click on this icon: 

Search all ontologies

[Search](#)[Advanced Search](#)

Find an ontology

[Explore](#)[Browse Ontologies >](#)

Most Viewed Ontologies (October, 2011)

Ontology	Views
National Drug File	7975
SNOMED Clinical Terms	3977
MedDRA	3293
Medical Subject Headings	1620
NCI Thesaurus	1227

Latest Notes

[RE: what's the difference between this note and notes on mappings? \(Biomedical Resource Ontology\)](#) about 1 month ago by whetzel
The Term Notes refer to comments or actions requested on Terms in the ontology. The Mapping Notes...

[what's the difference between this note and notes on mappings? \(Biomedical Resource Ontology\)](#) about 1 month ago by imposimon
Would you guys please answer this for me?

Goals of Semantic Technology

To support data reuse

To enable data registries

Metadata management

Support for Natural Language Understanding

Semantic Wikis

via **ontologies** formulated for example in
the Web Ontology Language (OWL)

Where we stand today

- html demonstrated the power of the Web to allow sharing of information
- increasing availability of semantically enhanced data
- increasing power of semantic software to allow automatic reasoning with online information
- increasing use of OWL in attempts to break down silos, and create useful integration of on-line data and information






as of September 2010

The result: the more Semantic Technology is successful, the more it fails to achieve its goals

As we break down silos via controlled vocabularies for the tagging of data the very success of the approach leads to the creation of ever new controlled vocabularies , **semantic silos** – because multiple ontologies are being created in ad hoc ways

The Semantic Web framework as currently conceived and governed by the W3C yields minimal standardization

Term Search

Search for a term across multiple ontologies 

Select ontologies to search

[select from list](#)

Search

534 results

TERM NAME		MATCHED IN filter	ONTOLOGY filter
Obesity	details visualize	Preferred Name	Bone Dysplasia Ontology
Obesity	details visualize	Preferred Name	ICD10
Obesity	details visualize	Preferred Name	Common Terminology Criteria for Adverse Events
obesity	details visualize	Preferred Name	Neuro Behavior Ontology
obesity	details visualize	Preferred Name	CRISP Thesaurus, 2006
Obesity	details visualize	Preferred Name	Read Codes, Clinical Terms Version 3 (CTV3)
obesity	details visualize	Preferred Name	Bleeding History Phenotype
Obesity	details visualize	Preferred Name	Human Phenotype Ontology
obesity	details visualize	Preferred Name	Human disease
obesity	details visualize	Preferred Name	eVOC (Expressed Sequence Annotation for Humans)
obesity	details visualize	Preferred Name	Experimental Factor Ontology
Obesity	details visualize	Preferred Name	NCI Thesaurus

Obesity	details visualize	Preferred Name	HEALTH INDICATORS
obesity	details visualize	Preferred Name	NIFSTD
Obesity	details visualize	Preferred Name	NIFSTD
Obesity	details visualize	Preferred Name	HOM ElixhauserScores
OBESITY	details visualize	Preferred Name	WHO Adverse Reaction Terminology
Obesity	details visualize	Preferred Name	National Drug File
Obesity	details visualize	Preferred Name	Medical Subject Headings
Obesity	details visualize	Preferred Name	PMA 2010
Obesity	details visualize	Preferred Name	MedlinePlus Health Topics
obesity	details visualize	Preferred Name	Physician Data Query
OBESITY	details visualize	Preferred Name	Online Mendelian Inheritance in Man
Obesity	details visualize	Preferred Name	Online Mendelian Inheritance in Man
Obesity	details visualize	Preferred Name	Cell line ontology
obesity	details visualize	Preferred Name	RadLex in OWL
Obesity	details visualize	Preferred Name	SNOMED Clinical Findings
Adiposity	details visualize	Synonym	SNOMED Clinical Terms
large body habitus artifact	details visualize	Synonym	RadLex
Large Body Habitus Artifact	details visualize	Synonym	NCI Metathesaurus
morbid obesity	details visualize	Synonym	Experimental Factor Ontology
Morbid obesity	details visualize	Synonym	NIFSTD
obese	details visualize	Synonym	Mammalian phenotype

Reasons for this effect

- Shrink-wrapped software mentality – you will not get paid for reusing old and good ontologies (Let a million ‘lite’ ontologies bloom)
- Belief that there are no ‘good’ ontologies (just arbitrary choices of terms and relations ...)
- Information technology (hardware) changes constantly, not worth the effort of getting things right
- We have done it this way for 30 years, we are not going to change now

Unified Medical Language System of the National Library of Medicine

- let a million ontologies bloom, each one close to the terminological habits of its authors
- in concordance with the “not invented here” syndrome
- then map these ontologies, and use these mappings to integrate your different pots of data

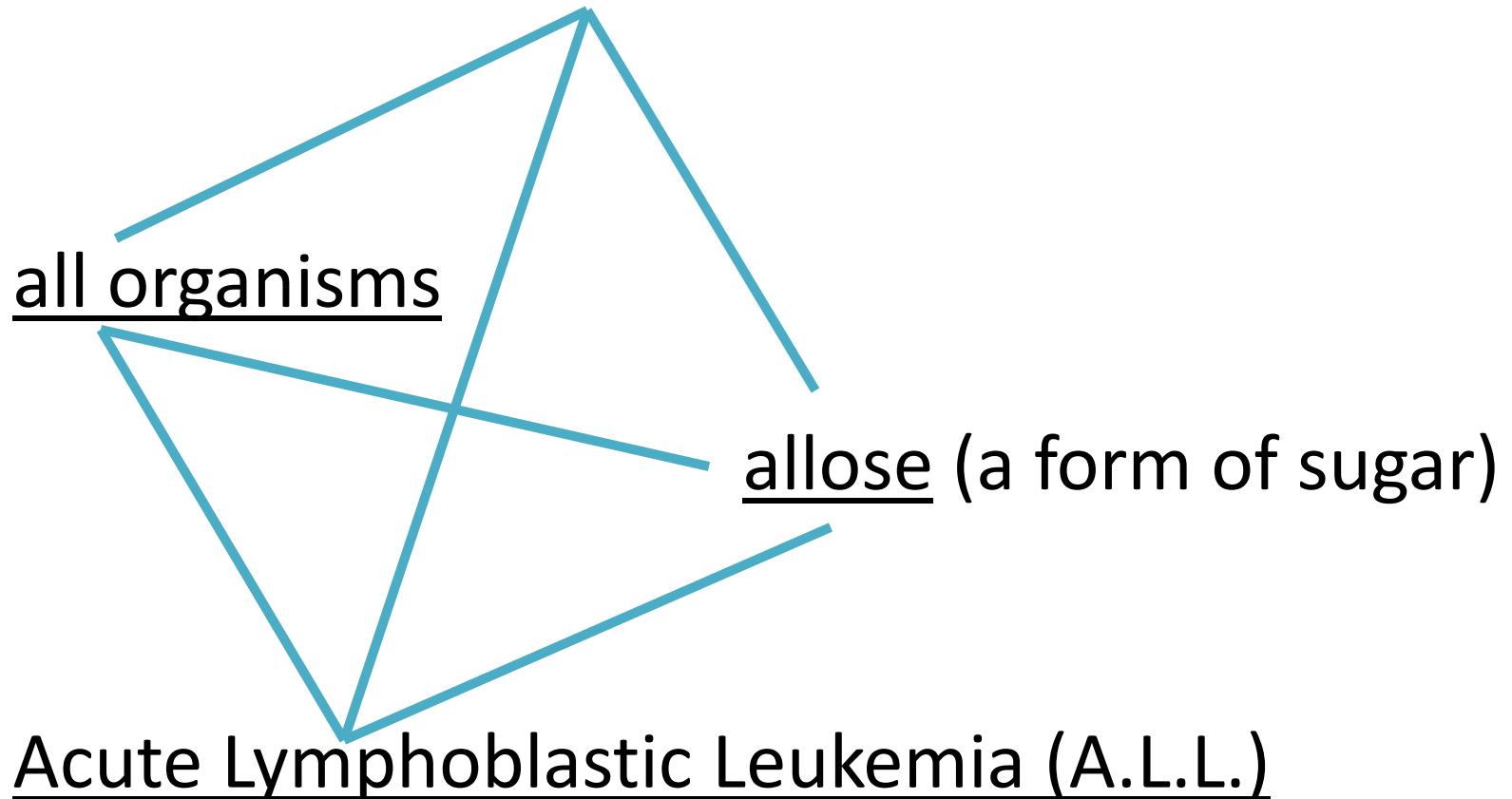
What you get with 'mappings'

all phenotypes (excess hair loss, duck feet)

all organisms

allose (a form of sugar)

Acute Lymphoblastic Leukemia (A.L.L.)



Mappings are hard

They are fragile, and expensive to maintain

Need a new authority to maintain, yielding new risk of forking

The goal should be to minimize the need for mappings

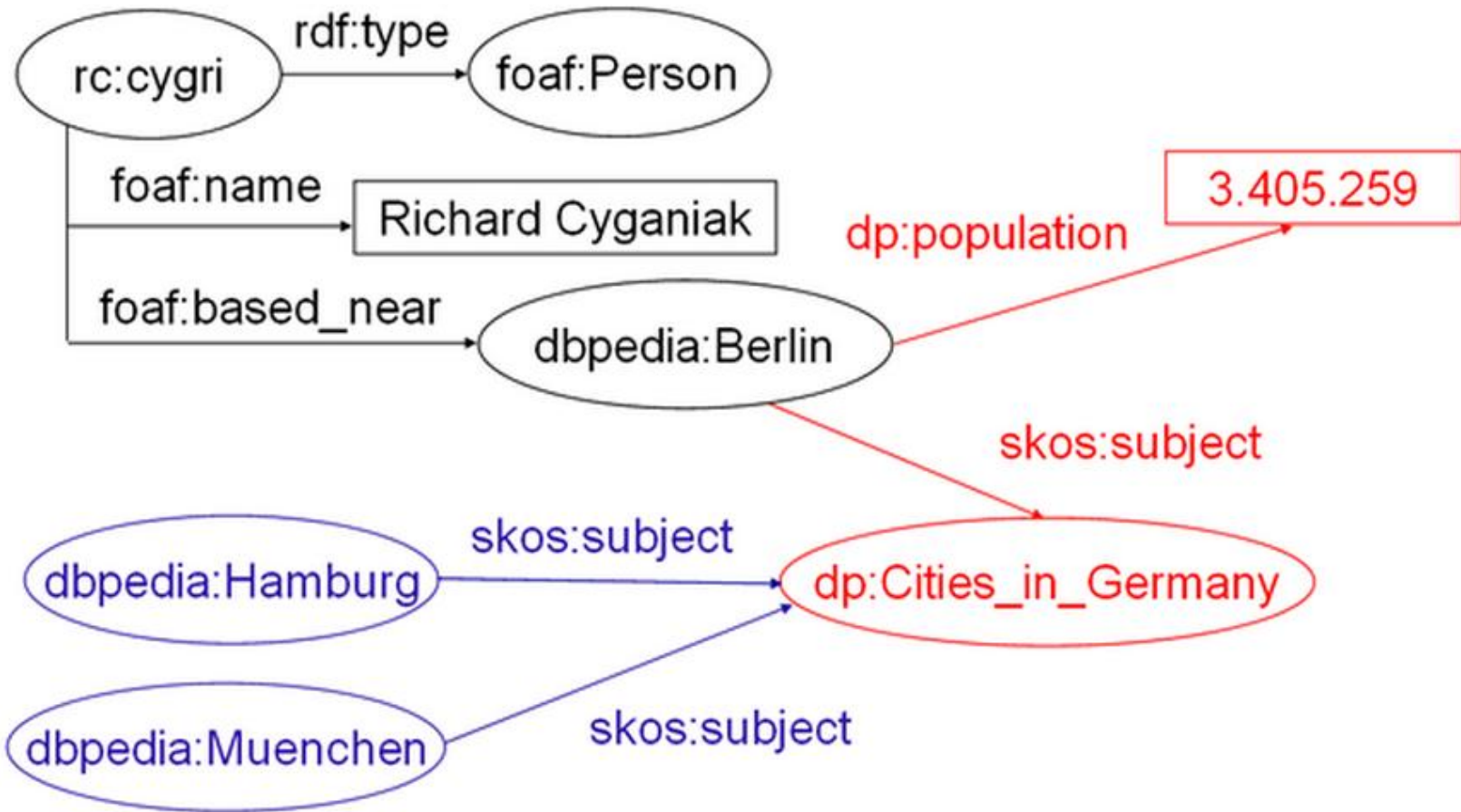
Invest resources in disjoint ontology modules which work well together

Why should you care?

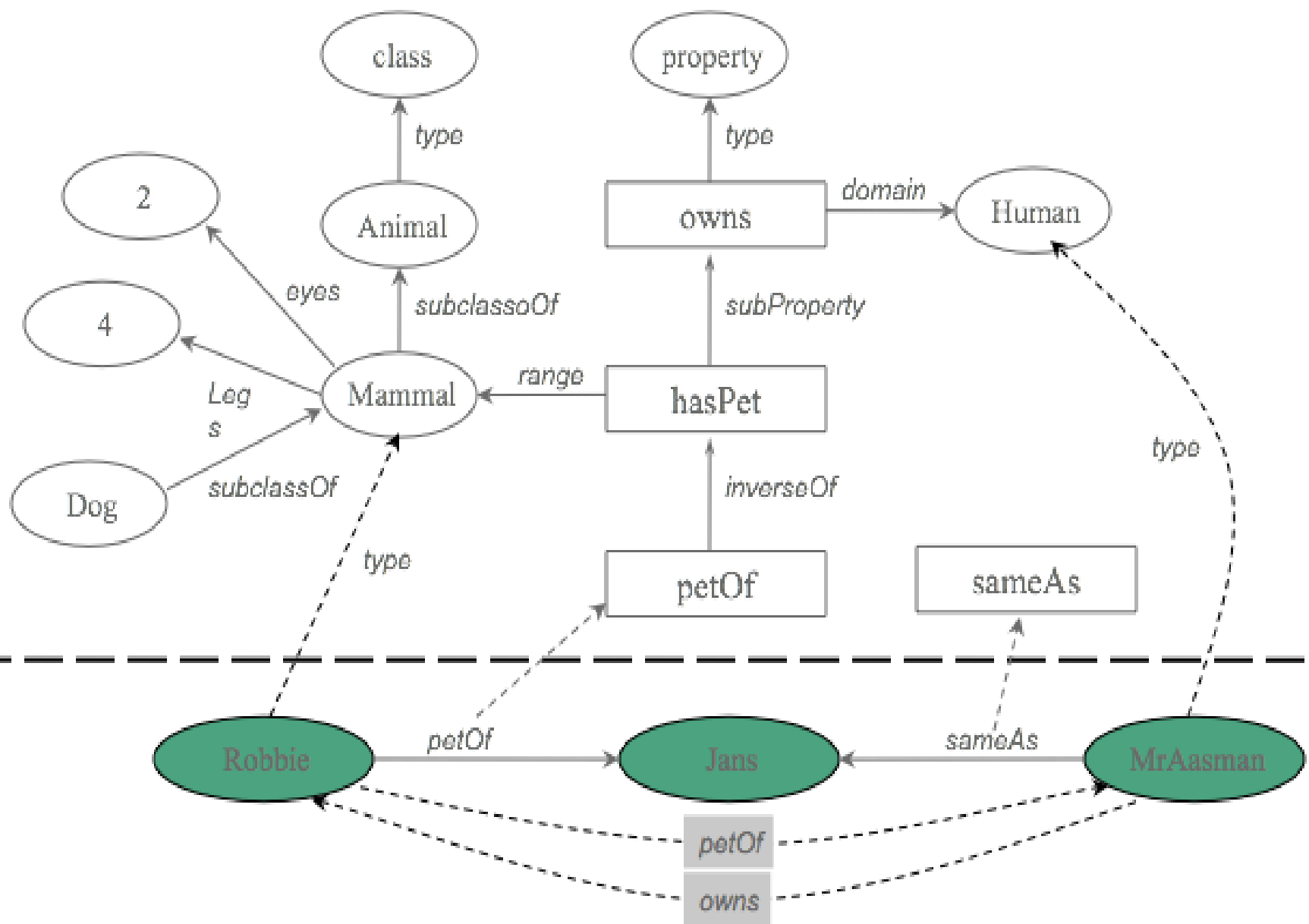
- you need to create systems for data mining and text processing which will yield useful output for library users
- if the codes you use are constantly in need of ad hoc repair huge resources will be wasted, manual effort will be needed on each occasion of use
- DoD alone spends \$6 billion per annum on this problem

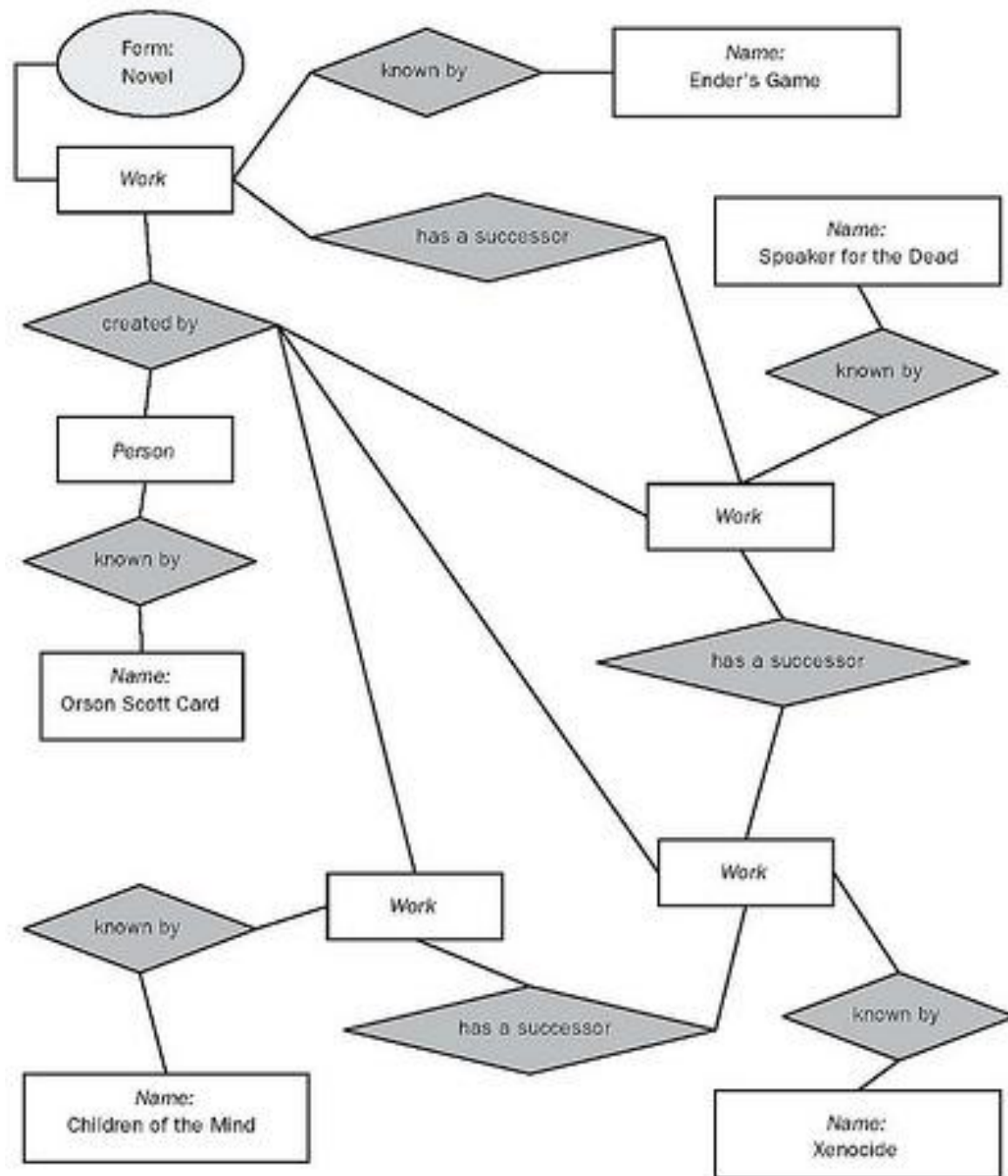
And there are other problems

- Weak expressivity of OWL
- Poor quality coding, poor quality ontologies, poor quality ontology management
- Strategy often serves only retrieval, not reasoning
- Confusion as to the meaning of ‘linked’



Uncontrolled proliferation of links

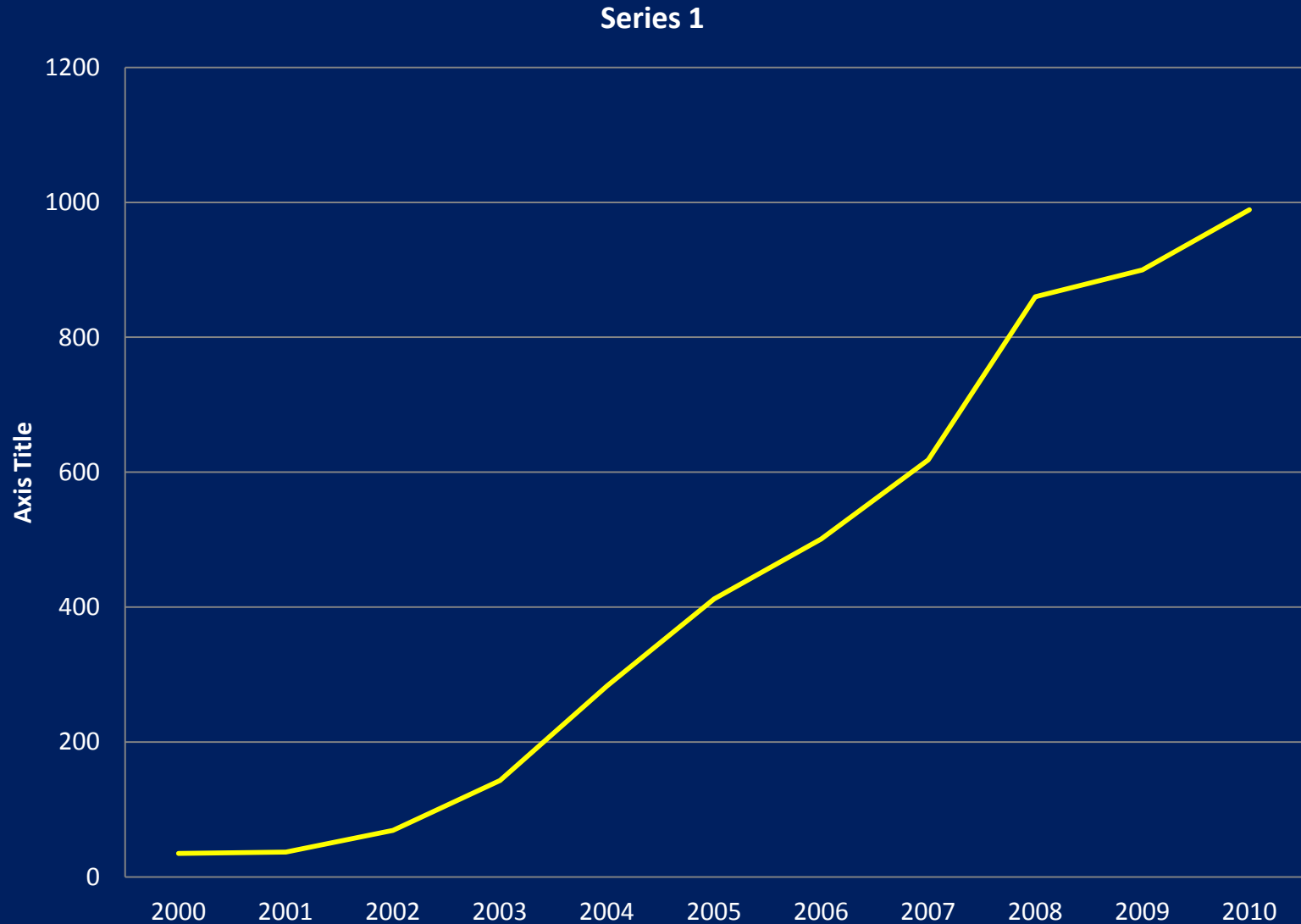




How to do it right?

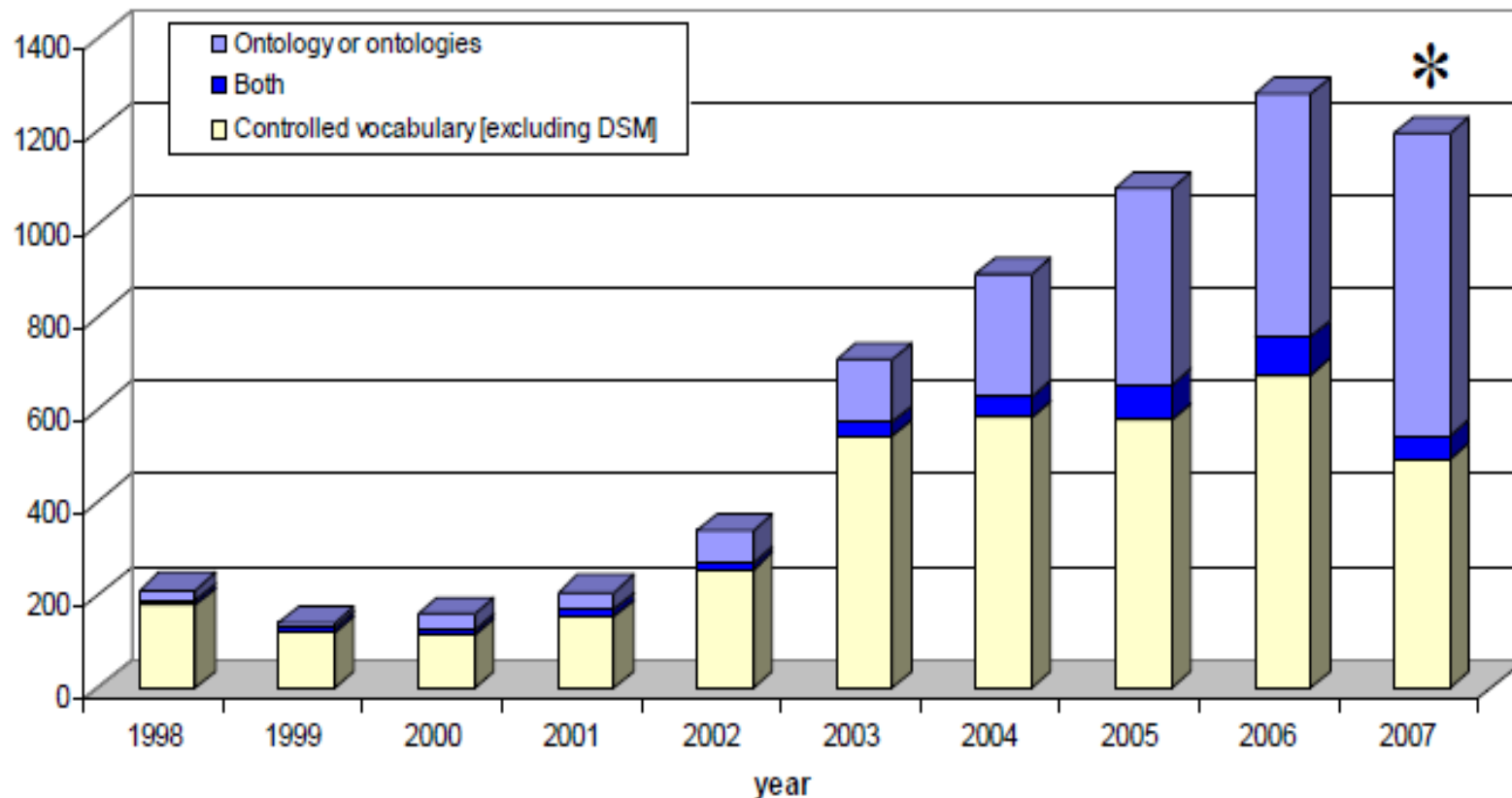
- how create an incremental, evolutionary process, where what is good survives, and what is bad fails
- create a scenario in which people will find it profitable to reuse ontologies, terminologies and coding systems which have been tried and tested
- silo effects will be avoided and results of investment in Semantic Technology will cumulate effectively

Ontology in PubMed



Biomedical ontology in PubMed

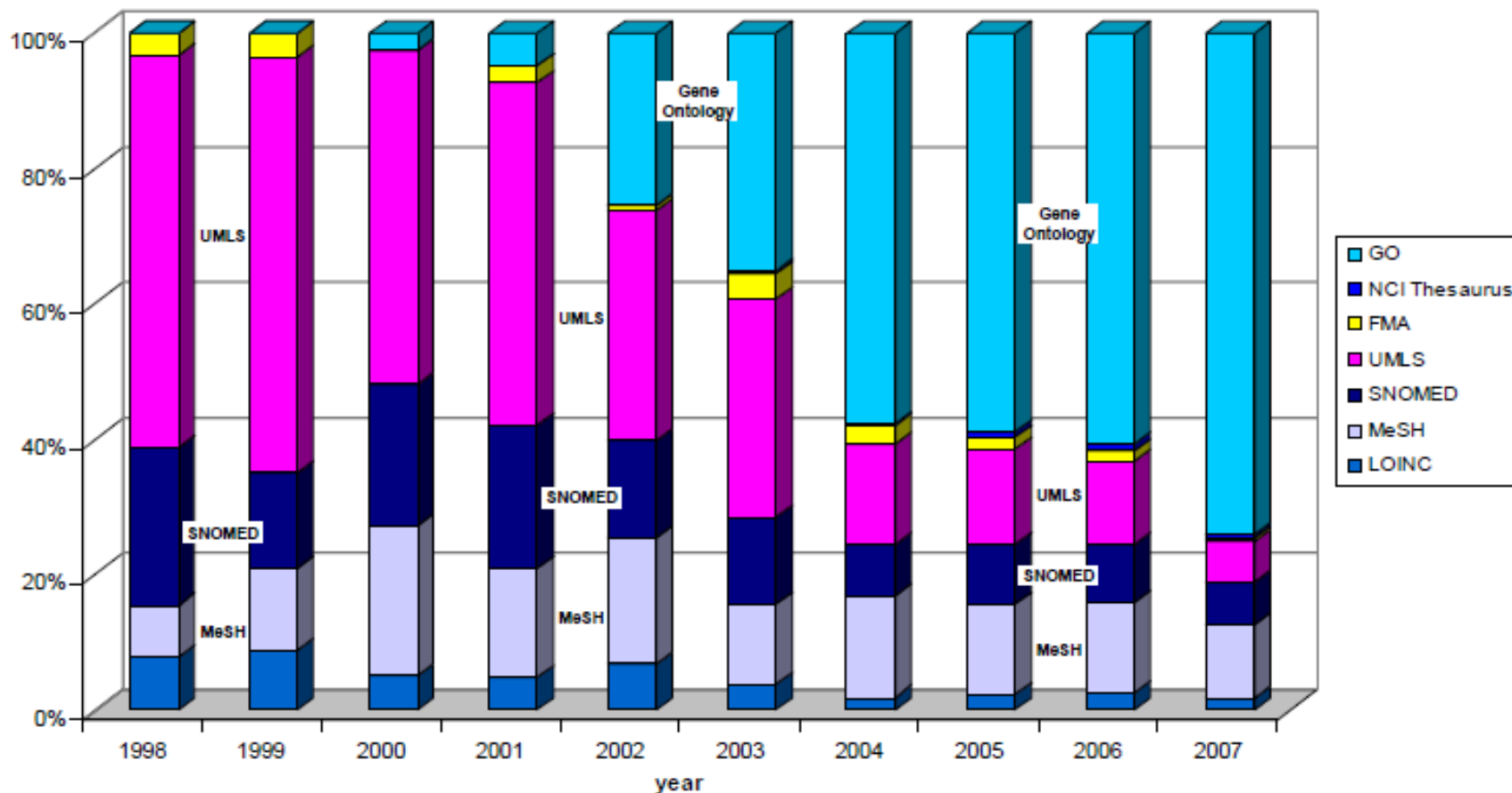
Number of articles in PubMed/MEDLINE on Ontology vs. Controlled vocabulary



(*) As of 2008/02/20
(Partial coverage for 2007, due to a slight lag in the indexing process)

By far the most successful: GO (Gene Ontology)

Proportion of citations in PubMed/MEDLINE by ontology



GO provides a controlled vocabulary of terms for use in annotating (describing, tagging) data

- multi-species, multi-disciplinary, open source
- contributing to the cumulativity of scientific results obtained by distinct research communities
- compare use of kilograms, meters, seconds in formulating experimental results
- natural language and logical definitions for all terms to support consistent human application and computational exploitation

You're interested
in which genes
control heart
muscle
development

17,536 results

The screenshot shows the NCBI PubMed website interface. At the top, the NCBI logo is on the left, and the PubMed logo with the URL 'www.pubmed.gov' is on the right. Below the logos, a navigation bar contains links to 'All Databases', 'PubMed', 'Nucleotide', 'Protein', 'Genome', and 'Structure'. A search bar is present with 'PubMed' selected in a dropdown menu and 'heart muscle development' entered in the text field. Below the search bar are buttons for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. Further down, there are controls for 'Display' (set to 'Summary'), 'Show' (set to '20'), 'Sort By', and 'Send to'. A summary bar indicates 'All: 17536' results and 'Review: 2524'. The main content area is titled 'Items 1 - 20 of 17536' and lists four search results, each with a checkbox, a citation number, a link to the full text, a document icon, and a brief description of the article.

NCBI PubMed
A service of the U.S. National Library of Medicine and the National Institutes of Health
www.pubmed.gov

All Databases PubMed Nucleotide Protein Genome Structure

Search PubMed for heart muscle development

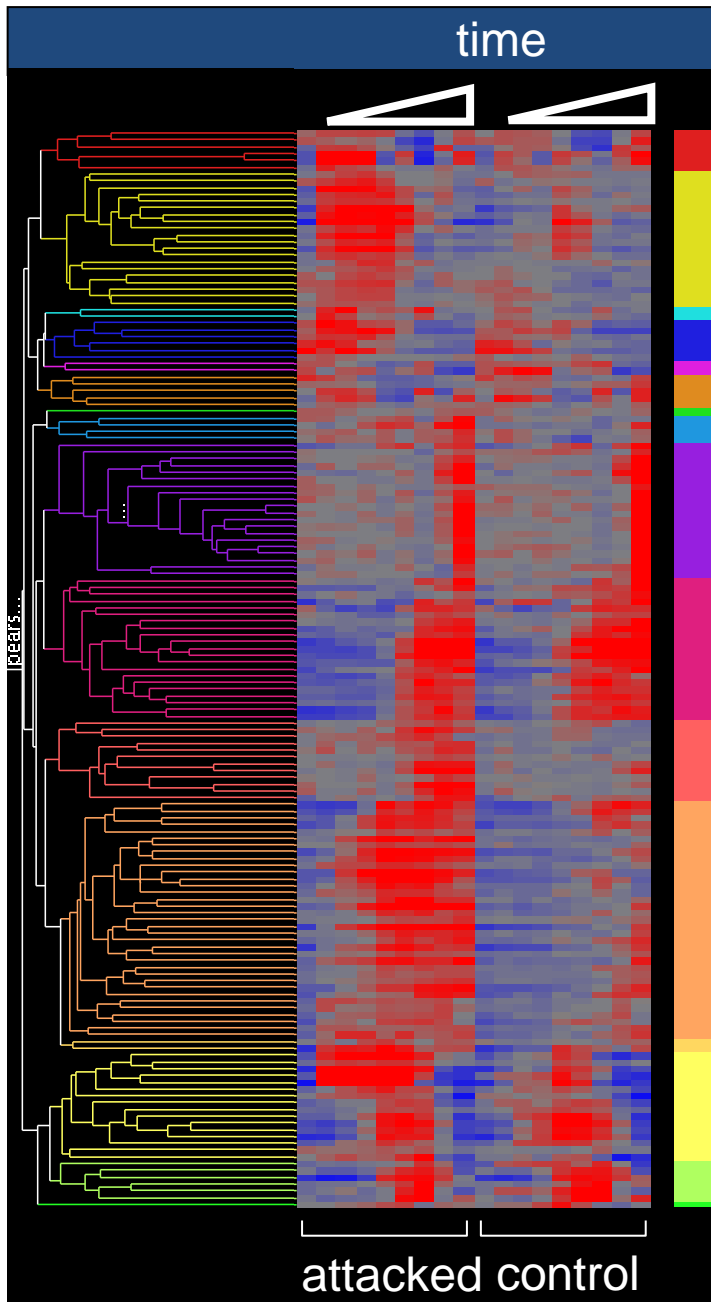
Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort By Send to

All: 17536 Review: 2524

Items 1 - 20 of 17536

- ☐ 1: [Tong CW, Stelzer JE, Greaser ML, Powers PA, Moss RL.](#)
Acceleration of Crossbridge Kinetics by Protein Kinase A Phosphorylation of Ca
Circ Res. 2008 Sep 18. [Epub ahead of print]
PMID: 18802026 [PubMed - as supplied by publisher]
- ☐ 2: [Pentassuglia L, Sawyer DE.](#)
The role of Neuregulin-1beta/ErbB signaling in the heart.
Exp Cell Res. 2008 Sep 3. [Epub ahead of print]
PMID: 18801360 [PubMed - as supplied by publisher]
- ☐ 3: [Takamatsu T.](#)
Arrhythmogenic substrates in myocardial infarct.
Pathol Int. 2008 Sep;58(9):533-43.
PMID: 18801068 [PubMed - in process]
- ☐ 4: [Lu S, Borst DE, Horowitz R.](#)
Expression and alternative splicing of N-RAP during mouse skeletal muscle devel
Cell Motil Cytoskeleton. 2008 Sep 15. [Epub ahead of print]
PMID: 18792955 [PubMed - as supplied by publisher]



Microarray data shows changed expression of thousands of genes.

How will you spot the patterns?

		Studies	Lesions	PET (%)	Studies	Lesions	PET (%)	Studies	Lesions	PET (%)	Studies	Lesions	EFFECT (%)
Bladder	Staging	136		76	98		87	98		83	12		17
	Dx/Staging	52		93	26		86	26		88			
	Recurrence	12		60							12		17
Brain	Dx	36		91									
	Staging	31		86									
	Recurrence	258		79	213		77	161		76	89		31
Breast	Mon Response	17		82	17		83						
	Other	34		93	19		67	19		84			
	Dx	202		91	97		93	105		95	6		100
			140	86		140	86		105	86			
	Staging	1407											24
	Dx/Staging	65	242										
	Recurrence	414											40
			41										
	Mon Response	206											
Colorectal			31										
	Staging		24										36
	Dx/Staging	101											
	Recurrence	1426											32
			981										
	Mon Response		34										
Gastro-Esoph	Dx	120											14
			276										
	Staging	545											20
			15										
	Dx/Staging	109											14
	Recurrence	41											
Head&Neck	Mon Response	13											
	Dx	129											
			311										
	Staging	363											
			2020										
	Dx/Staging	296											33
	Recurrence	342											33
			278	84		241	92		241	90			
	Mon Response	128		84	122		95	81		96			
			16	44									
Hepatocellular	Staging	292		77	249		97	249		93	20		60
	Dx/Staging	22		64									
	Recurrence		9	88									
Lung	Dx	919		96	797		73	719		90			
			278	91		259	68		101	82			
	Staging	1867		83	1495		91	1272		82	1565		37
			1721	83		1553	92		1478	90			
	Recurrence	209		98	193		92	180		96			
			39	100		39	62		39	87			
	Mon Response	161		94	161		90	126		96			
Lymphoma	Other	101		83									
	Dx	11		100									
	Staging	1179		90	826		93	158		88	407		21
			1156	91		58	100		32	95			
	Dx/Staging	254		92	177		93				62		5
	Recurrence	557		87	453		93	155		88	158		10
			114	100									
	Mon Response	257		90	279		93	13		69			
Melanoma									32	95			
	Staging	888		83	863		91	125		91	283		26
			899	87		461	68		83	84			

You're interested in which of your hospital's patient data is relevant to understanding how genes control heart muscle development

		Studies	Lesions	PET (%)	Studies	Lesions	PET (%)	Studies	Lesions	PET (%)	Studies	Lesions	EFFECT (%)
Bladder	Staging	136		76	98		87	98		83	12		17
	Dx/Staging	52		93	26		86	26		88			
	Recurrence	12		60							12		17
Brain	Dx	36		91									
	Staging	31		86									
	Recurrence	258		79	213		77	161		76	89		31
	Mon Response	17		82	17		83						
	Other	34		93	19		67	19		84			
	Dx	202		91	97		93	105		95	6		100
			140										
	Staging	1407											24
			242										
	Dx/Staging	65											
	Recurrence	414											40
			41										
	Mon Response	206											
			31										
	Staging		24										36
Colorectal													
	Dx/Staging	101											
	Recurrence	1426											32
			981										
	Mon Response		34										
Gastro-Esoph	Dx	120											14
			276										
	Staging	545											20
			15										
	Dx/Staging	109											14
	Recurrence	41											
Head&Neck	Mon Response	13											
	Dx	129											
			311										
	Staging	363											
			2020										
	Dx/Staging	296											33
			179										
	Recurrence	342											33
			278										
	Mon Response	128											
			16										
Hepatocellular	Staging	292											60
	Dx/Staging	22											
	Recurrence		9										
Lung	Dx	919											
			278										
	Staging	1867											37
			1721										
	Recurrence	209											
			39										
	Mon Response	161											
	Other	101											
	Dx	11											
Lymphoma	Staging	1179											21
			1156										
	Dx/Staging	254											5
	Recurrence	557											10
			114										
	Mon Response	257											
Melanoma				83	863		91	125	32	95			
	Staging	888		87		461	68		83	84	283		26
			899										

Lab / pathology data

EHR data

Clinical trial data

Family history data

Medical imaging

Microarray data

Model organism data

Flow cytometry

Mass spec

Genotype / SNP data

How will you spot the patterns?

How will you find the data you need?

One strategy for bringing order into this huge conglomeration of data is through the use of Common Data Elements

- Discipline-specific (cancer, NIAID, ...)
- Do not solve the problems of balkanization (data siloes)
- Do not evolve gracefully as knowledge advances
- Support data cumulation, but do not readily support data integration and computation

How does the Gene Ontology work?

with thanks to
Jane Lomax, Gene Ontology Consortium

GO provides a controlled system of representations for use in annotating data

- multi-species, multi-disciplinary, open source
- contributing to the cumulativity of scientific results obtained by distinct research communities
- compare use of kilograms, meters, seconds ... in formulating experimental results

Term Search Results

11 results for **heart muscle development** in terms fields **term accession, term name, and synonyms**

▼ **Filter search results** [?](#)

Ontology

All
 biological process
 cellular component
 molecular function

Results are sorted by **relevance**. To change the sort order, click on the column headers.

<div> <input type="button" value="Select all"/> <input type="button" value="Clear all"/> <input type="text" value="Perform an action with the selected terms..."/> <input type="button" value="Go!"/> </div>			
rel ↓	Accession , Term		Ontology
<input checked="" type="checkbox"/>	GO:0048738 : cardiac muscle development [show def]	31 gene products view in tree	biological process
	Query matches synonym " heart muscle development " [exact synonym]		
<input checked="" type="checkbox"/>	GO:0055013 : cardiac muscle cell development [show def]	27 gene products view in tree	biological process
	Query matches synonym " heart muscle cell development " [exact synonym]		
<input checked="" type="checkbox"/>	GO:0055014 : atrial cardiac muscle cell development [show def]	2 gene products view in tree	biological process
	Query matches synonym "atrial heart muscle development " [exact synonym]		
<input type="checkbox"/>	GO:0055024 : regulation of cardiac muscle development [show def]	0 gene products	biological process

Term Search Results

11 results for **heart muscle development** in terms fields **term accession, term name, and synonyms**

Current filters

Species: [Homo sapiens](#), [Human immunodeficiency virus 1](#)

▼ Filter search results ?

Ontology

Set filters

All
biological process
cellular component
molecular function

Remove all filters

Results are sorted by **relevance**. To change the sort order, click on the column headers.

Select all

Clear all

Perform an action with the selected

Go!

Definitions

rel ↓

Accession , Term

Ontology

☐
GO:0048738 : [cardiac muscle development](#) [\[hide term\]](#)
2 gene products
[view in tree](#)
[biological process](#)

The process whose specific outcome is the progression of cardiac muscle over time, from its formation to the mature structure.

Query matches synonym "**heart muscle development**" [exact synonym]

40

Definitions

Gene Product Associations to cardiac muscle development ; GO:0048738 and children

[gene association format](#) [RDF/XML](#)

Current filters

Species: [Homo sapiens](#), [Human immunodeficiency virus 1](#)

Filter associations displayed ?

Filter by Gene Product

Gene Product Type

All
complex
gene
protein

Data source

All
CGD
dictyBase
EcoCyc

Species

All
Anaplasma phagocy...
Arabidopsis thaliana
Bacillus anthraci...

Filter by Association

Evidence Code

All
IC
IDA
EXP

View associations

☒ All ☐ Direct associations

Set filters

Remove all filters

cardiac muscle development ; GO:0048738 [\[show def\]](#) [\[view in tree\]](#)

Symbol, full name	Information	Qualifier	Evidence	Reference	Assigned by
<input type="checkbox"/> TAZ_HUMAN view associations TAZ, EFE2, G4.5: Tafazzin	protein from <i>Homo sapiens</i>		IMP	PMID:17043667	HGNC (via UniProtKB)

cardiac muscle fiber development ; GO:0048739 [\[show def\]](#) [\[view in tree\]](#)

Symbol, full name	Information	Qualifier	Evidence	Reference	Assigned by
<input type="checkbox"/> MYH11_HUMAN view associations MYH11, KIAA0866: Myosin-11	protein from <i>Homo sapiens</i>		BLAST		

Gene products involved in cardiac muscle development in humans

GO provides answers to three types of questions

for each gene product

- in what parts of the cell has it been identified?
- exercising what types of molecular functions?
- with what types of biological processes?

when is a particular gene product involved

- in the course of normal development?
- in the process leading to abnormality

with what functions is the gene product associated in other biological processes?

Some pain-related terms in GO

<u>GO:0048265</u>	response to pain
<u>GO:0019233</u>	sensory perception of pain
<u>GO:0048266</u>	behavioral response to pain
<u>GO:0019234</u>	sensory perception of fast pain
<u>GO:0019235</u>	sensory perception of slow pain
<u>GO:0051930</u>	regulation of sensory perception of pain
<u>GO:0050967</u>	detection of electrical stimulus during sensory perception of pain
<u>GO:0050968</u>	detection of chemical stimulus involved in sensory perception of pain
<u>GO:0050966</u>	detection of mechanical stimulus involved in sensory perception of pain

GO allows a new kind of biological research, based on analysis and comparison of the massive quantities of annotations linking GO terms to gene products

One standard method

Sjöblöm T, *et al.* analyzed 13,023 genes in 11 breast and 11 colorectal cancers

using functional information captured by GO for given gene product types

identified 189 as being mutated at significant frequency and thus as providing targets for diagnostic and therapeutic intervention.

Science. 2006 Oct 13;314(5797):268-74.

A new kind of biological research

based on analysis and comparison of the massive quantities of annotations **linking ontology terms to raw data**, including genomic data, clinical data, public health data

What 10 years ago took multiple groups of researchers months of data comparison effort, can now be performed in milliseconds

What is the key to GO's success?

- GO is developed, maintained and by experts who adhere to ontology best practices
- over 11 million annotations relating gene products described in the UniProt, Ensembl and other databases to terms in the GO
- experimental results reported in 52,000 scientific journal articles manually annotated by expert biologists using GO
- ontology building and ontology QA are two sides of the same coin

If controlled vocabularies are to serve data interoperability

they have to be used in annotations by many
owners of data

they have to be updated by respected experts who
are trained in best practices of ontology
maintenance

they have to be respected by many owners of data
as a framework for semantic enhancement that
ensures accurate description of their data

the benchmark for accuracy (the ground truth) is
provided by the results of scientific experiment

GO maintained not by computer scientists but by
biologists

The new profession of biocurator



ISB International Society
for Biocuration


[about the ISB](#)

[contact us](#)

- home
- what is biocuration?
- conferences
- education
- email list
- jobs
- join the ISB
- journals
- members
- members' affiliations
- mission
- networking
- publications

Welcome to the ISB

Welcome to the International Society for Biocuration (ISB). The ISB is a non-profit organisation for biocurators, developers, and researchers with an interest in biocuration. The society promotes the field of biocuration and provides a forum for information exchange through meetings and workshops.

For questions or comments, please email [the ISB helpdesk](#) .



ISB

International Society
for Biocuration



[home](#)



[what is biocuration?](#)



[conferences](#)



[education](#)



[email list](#)



[jobs](#)



[join the ISB](#)



[journals](#)



[members](#)



[members' affiliations](#)



[mission](#)



[networking](#)



[publications](#)

Welcome to the International S
biocurators, developers, and re
biocuration and provides a foru

How to do it right?

- how create an incremental, evolutionary process, where what is good survives, and what is bad fails
- where the number of ontologies needing to be linked is small
- where links are stable
- create a scenario in which people will find it profitable to reuse ontologies, terminologies and coding systems which have been tried and tested
- and in which ontologies will evolve on the basis of feedback from users

Reasons why GO has been successful

It is a system for **prospective standardization** built with coherent top level but with content contributed and monitored by domain specialists

Based on community consensus

Clear versioning principles ensure **backwards compatibility**; prior annotations do not lose their value

Initially low-tech to encourage adoption by new communities of users

Tracker for user input with rapid turnaround and help desk

But GO is limited in its scope

it covers only generic biological entities of three sorts:

- cellular components
- molecular functions
- biological processes

no diseases, symptoms, disease biomarkers,
protein interactions, experimental processes ...

Extending the GO methodology to other domains of biology and medicine

RELATION TO TIME GRANULARITY	CONTINUANT				OCCURRENT
	INDEPENDENT		DEPENDENT		
ORGAN AND ORGANISM	Organism (NCBI Taxonomy)	Anatomical Entity (FMA, CARO)	Organ Function (FMP, CPRO)	Phenotypic Quality (PaTO)	Biological Process (GO)
CELL AND CELLULAR COMPONENT	Cell (CL)	Cellular Component (FMA, GO)	Cellular Function (GO)		
MOLECULE	Molecule (ChEBI, SO, RnaO, PrO)		Molecular Function (GO)		Molecular Process (GO)

OBO (Open Biomedical Ontology) Foundry proposal
(Gene Ontology in yellow)

RELATION TO TIME GRANULARITY	CONTINUANT				OCCURRENT
	INDEPENDENT		DEPENDENT		
ORGAN AND ORGANISM	Organism (NCBI Taxonomy)	Anatomical Entity (FMA, CARO)	Organ Function (FMP, CPRO)	Phenotypic Quality (PaTO)	Biological Process (GO)
CELL AND CELLULAR COMPONENT	Cell (CL)	Cellular Component (FMA, GO)	Cellular Function (GO)		
MOLECULE	Molecule (ChEBI, SO, RnaO, PrO)		Molecular Function (GO)		Molecular Process (GO)

The strategy of orthogonal modules

Ontology	Scope	URL	Custodians
Cell Ontology (CL)	cell types from prokaryotes to mammals	obo.sourceforge.net/cgi-bin/detail.cgi?cell	Jonathan Bard, Michael Ashburner, Oliver Hofman
Chemical Entities of Biological Interest (ChEBI)	molecular entities	ebi.ac.uk/chebi	Paula Dematos, Rafael Alcantara
Common Anatomy Reference Ontology (CARO)	anatomical structures in human and model organisms	(under development)	Melissa Haendel, Terry Hayamizu, Cornelius Rosse, David Sutherland,
Foundational Model of Anatomy (FMA)	structure of the human body	fma.biostr.washington.edu	JLV Mejino Jr., Cornelius Rosse
Functional Genomics Investigation Ontology (FuGO)	design, protocol, data instrumentation, and analysis	fugo.sf.net	FuGO Working Group
Gene Ontology (GO)	cellular components, molecular functions, biological processes	www.geneontology.org	Gene Ontology Consortium
Phenotypic Quality Ontology (PaTO)	qualities of anatomical structures	obo.sourceforge.net/cgi-bin/detail.cgi?attribute_and_value	Michael Ashburner, Suzanna Lewis, Georgios Gkoutos
Protein Ontology (PrO)	protein types and modifications	(under development)	Protein Ontology Consortium
Relation Ontology (RO)	relations	obo.sf.net/relationship	Barry Smith, Chris Mungall
RNA Ontology (RnaO)	three-dimensional RNA structures	(under development)	RNA Ontology Consortium
Sequence Ontology (SO)	properties and features of nucleic sequences	song.sf.net	Karen Eilbeck

OBO Foundry

recognized by NIH as framework to address mandates for re-usability of data collected through Federally funded research

see NIH PAR-07-425: Data Ontologies for Biomedical Research (R01)

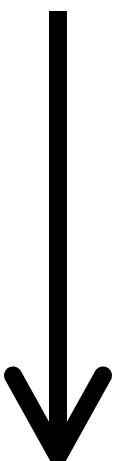
OBO Foundry provides

- tested **guidelines** enabling new groups to develop the ontologies they need in ways which counteract forking and dispersion of effort
- an incremental bottoms-up approach to evidence-based terminology practices in medicine that is **rooted in basic biology**
- automatic web-based linkage between biological knowledge resources (massive integration of databases across species and biological system)

RELATION TO TIME GRANULARITY	CONTINUANT				OCCURRENT
	INDEPENDENT		DEPENDENT		
ORGAN AND ORGANISM	Organism (NCBI Taxonomy)	Anatomical Entity (FMA, CARO)	Organ Function (FMP, CPRO)	Phenotypic Quality (PaTO)	Biological Process (GO)
CELL AND CELLULAR COMPONENT	Cell (CL)	Cellular Component (FMA, GO)	Cellular Function (GO)		
MOLECULE	Molecule (ChEBI, SO, RnaO, PrO)		Molecular Function (GO)		Molecular Process (GO)

The Open Biomedical Ontologies (OBO) Foundry

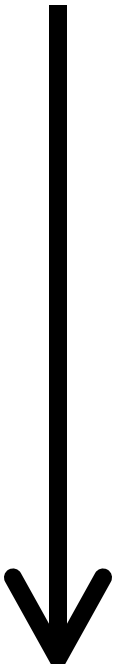
Governance



Basic Formal Ontology (BFO)

Information Artifact Ontology (IAO)		Ontology for Biomedical Investigations (OBI)		Ontology of General Medical Science (OGMS)	
Anatomy Ontology (FMA*, CARO)		Environment Ontology (EnvO)	Infectious Disease Ontology (IDO*)		Biological Process Ontology (GO*)
Cell Ontology (CL)	Cellular Component Ontology (FMA*, GO*)		Phenotypic Quality Ontology (PaTO)		
Subcellular Anatomy Ontology (SAO)					
Sequence Ontology (SO*)					
Protein Ontology (PRO*)					
		Molecular Function (GO*)			

Training



Basic Formal Ontology (BFO)

Information Artifact
Ontology
(IAO)

Ontology for Biomedical
Investigations
(OBI)

Ontology of General
Medical Science
(OGMS)

Anatomy Ontology
(FMA*, CARO)

Cell
Ontology
(CL)

Cellular
Component
Ontology
(FMA*, GO*)

Environment
Ontology
(EnvO)

Infectious
Disease
Ontology
(IDO*)

Phenotypic
Quality
Ontology
(PaTO)

Biological
Process
Ontology (GO*)

Subcellular Anatomy Ontology (SAO)

Sequence Ontology
(SO*)

Protein Ontology
(PRO*)

Molecular
Function
(GO*)

top level

Basic Formal Ontology (BFO)

mid-level

**Information Artifact
Ontology
(IAO)**

**Ontology for
Biomedical
Investigations
(OBI)**

**Spatial Ontology
(BSPO)**

**domain
level**

Anatomy Ontology
(FMA*, CARO)

Cell
Ontology
(CL)

Cellular
Component
Ontology
(FMA*, GO*)

Environment
Ontology
(EnvO)

Infectious
Disease
Ontology
(IDO*)

Phenotypic
Quality
Ontology
(PaTO)

Subcellular Anatomy Ontology (SAO)

Sequence Ontology
(SO*)

Protein Ontology
(PRO*)

Molecular
Function
(GO*)

Biological
Process
Ontology (GO*)

Extension Strategy + Modular Organization 64

How to build an ontology

1. due diligence: identify the existing ontology content that is most relevant to your needs
2. work with domain experts to identify parts of the domain not covered by this ontology
3. find ~50 most commonly used terms corresponding to types of entities in this domain
4. arrange these terms into a taxonomical hierarchy using the strategy of downward population
5. work with domain experts to populate the lower levels of the hierarchy

Example: The Cell Ontology

SUBCLASS EXPLORER

For Project:  DC_CL

Asserted Hierarchy

 owl:Thing

  Entity ≡ Entity

  Continuant

  DependentContinuant

  IndependentContinuant

 FiatObjectPart

  Object

  Biological_Macromolecule ≡ Biological_Macromolecule

  Cell

  CD11c_Low__Plasmacytoid_Dendritic_Cell

  CD11c_Negative_Plasmacytoid_Dendritic_Cell

  Conventional_Dendritic_Cell

  CD8_alpha_Neg_CD11b_Neg_Dendritic_Cell

 Immature_CD8_alpha_Neg_CD11b_Neg_Dendritic_Cell

 Mature_CD8_alpha_Neg_CD11b_Neg_Dendritic_Cell

Ontology and Library Science

- Nanopublishing
- FaBRO
- Semantically enhanced publishing
- eagle-I and VIVO resource registry i

Nanopublishing

- **Definition** – An online publishing model that uses a scaled-down, inexpensive operation to reach a targeted audience, especially by using blogging techniques
- **Applied to ontologies** – gives credit to authors of fragments of ontologies, including single ontology terms and definitions
- **Applied to annotations** – gives credit to curators for use of ontology terms in literature tagging

Functional Requirements for Bibliographic Records (FRBR)

- **Group 1 entities:** user interests in intellectual or artistic products
 - **Work:** a distinct intellectual or artistic creation
 - **Expression:** its intellectual or artistic realization
 - **Manifestation:** the physical embodiment of an expression of a work
 - **Item:** a single exemplar of a manifestation
- **Group 2 entities:** are responsible for content, production, ..., of group 1 entities
 - **Person:** an individual
 - **Corporate body:** an organization or group of individuals or organizations
- **Group 3 entities:** serve as the subjects of works
 - **Concept:** an abstract notion or idea
 - **Object:** a material thing
 - **Event:** an action or occurrence
 - **Place:** a location

FaBiO

- FRBR (Functional Requirements for Bibliographic Records) model from to OWL format.

[FaBiO \(FRBR-aligned Bibliographic Ontology\).](#)

- Paolo Ciccarese

<http://www.paolociccarese.info/>

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

Information Artifact Ontology

Terms ▾

Jump To:

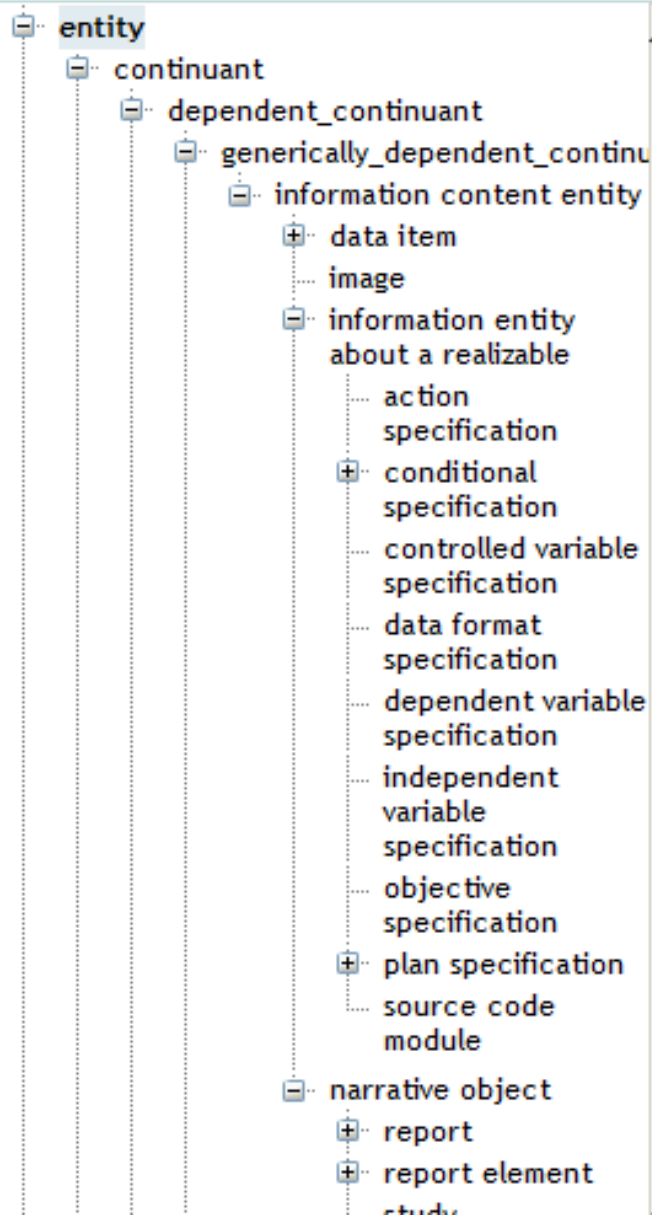
Details

Visualization

Notes (1)

Term Mappings (21)

Term R



Preferred Name

entity

ID

bfo:Entity

Full Id

<http://www.ifomis.org/bfo/1.1#Entity>

Equivalent Class

continuant or occurrent

Label

entity

<http://code.google.com/p/information-artifact-ontology/>

Semantically enhanced publishing

turn all highlighting on

date

disease

habitat

institution

organism

person

place

protein

taxon

[Top](#) | [Abstract](#) | [Author Summary](#) | [Introduction](#) | [Methods](#) | [Results](#) | [Discussion](#) | [Supporting Information](#) | [Acknowledgements](#) | [References](#) | [Data Fusion Supplements](#)

SEMANTICALLY ENHANCED VERSION OF A RESEARCH ARTICLE FROM PLOS NEGLECTED TROPICAL DISEASES

Impact of Environment and Social Gradient on *Leptospira* Infection in Urban Slur

document summary

Renato B. Reis ^{1#}, Guilherme S. Ribeiro ^{1#}, Ridalva D. M. Felzemburgh ¹, Francisco S. Santana ^{1, 2}, Sharif Mohr ¹, Astrid X. T. O. Melendez ¹, Adriano Queiroz ¹, Andréia C. Santos ¹, Romy R. Ravines ³, Wagner S. Tassinari ^{3, 4}, Marília S. Carvalho ³, Mitermayer G. Reis ¹, Albert I. Ko ^{1, 5 *}

¹ Centro de Pesquisas Gonçalo Moniz, Fundação Oswaldo Cruz, Ministério da Saúde, Salvador, Brazil ² Secretária Estadual de Saúde da Bahia, Salvador, Brazil ³ Escola Nacional da Saúde Pública, Fundação Oswaldo Cruz, Ministério da Saúde, Rio de Janeiro, Brazil ⁴ Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, Brazil ⁵ Division of International Medicine and Infectious Diseases, Weill Medical College of Cornell University, New York, New York, United States of America

Abstract

Background

Leptospirosis has become an urban health problem as slum settlements have expanded worldwide. Efforts to identify interventions for urban leptospirosis have been hampered by the lack of population-based information on *Leptospira* transmission determinants. The aim of the study was to estimate the prevalence of *Leptospira* infection and identify risk factors for infection in the urban slum setting.

Methods and Findings

We performed a community-based survey of 3,171 slum residents from Salvador, Brazil. *Leptospira* agglutinating antibodies were measured as a marker for prior infection. Poisson regression models evaluated the association between the presence of *Leptospira* antibodies and environmental attributes obtained from Geographical Information System surveys and indicators of socioeconomic status and exposures for individuals. Overall prevalence of *Leptospira* antibodies was 15.4% (95% confidence interval [CI], 14.0–16.8). Households of subjects with *Leptospira* antibodies clustered in squatter areas at the bottom of valleys. The risk of acquiring *Leptospira* antibodies was associated with household environmental factors such as residence in flood-risk regions with open sewers (prevalence ratio [PR] 1.43, 95% CI 1.14–1.75) and proximity to accumulated refuse (1.43, 1.04–1.98), sighting

With highlighting on

turn all highlighting off

date

disease

habitat

institution

organism

person

place

protein

taxon

[Top](#) | [Abstract](#) | [Author Summary](#) | [Introduction](#) | [Methods](#) | [Results](#) | [Discussion](#) | [Supporting Information](#) | [Acknowledgements](#) | [References](#) | [Data Fusion Supplements](#)

SEMANTICALLY ENHANCED VERSION OF A RESEARCH ARTICLE FROM PLOS NEGLECTED TROPICAL DISEASES

Impact of Environment and Social Gradient on *Leptospira* Infection in Urban Slum

document summary

Renato B. Reis ^{1#}, Guilherme S. Ribeiro ^{1#}, Ridalva D. M. Felzemburgh ¹, Francisco S. Santana ^{1, 2}, Sharif Mohr ¹, Astrid X. T. O. Melendez ¹, Adriano Queiroz ¹, Andréia C. Santos ¹, Romy R. Ravines ³, Wagner S. Tassinari ^{3, 4}, Marília S. Carvalho ³, Mitermayer G. Reis ¹, Albert I. Ko ^{1, 5*}

¹ Centro de Pesquisas Gonçalo Moniz, Fundação Oswaldo Cruz, Ministério da Saúde, Salvador, Brazil ² Secretária Estadual de Saúde da Bahia, Salvador, Brazil ³ Escola Nacional da Saúde Pública, Fundação Oswaldo Cruz, Ministério da Saúde, Rio de Janeiro, Brazil ⁴ Universidade Federal Rural do Rio de Janeiro, Rio de Janeiro, Brazil ⁵ Division of International Medicine and Infectious Diseases, Weill Medical College of Cornell University, New York, New York, United States of America

Abstract

Background

Leptospirosis has become an urban health problem as *slum settlements* have expanded worldwide. Efforts to identify interventions for urban *leptospirosis* have been hampered by the lack of population-based information on *Leptospira* transmission determinants. The aim of the study was to estimate the prevalence of *Leptospira* infection and identify risk factors for infection in the urban slum setting.

Methods and Findings

We performed a community-based survey of 3,171 slum residents from *Salvador, Brazil*. *Leptospira* agglutinating antibodies were measured as a marker for prior infection. Poisson regression models evaluated the association between the presence of *Leptospira* antibodies and environmental attributes obtained from Geographical Information System surveys and indicators of socioeconomic status and exposures for individuals. Overall prevalence of *Leptospira* antibodies was 15.4% (95% confidence interval [CI], 14.0–16.8). Households of subjects with *Leptospira* antibodies clustered in squatter areas at the bottom of *valleys*. The risk of acquiring *Leptospira* antibodies was associated with household environmental factors such as residence in flood-risk regions with *open sewers* (prevalence ratio [PR] 1.42, 95% CI 1.14–1.75) and proximity to *accumulated refuse* (1.43, 1.04–1.88), sighting *rats* (1.32, 1.10–1.58), and the presence of *chickens* (1.26, 1.05–1.51). Furthermore, low income and black race (1.25, 1.03–1.50) were independent risk factors. An increase of US\$1 per day in per capita household income was associated with an 11% (95% CI 5%–18%) decrease in infection risk.

Conclusions

turn all highlighting off

date

disease

habitat

institution

organism

person

place

protein

taxon

[Introduction](#) | [Methods](#) | [Results](#) | [Discussion](#) | [Supporting Information](#) | [Acknowledgements](#) | [References](#) | [Data Fusion Supplements](#)

ARTICLE FROM [PLOS NEGLECTED TROPICAL DISEASES](#)

Spatial Gradient on *Leptospira* Infection in Urban Slums

document summary

D. M. Felzemburgh ¹, Francisco S. Santana ^{1, 2}, Sharif Mohr ¹, Astrid X. T. O. Melendez ¹,
Cavines ³, Wagner S. Tassinari ^{3, 4}, Marília S. Carvalho ³, Mitermayer G. Reis ¹, Albert I. Ko ¹,

¹ Instituto de Saúde, Salvador, Brazil ² Secretária Estadual de Saúde da Bahia, Salvador, Brazil ³ Escola Nacional da Saúde Pública, Fundação
e Federal Rural do Rio de Janeiro, Rio de Janeiro, Brazil ⁵ Division of International Medicine and Infectious Diseases, Weill Medical
College

eagle-i and VIVO resource registry initiatives

eagle-i: Ontology for indexing and querying
biomedical research resources

<http://code.google.com/p/eagle-i/>

VIVO: An interdisciplinary national network
enabling collaboration and discovery
among scientists across all disciplines

<http://vivoweb.org/>

Shared ontology resources in OBO Foundry

Librarians will take over the world

shared VIVO and eagle-I ontologies inventorying

laboratories

services

instruments

reagents

organisms

images and videos

persons

protocols

patents

human studies

tissue samples

DNA sample

sample repositories

training opportunities

databases

papers

journals

expertise