Ontologies for e-Science Data

MELISSA HAENDEL, PH.D.
CHRIS SHAFFER, M.S. AHIP
OREGON HEALTH & SCIENCE UNIVERSITY LIBRARY
DECEMBER 9, 2011

OREGON HEALTH & SCIENCE UNIVERSITY
OHSU
OHSU Library

Outline

1. Why (as we all know) we need metadata
2. How ontologies can be used as a type of metadata
3. Science is messy
4. The Semantic Web
5. Ontologies for data inference
6. How libraries can be involved

The problem

Find the information:

OREGON HEALTH & SCIENCE UNIVERSITY
OHSU Library

Information retrieval from text-based resources is hard:

OMIM Query	# of records
“large bone”	785
"enlarged bone"	156
"big bones"	16
"huge bones"	4
"massive bones"	28
"hyperplastic bones"	12
"hyperplastic bone"	40
"bone hyperplasia"	134
"increased bone growth"	612

As librarians, you know that metadata standards are used in support of information retrieval

"Now! *That* should clear up a few things around here!"

Larson, October 1987

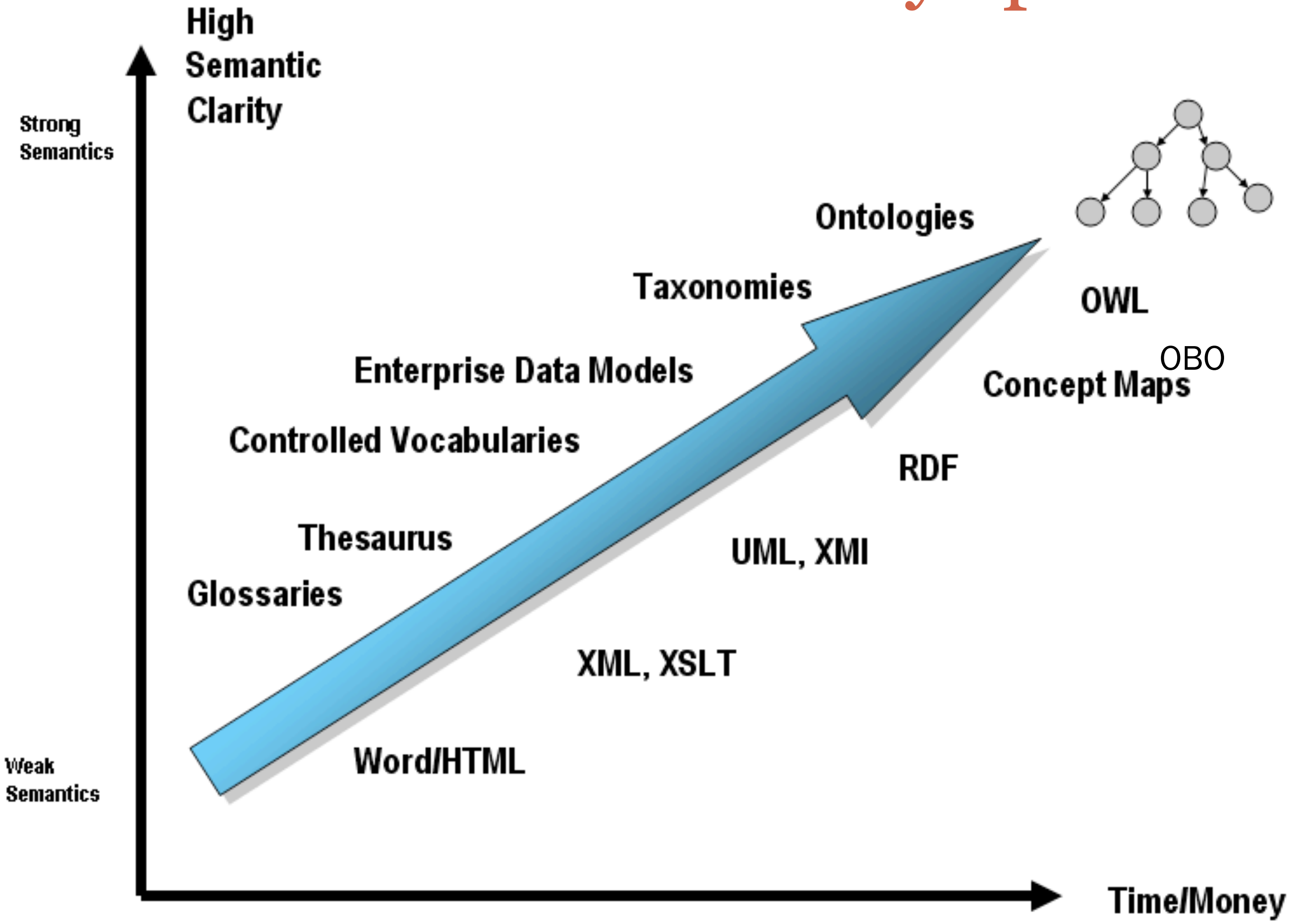
The use of an ontology to annotate data can *further* enhance retrieval, analysis and data sharing

What is an ontology?

Philosophers:
Ontology = The study of *being* as a branch of philosophy

Informaticists:
Domain ontology = representing a specific knowledge base

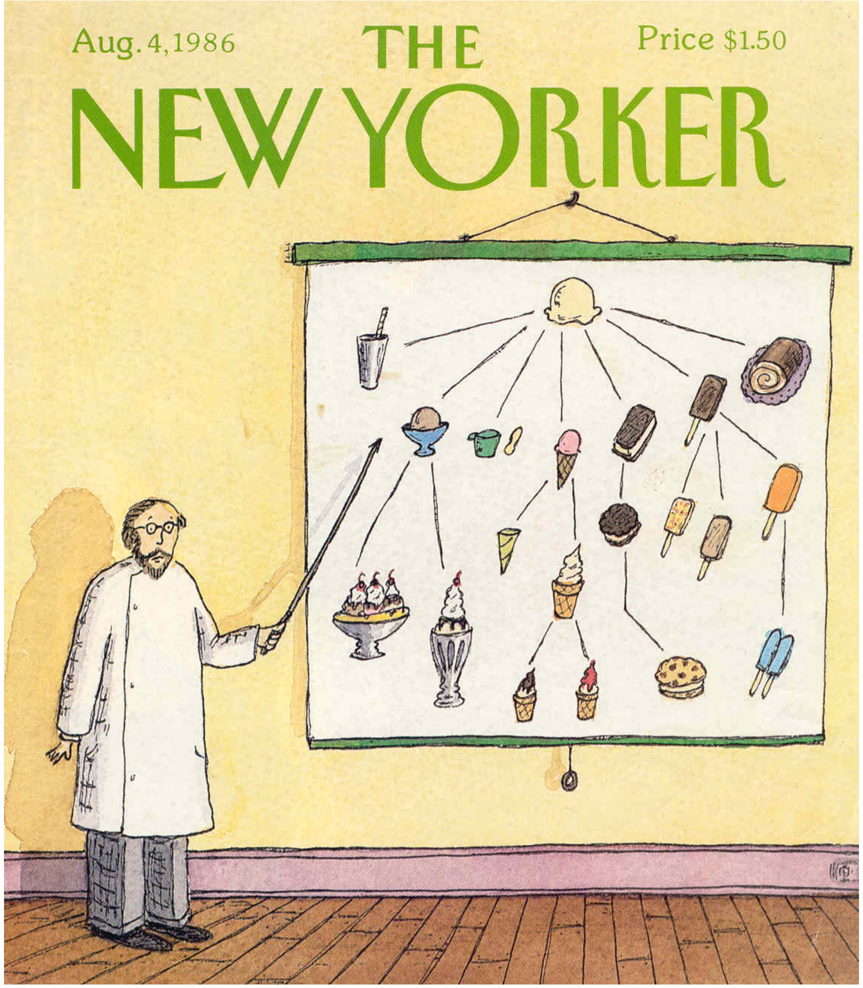
The controlled vocabulary spectrum



http://www.mkbergman.com/?m=20070516

* Reuse of ontologies can help reduce time/money. Libraries can help with this!

How does an ontology differ from other hierarchical vocabularies?



1. **Hierarchical terms are defined and annotations are made to the definitions**
2. **Relationships between the terms are also defined**
3. **Expressed in a language that can be reasoned across by computers**
4. **Data can easily be published as Linked Open Data**

Roz Chast, 1986

In order to understand the need for ontologies, we must first understand researcher behavior and needs

Research pre-Web:

Do an experiment

Document in a lab notebook

Publish your results

Research now:

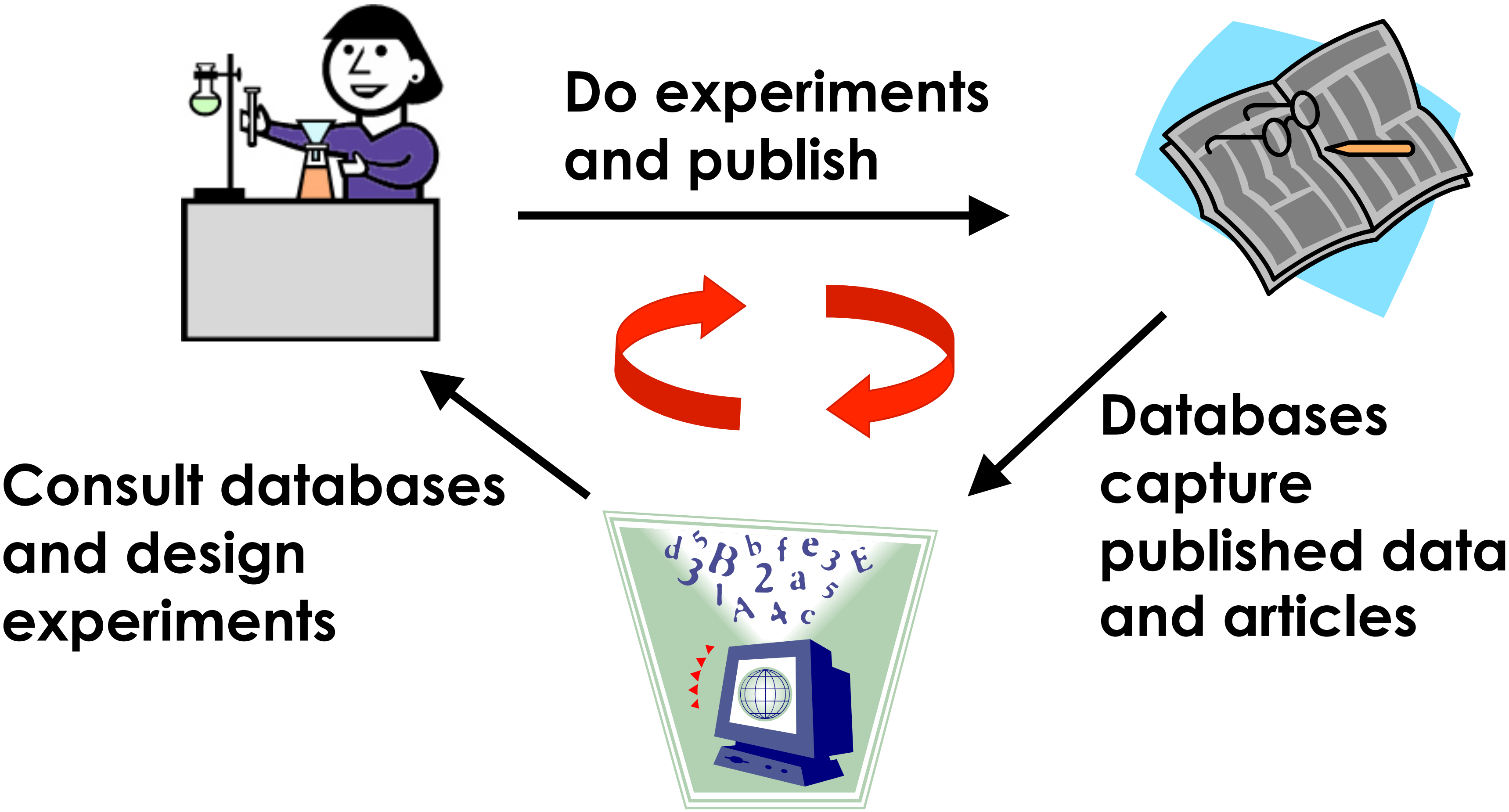
Do experiments and publish

Databases capture published data and articles

Consult databases and design experiments

Research databases save time and money

How do we facilitate this information cycle?



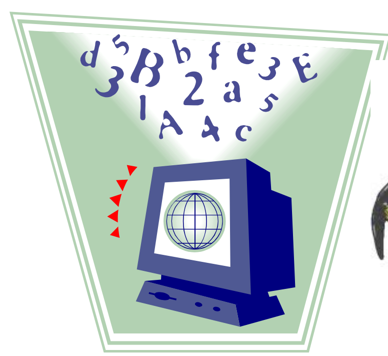
There are bottlenecks at all three steps ….

Text-mining

Do experiments and publish

Databases capture published data and articles

Consult databases and design experiments

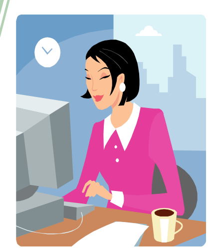


Text mining can be used to extract pertinent information into a database

Biocuration

Do experiments and publish

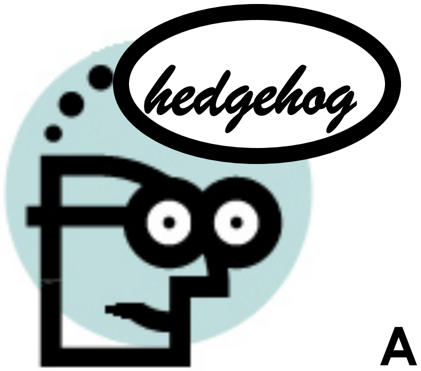
Databases capture published data and articles



Consult databases and design experiments

Biocurators are scientists who extract data from publications into a database

One problem:



Researchers use natural language when they publish

Author states: polyclonal anti-Mypt1
Santa Cruz Biotechnology

Which reagent did they mean?

Supplier lists:

PRODUCT NAME	CATALOG #	ISOTYPE	EPITOPE	APPLICATIONS	SPECIES
MYPT1 (E-19) Antibody	sc-17434	goat IgG	N-terminus (h)	WB, IP, IF, ELISA	m, r, h
MYPT1 (N-15) Antibody	sc-17433	goat IgG	N-terminus (h)	WB, IP, IF, ELISA	m, r, h
MYPT1 (H-130) Antibody	sc-25618	rabbit IgG	711-840 (h)	WB, IP, IF, IHC(P), ELISA	m, r, h
MYPT1 (K-18) Antibody	sc-34142	goat IgG	C-terminus (h)	WB, IF, ELISA	m, r, h
MYPT1/2 (C-18) Antibody	sc-34143	goat IgG	C-terminus (h)	WB, IF, ELISA	m, r, h
p-MYPT1 (Thr 853) Antibody	sc-17432	goat IgG	Thr 853 (h)	WB, IP, IF, ELISA	m, r, h
p-MYPT1 (Ser 695) Antibody	sc-33360	rabbit IgG	Ser 695 (h)	WB, IP, IF, ELISA	m, r, h
p-MYPT1 (Thr 696) Antibody	sc-17556	goat IgG	Thr 696 (h)	WB, IF, ELISA	m, r, h
p-MYPT1 (Ser 903) Antibody	sc-17557	goat IgG	Ser 903 (h)	WB, IF	m, r, h

Biocurators nor mining software can read minds

Lack of specificity results in databases missing data

ZFIN

Research | General Information | ZIRC

Home | Genes / Markers / Clones | BLAST | GBrowse | Expression | Antibodies | Mutants / Morphants / Tg | Anatomy | Publications | Maps

ZFIN ID: ZDB-GENE-980526-166

Gene Name: *sonic hedgehog a*
Gene Symbol: *shha*
Previous Names: shh, syu, vhh1(1), fc83d08, sonic you, sonic ... ehog homolog(1), wu:fc83d08(1)
Nomenclature History

Don't see your data here?

GENE EXPRESSION

All Expression Data:	254 figures from 195 publications
Wild-type Stages, Structures:	Blastula:30%-epiboly (4.66h-5.25h) to Adult (90d-730d, breeding adult) anterior neural keel, anterior neural rod, axial chorda mesoderm, axial mesoderm (all 85)
Curated Microarray Expression:	GEO (1)

MUTANTS AND TARGETED KNOCKDOWNS

Mutant lines:	32 genotypes
Alleles:	b240, shha_unspecified, t4, tbq70, tbx392 (all 6)
Knockdown reagents:	MO1-shha (1), MO2-shha (1), MO3-shha (1), MO4-shha (1)

PHENOTYPE

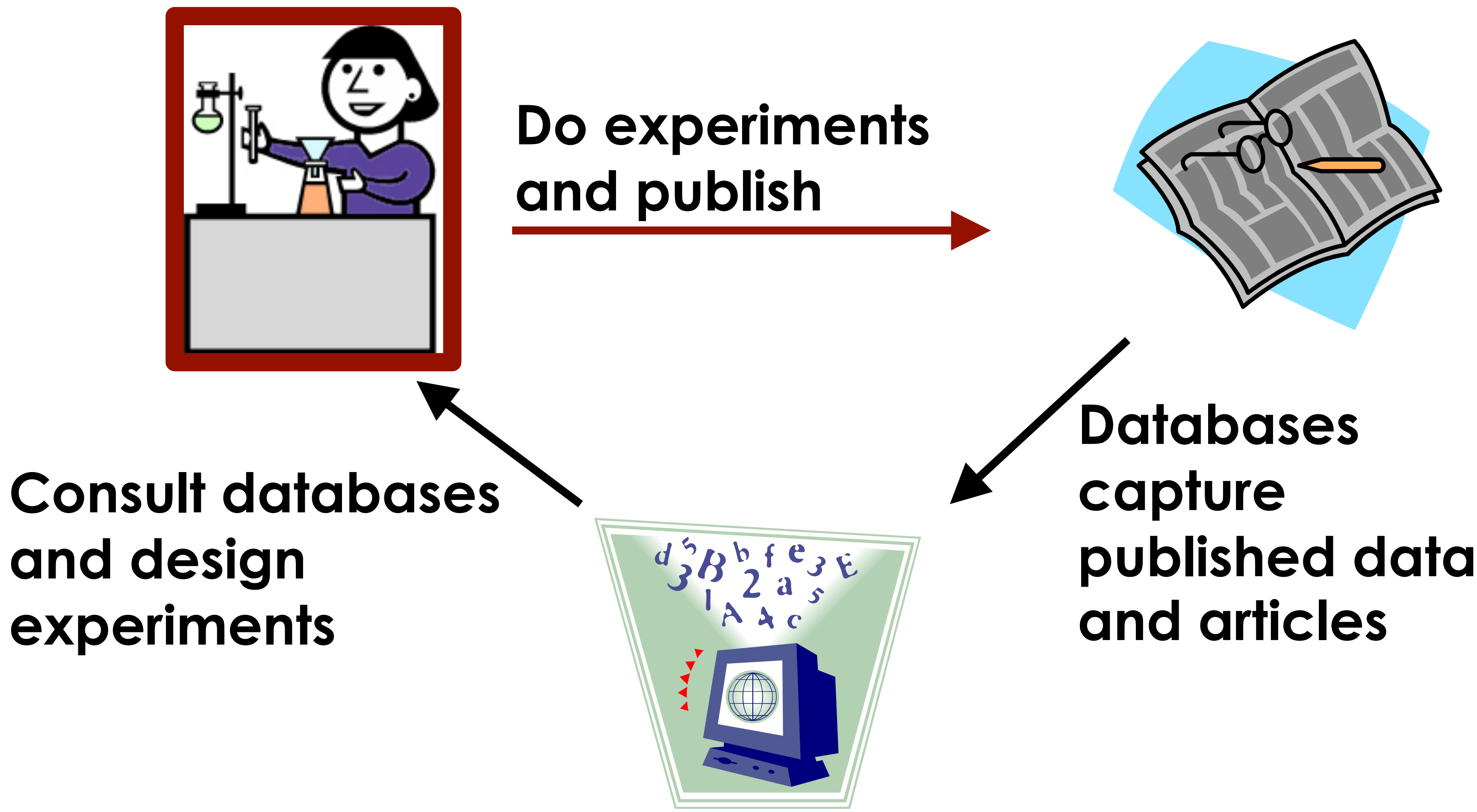
Data:	13 figures from 8 publications
Observed in:	adaxial cell, anatomical system, eye, fin (all 28)

GENE ONTOLOGY

Ontology	**GO Term**
Biological Process	adenohypophysis development (more)
Cellular Component	extracellular region (more)
Molecular Function	calcium ion binding (more)

GO Terms (all 65)

Science is messy



Science is messy

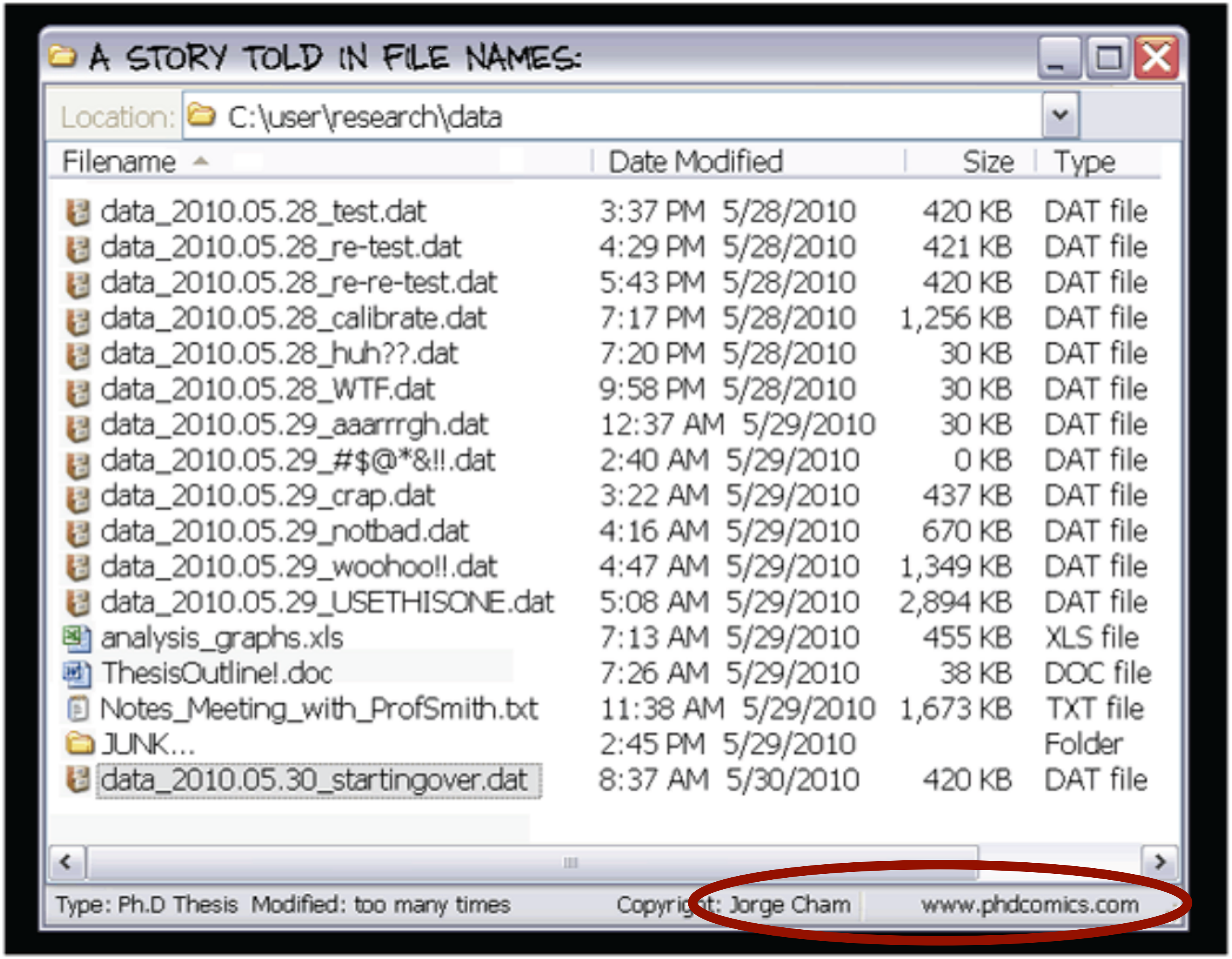
Researchers don't keep track of their activities or resources very consistently

A survey of 48 ecology programs revealed:

- **Over 75% did not require students to use lab notebooks**
- **Over 50% did not include data management-related instruction in the curriculum**

(Carly Strasser, 2011)

Today's lab notebook is often a collection of poorly organized files



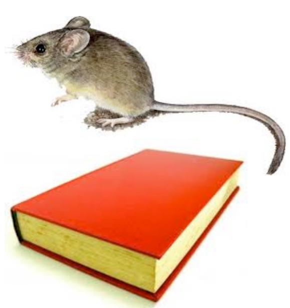
The eagle-i Consortium

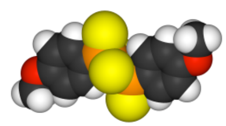
NIH funded pilot project to:

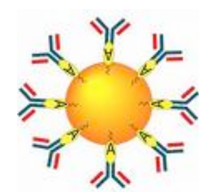
- Help researchers find scientific resources more easily
- Reduce time-consuming and expensive duplication of resources
- Provide meaningful semantic relationships between them using an ontology

Biologists went into labs to collect information about:

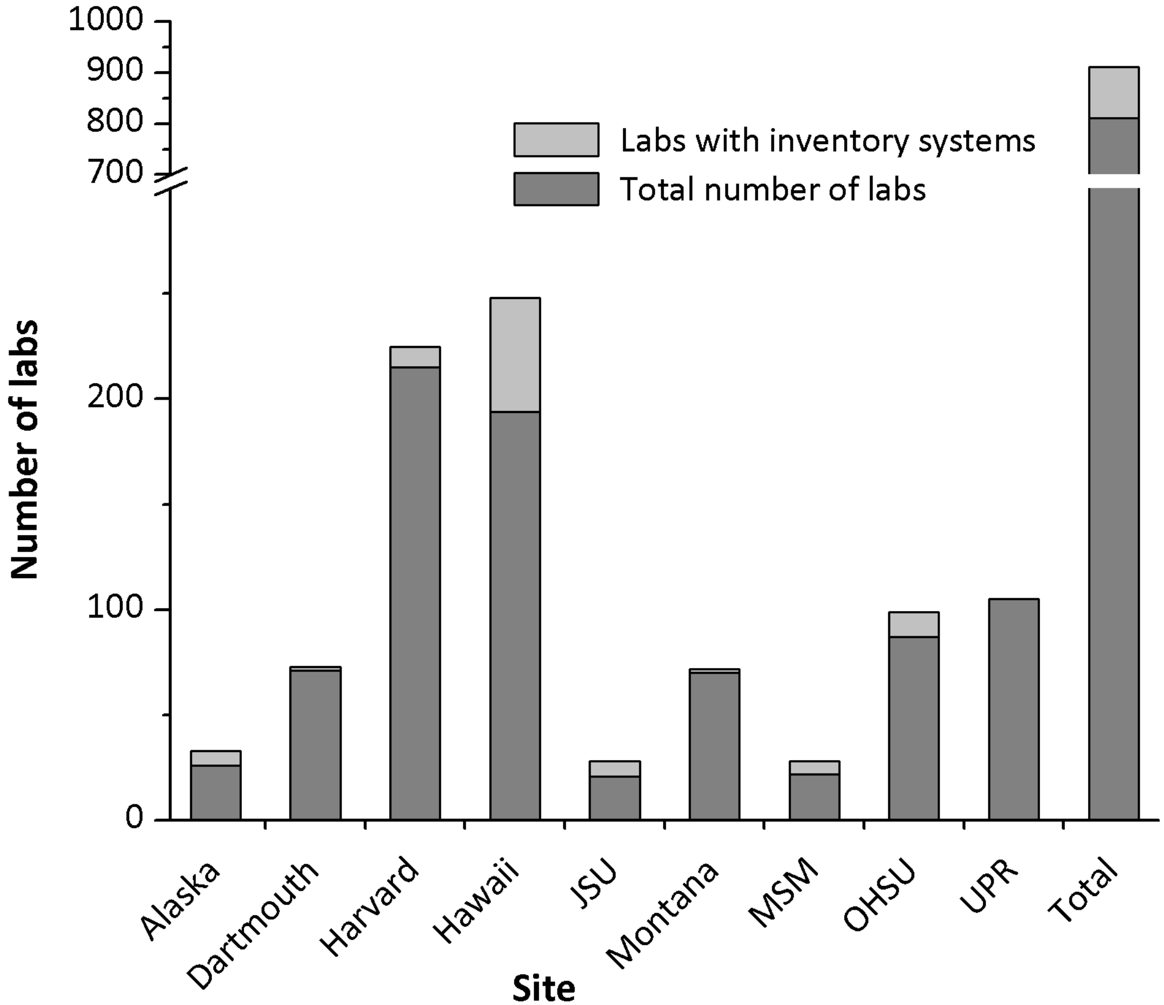
Reagents, protocols, services, instruments, expertise, organisms, training opportunities, software, human study metadata, biological specimens, etc.



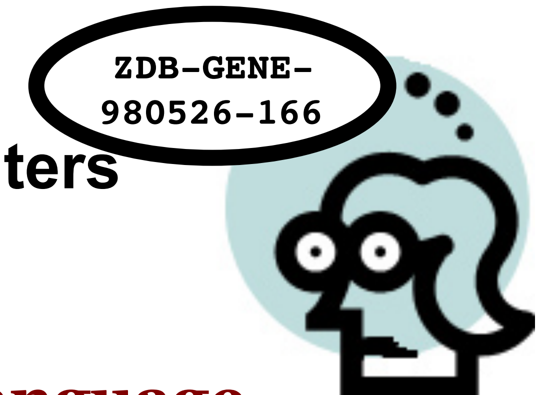
Today's labs are similarly disorganized



In an eagle-i survey of labs, 88% of labs had no inventory system of any kind

How do we create a culture of semantic scientists?

Authors, researchers



Databases, computers

We need to speak the same language

Literature/data specificity ↔ database quality and experimental reproducibility

Researchers need our help

Libraries are well-positioned to:

- **Facilitate semantic awareness**
- **Teach information management strategies**
- **Develop tools and ontologies**
- **Curate and publish semantically structured data**

Questions?

So….what about ontologies?

Ontologies enable organizing, filtering, connecting and suggesting data.

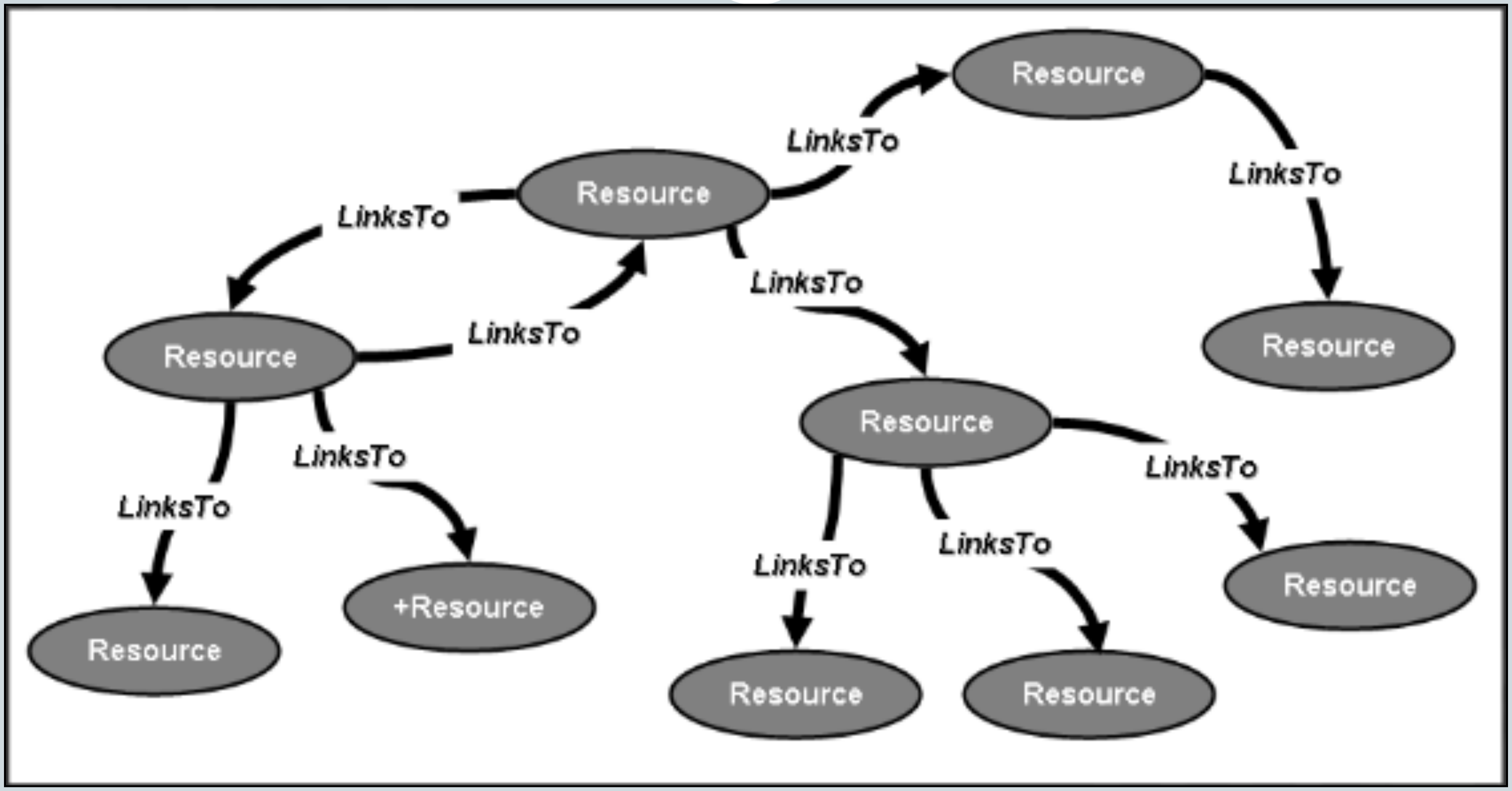
The Semantic Web is a way of sharing and reusing structured information.

Semantic Web Vision

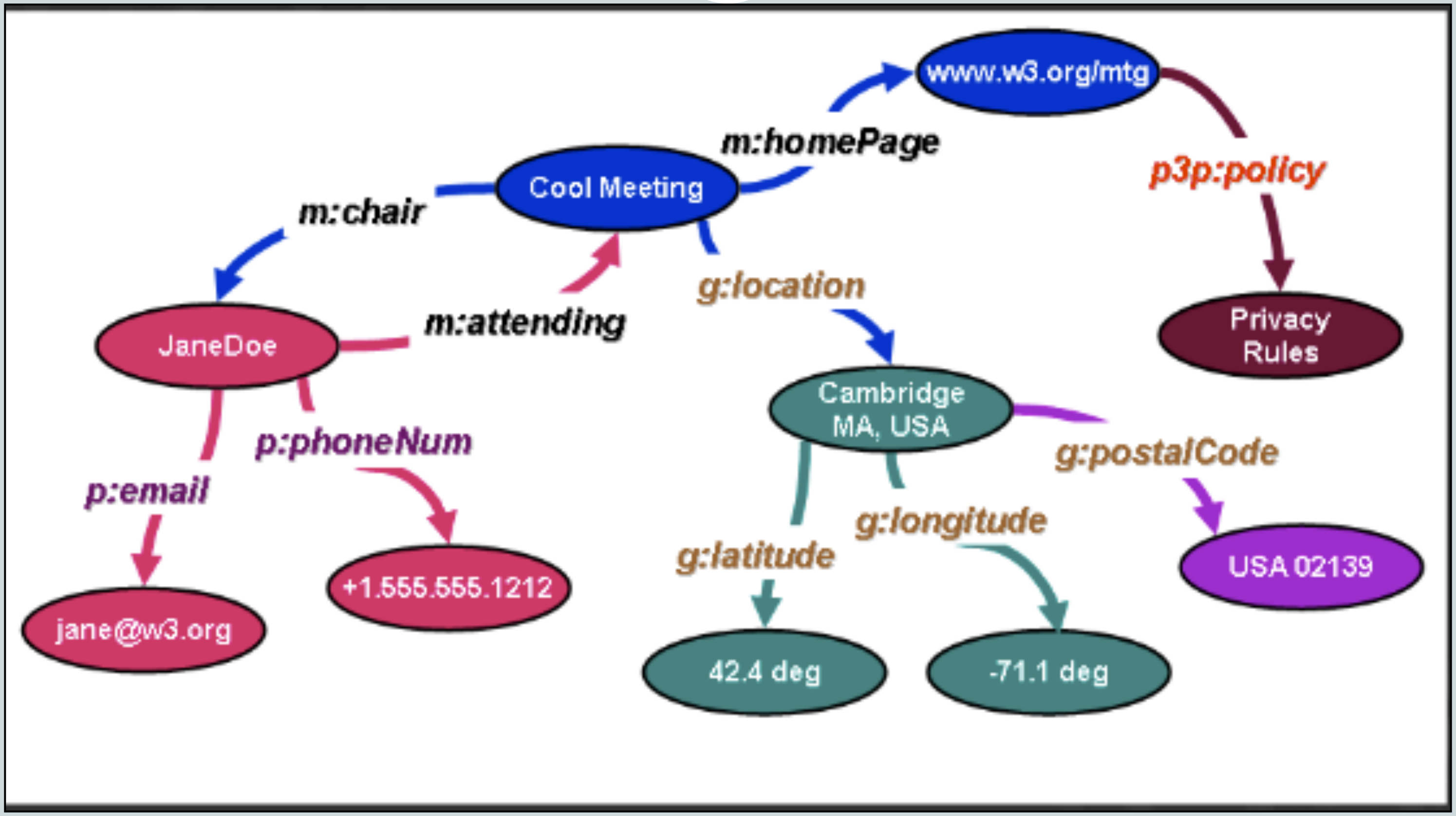
"The Semantic Web is an extension of the current web in which **information** is given well-defined **meaning**, better enabling computers and people to work in **cooperation**"

Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001

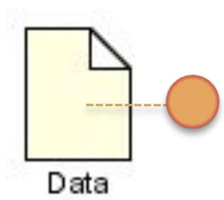
From web of documents….



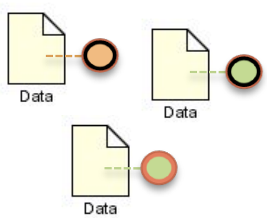
...to the web of things.



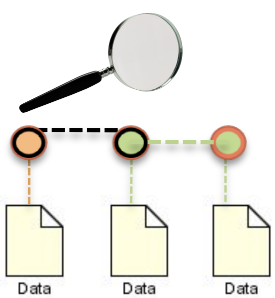
Using an ontology for annotation is similar to other metadata standards



- An ontology term is used as a tag on a piece of data similar to other metadata methods



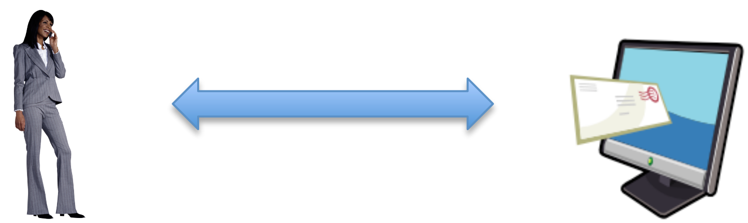
- The goal of the annotation is to add value by enabling:
  - Indexing data
  - Linking data



- Annotation of data using an ontology makes it easier to find and group data via semantic search

What can we do with ontologies that we can't do with simple metadata?

Ontologies are intelligible to both:

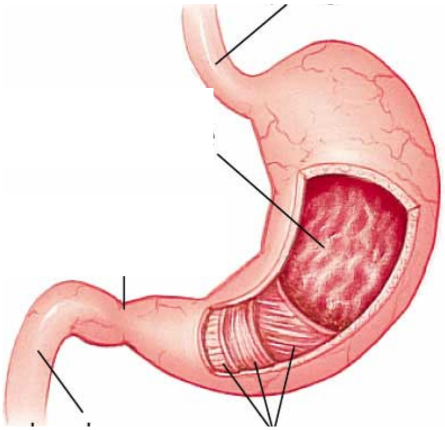


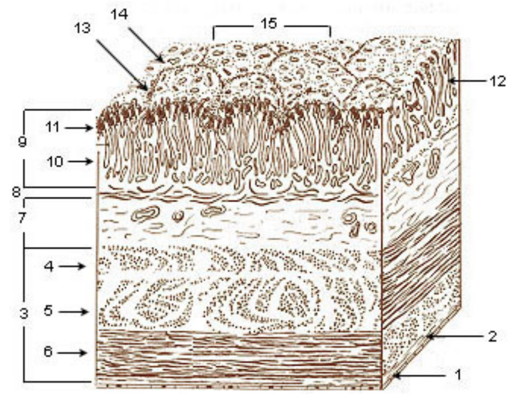
Humans **Machines**

Ontologies enable:

- **Automatic reasoning to infer related classes**
- **Annotation consistency**
- **Error checking**
- **Alignment with other ontologies**
- **Computation**

Common controlled vocabularies indicate the same meaning

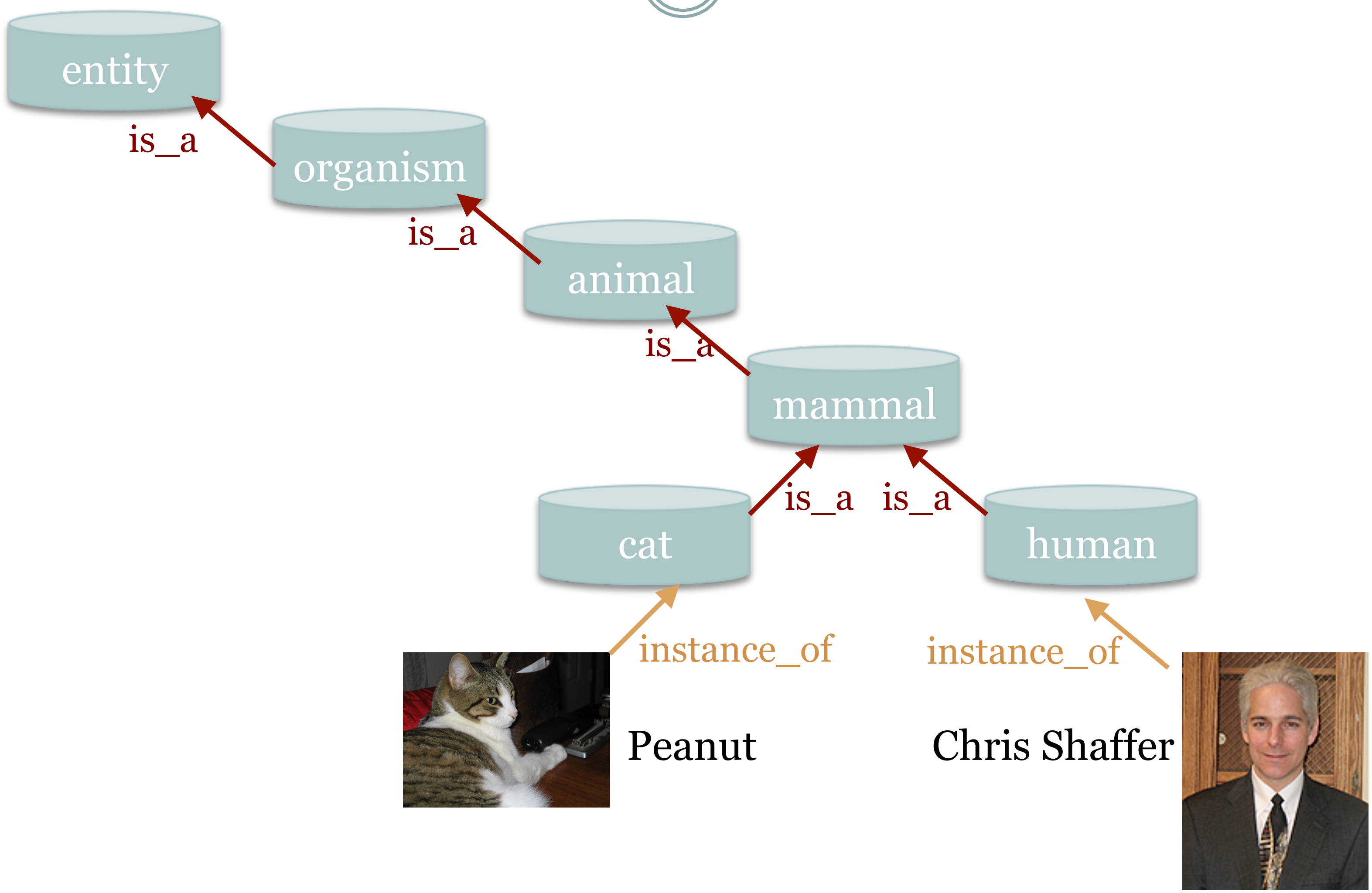




Is stomach defined by its gross morphology and location, or by the presence/absence of specific cell types?

- **Definitions lead to more consistent annotation**
- **Reusable classes make data interoperable**

A simple ontology example:
a machine can compute this



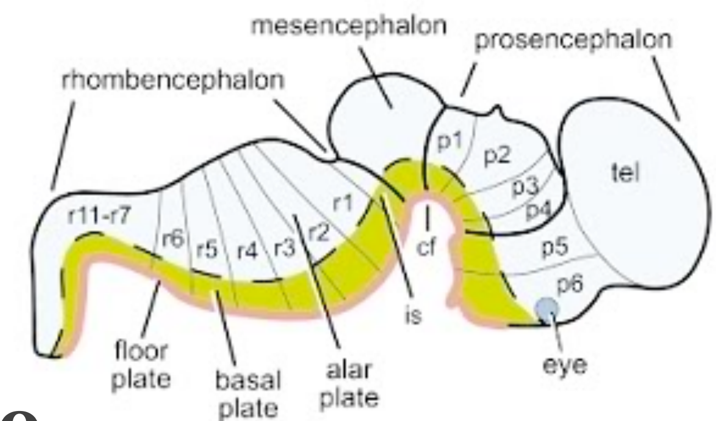
Searching using an ontology: A simple example

Number of genes annotated to each of the following brain parts in an ontology:

brain 20

part_of **hindbrain 15**

part_of **rhombomere 10**

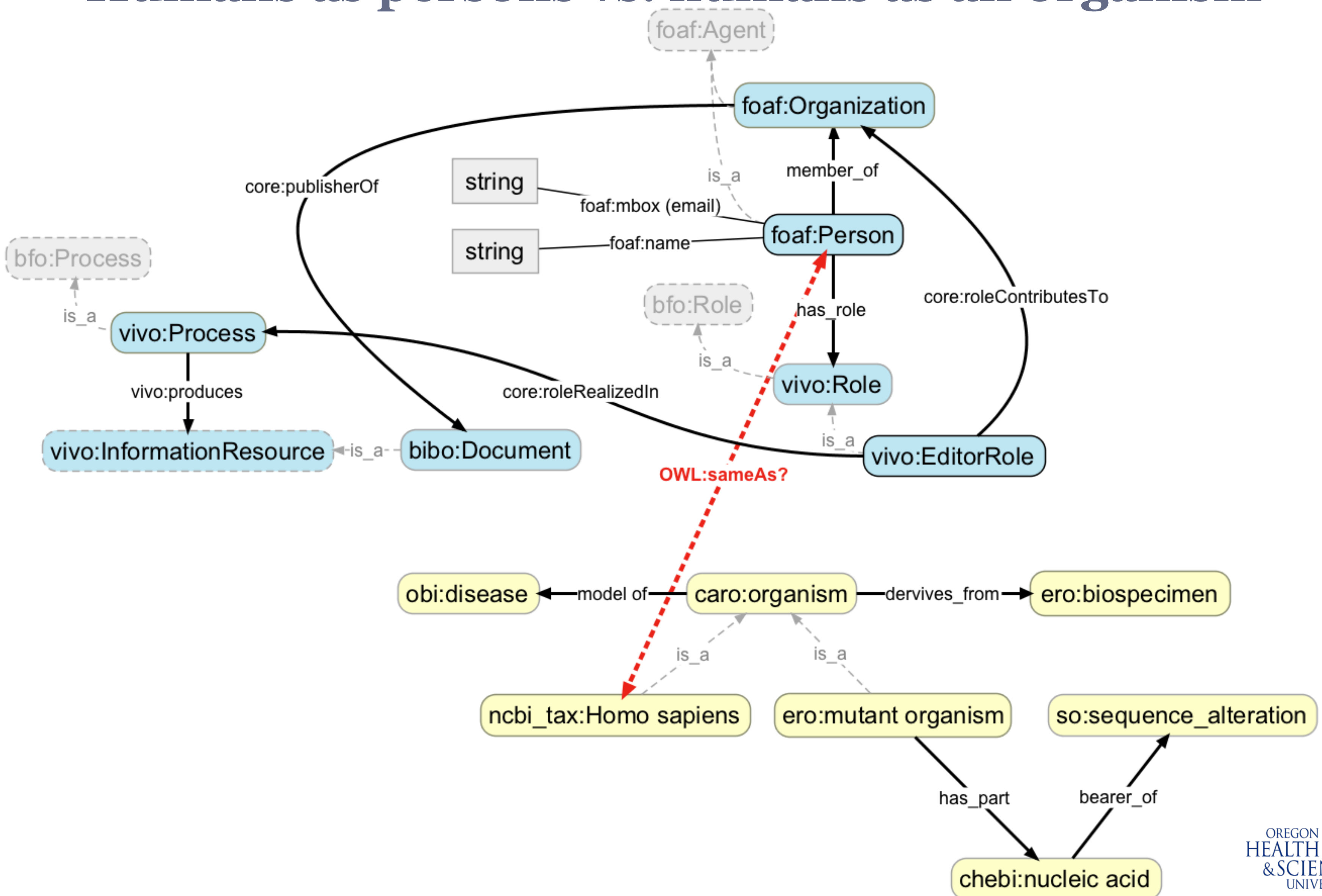


Query brain without ontology 20
Query brain with ontology 45

Ontologies facilitate grouping and retrieval of data

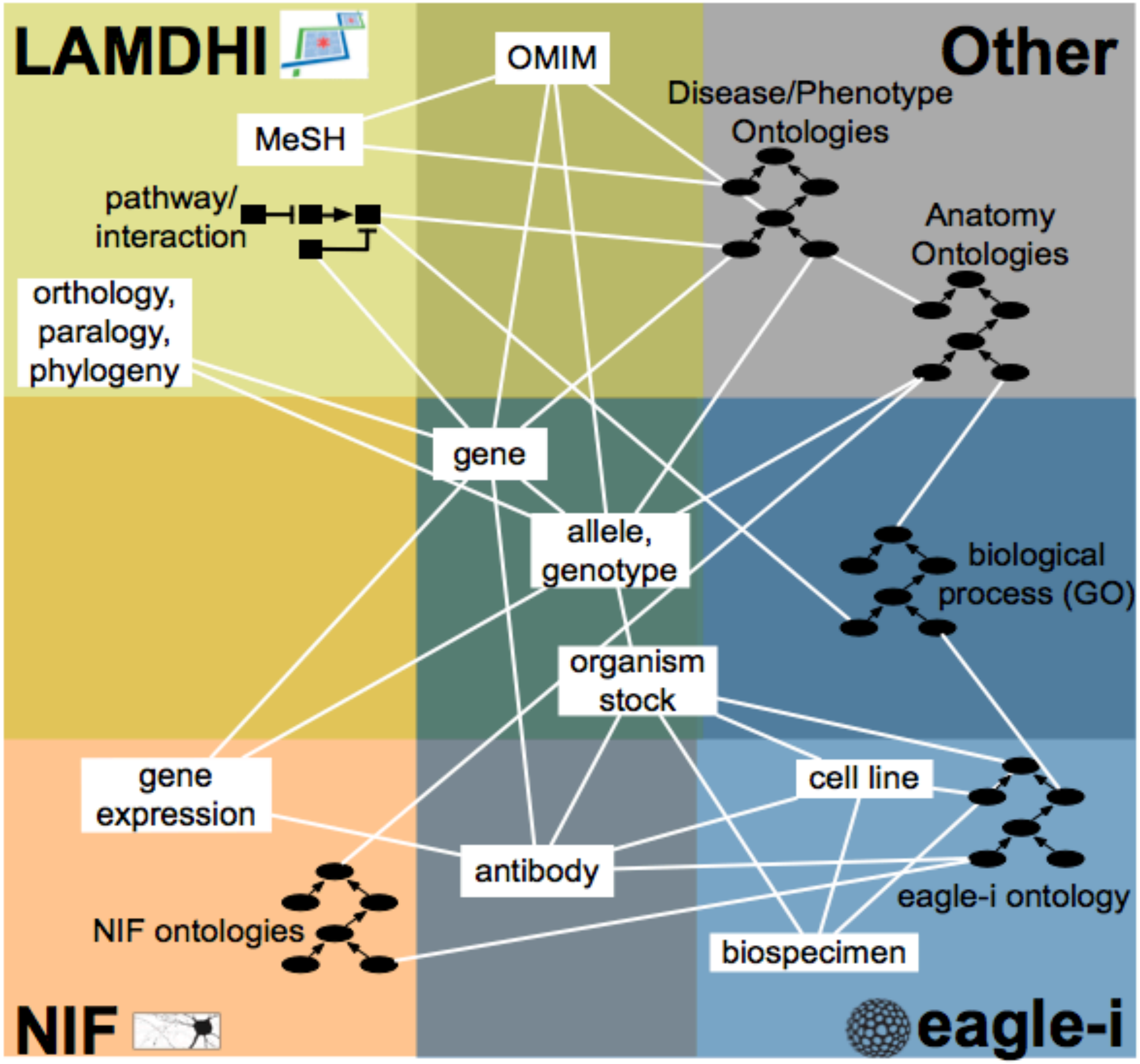
Ontology alignment issues and the need for common representation

Humans as persons vs. humans as an organism



OREGON HEALTH & SCIENCE UNIVERSITY
OHSU Library

There exist many types of relationships between entities of interest



OREGON HEALTH & SCIENCE UNIVERSITY
OHSU Library

We need global instances for common use

Establishing permanent URIs will be essential for achieving the goals of linked open data

Persons

ORCID Open Researcher & Contributor ID

isni

OCLC®

Organizations

NISO

D&B

ANSI American National Standards Institute

Geographic locations

Gazetteer (GAZ)

U.N. Geopolitical Ontology

FAO FIAT PANIS

GeoNames

Controlled vocabularies

Unified Medical Language System® (UMLS®)

AIMS Agriculture Information Management Standards AGROVOC

USDA United States Department of Agriculture National Agricultural Library

LIBRARY OF CONGRESS LC Subject Headings

Scientific inquiry is dependent on the resources at hand

This is what is in your kitchen, what are you going to make for dinner?

Scientific inquiry is dependent on the resources at hand

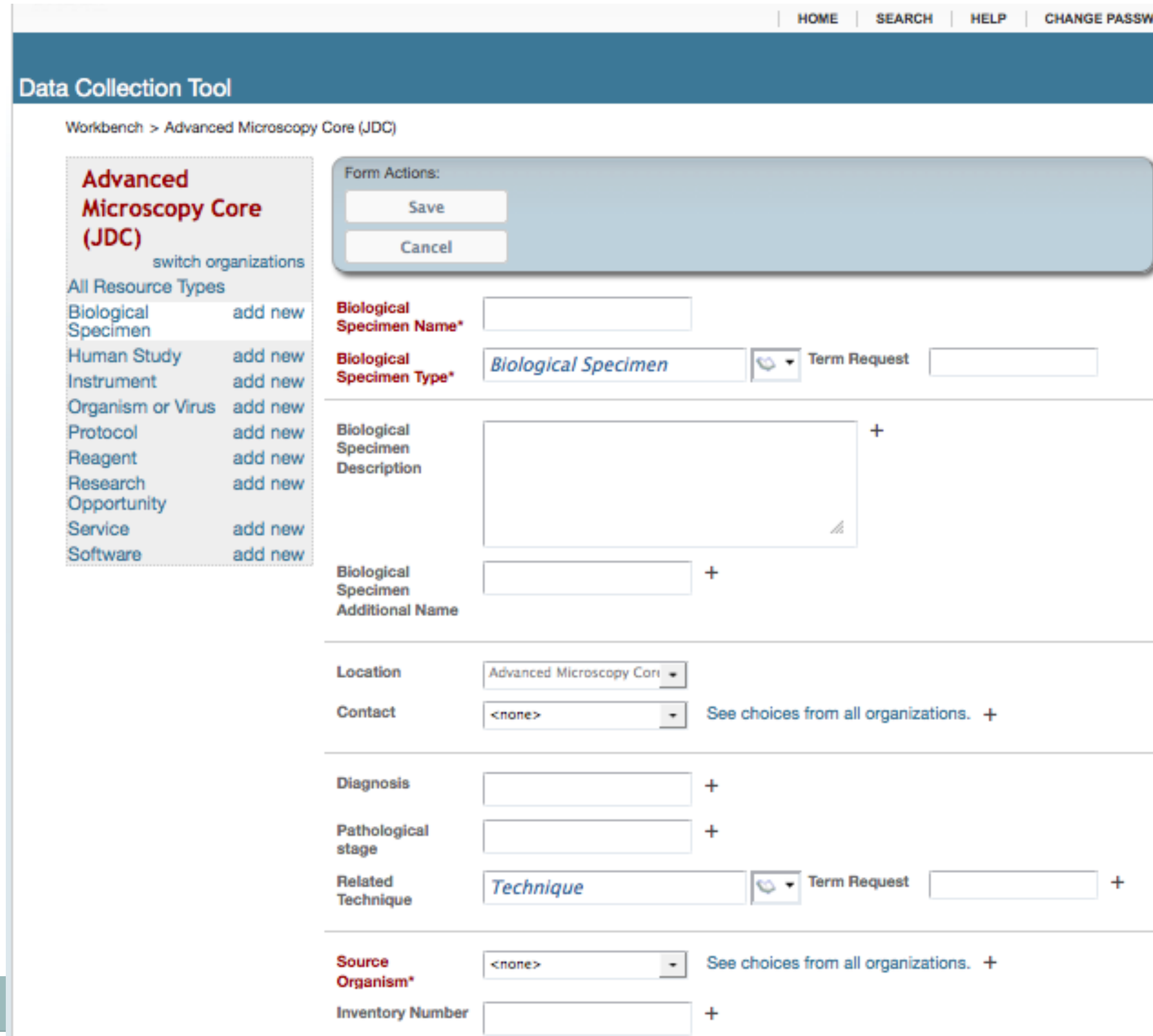
This is what is in your garden, what are you going to make for dinner now?

How do we get this wealth of data *to* researchers and how do we get this data *from* them?

- Scientists don’t often realize that providing the most basic annotation can be valuable for others to retrieve information

- Scientists have few incentives or tools to provide well annotated data

Example ontology driven application



OREGON HEALTH & SCIENCE UNIVERSITY
OHSU Library

Libraries can help scientists

- **How do we help researchers keep better track of their data?**
  - Online lab notebooks, lab inventory systems, data indexing, etc.

- **How can we improve the scholarly communication cycle to have more specific data?**
  - PDF markup tools, better journal requirements, etc.

Libraries can help:
- **design tools**
- **build ontologies**
- **promote semantically aware tools and interoperability**

Some examples of how the OHSU library is helping scientists

- **Post-traumatic Stress Disorder project to determine the effectiveness of different treatment strategies**
- **Clinical dental ontology to infer knowledge about long various kinds of restorations last**
- **Biospecimen representation to support identification of relevant biosamples**
- **Resource discovery**
- **Phenotype query across species to identify disease candidate genes**
- **Cell typing using existing ontologies to identify relevant biological processes based on gene expression**

Why libraries should care about ontologies

- Ontologies can be used to support scientific inference and new hypotheses

- Ontologies can be used to publish Linked Data in support of discovery and NIH/NSF-mandated data sharing

- Ontologies can be used to link disparate data in support of inference across them

Acknowledgements

OHSU

Nicole Vasilevsky
Erik Segerdell
Carlo Torniai
Matthew Brush
Scott Hoffmann
Jackie Wirz
Carla Pealer
Melanie Wilson
Shahim Essaid

Harvard

Daniela Bourges
Julie McMurry
Ted Bashor

Cornell

Jon Corson-Rikert
Stella Mitchell
Brian Lowe

LBNL

Christopher Mungall