# A biologists' perspective on ontology utility

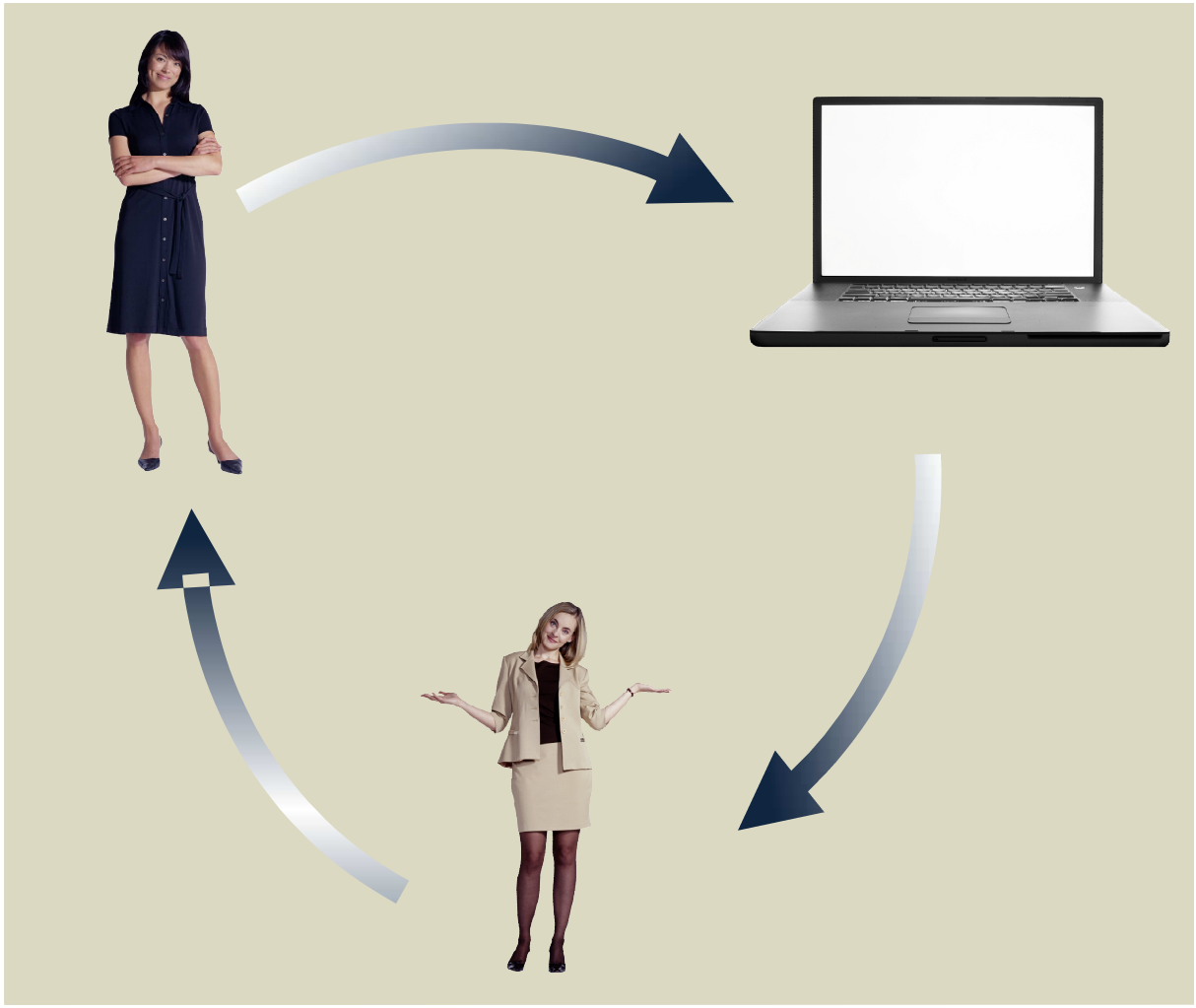
Melissa Haendel

Ontology Summit 2013

3.7.13

# What do biologists need from an ontology?

## Need it to communicate with each other and computers



# What do biologists need from an ontology?

## Need to apply ontologies to data:

- Annotation
- Structuring
- Mapping

IN A CONSISTENT MANNER

| Name | Type | Strain |
|---|---|---|
| CD4 knockout | Mouse | C57BL/6 |
| OTI | Mouse | C57BL/6 |
| CD4 knockout | Rat | Wistar |

## Search and classify data:

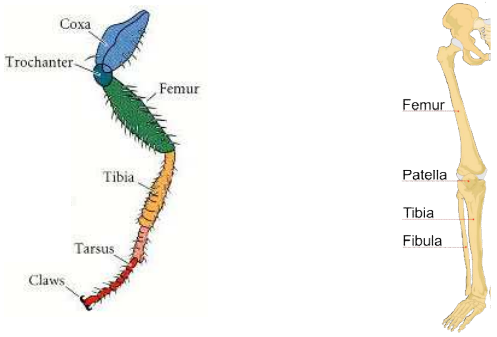
Query:
Find organisms on C57BL/6 background

| Name | Type | Strain |
|---|---|---|
| CD4 knockout | Mouse | C57BL/6 |
| OTI | Mouse | C57BL/6 |

# Biologists have issues

Biology is particularly challenging due to moving conceptual targets (e.g. gene, allele)

Biologists like to re-use the same labels left and right







**Biologists often don't agree on how to define key concepts!**

# How to achieve these goals and address these issues?

# Quality textual documentation and definitions

Seems simple . . .

# Quality textual documentation and definitions

- Definitions must not only describe a concept, but allow determination of real instances
- Requires consideration of specific and precise concepts to be defined
- This is where listing out properties and their use for definition are very helpful

**=> Need a 'concept first approach' to definition creation**

# A ‘Concept-First Approach’ Example

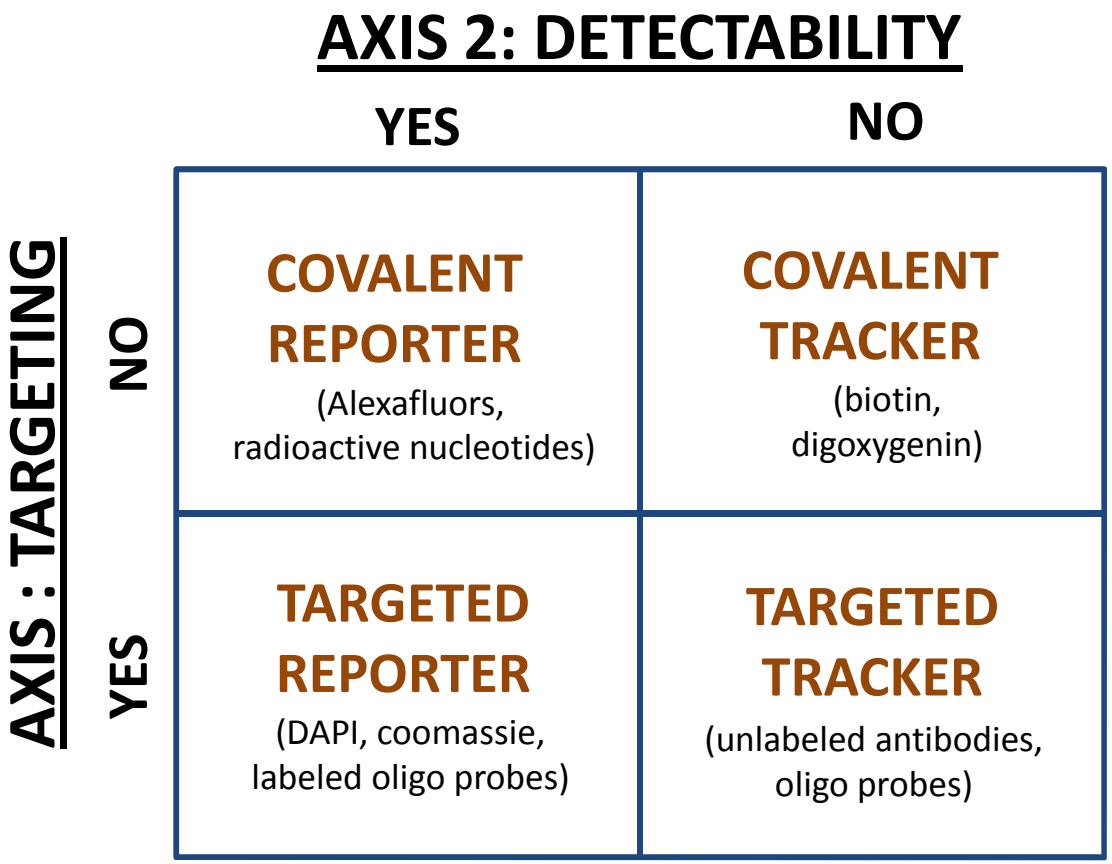
- Terminology of **molecular labels** is inconsistent: ‘probe’, ‘tracer’, ‘detector’, ‘reporter’, used variably to describe reagents with different characteristics
- Conformance to varied label conventions has led to confusion, ambiguity, and inconsistency in different ontologies
- Throw out the labels and consider the biologically important axes:

**Axis 1 = Targeting:** ability to specifically associate with a molecular target

**Axis 2 = Detectability:** ability to emit or produce some detectable signal

# A ‘Concept-First Approach’ Example

Axes applied to yield a set of principled subclasses, and a more descriptive labeling scheme is applied




Results vetted by members of different stakeholder communities

# Label Bias

- A common source of problems is developers imposing assumptions about the semantic content of a label.

- Best to avoid using terms with varied and ambiguous meaning as primary labels . . .
    - e.g. ‘molecular probe’, ‘cell line’,

  . . . or be careful to document this ambiguity, and give a precise definition that is clear about its specific view

- OBO Foundry principle suggests use of numeric URIs which helps avoid label bias.
    - e.g. URIs not like http://xyzweb.org/ont/core#Position
    - But instead like: http://xyzweb.org/ont/core_43888887

# Evaluating text definitions

- Are all classes and properties defined? Do they avoid use of figurative or obscure language? Are there citations?

- Are essential features for distinguishing from other classes included?
  - Aristotelian/genus-differentia structure is good for this
  - Especially important for imported classes

- Is circularity avoided?
  - e.g. ‘specimen collection objective’ defined as the ‘objective to collect a specimen’

- Are there clear instances in reality? Test against a set of candidate instances

# Testing Definitions Against Instances

Example: ‘genetic material’

- complex and nuanced concept, subject to varied interpretations
- precise definition needed for use in diverse research communities (genomics, experimental biology, model organisms)
- many definitions found on web are correct but vague and insufficient (not precise enough to delineate instances in reality)
- e.g. *“material of plant, animal, microbial or other origin containing functional units of heredity”*

# Testing Definitions Against Instances

- A more precise definition

"a nucleic acid macromolecule that is part of a cell or virion and has the disposition to be replicated and inherited by descendants."

- Start testing against instances
  1. chromosomal DNA in dividing cells
  2. chromosomal DNA in post-mitotic cells

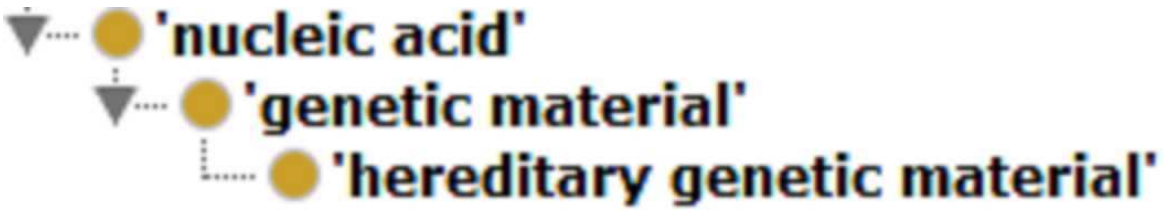
- Problem encountered already: a disposition for replication is not relevant in post-mitotic cells

- Solution: Refine Definition

"a nucleic acid macromolecule that is part of a cell or virion and is inherited from an immediate ancestor, or incorporated in a manner that it has the disposition to be replicated and inherited by descendants."

# Testing Definitions Against Instances

- Continue testing against instances
  3. a gene targeting DNA construct transfected into a cell
  4. a transiently transfected DNA expression construct
  5. a microinjected siRNA oligo in a cell
- Problem: some instances expected to be grouped with genetic material, but don't meet our strict definition
- Solution:
  1. Rename original class with more descriptive label '<u>hereditary genetic material</u>', and
  2. Implement new '<u>genetic material</u>' as a more inclusive parent class



# Other Textual Documentation

- Does the ontology include extensive synonyms?
- Does the ontology include other internal documentation such as edit history? Creator/editor? Status? Modeling notes?
- Are there examples of usage? Counter examples?
  - Sometimes a concept is so complex that a definition cannot disambiguate exclusion or inclusion.

# Example of a difficult to define entity

**Reagent:**

a material entity that bears a reagent role by virtue of it being intended for application in a scientific technique to participate in (or have molecular parts that participate in) a chemical reaction that facilitates the generation of data about some distinct entity, or the generation of some distinct material specified output

A ph meter probe fits our best definition, but is not considered a reagent by biologists =>> document!

# Is there complete metadata?

- Can use scripts to check ontology for a number of important metadata bits
  - missing definitions, labels, source, example of usage
  - duplicate labels
  - mismatched synonyms
- Inclusion of some sort of curation status or metadata complete annotation to indicate readiness of entities

# 2. Quality logical expressions

- No true path violations
  - Common mistakes are to forget that the distant ancestors definitions apply, and misunderstandings about transitive properties. These are not always caught by reasoners.
- A balance of realism and practicality
  - Are the concepts as expressed in the text definition aligned with the logical expressions?
  - Are there many concepts used in logical expressions for which there would be no instances collected?
  - Does the inferred classification makes sense biologically?

# 3. Ontology reuse

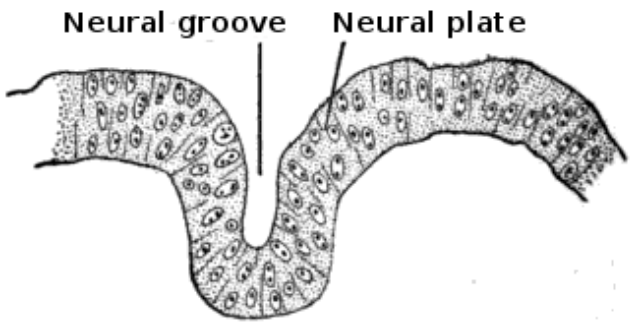
- Is the ontology orthogonal to others (an OBO priniciple)?
- Does the ontology reuse entities from other ontologies at key intersection points?
- Does the ontology include a subset of external entities with the closure, or import entire external ontologies?

# Testing the ontology against data

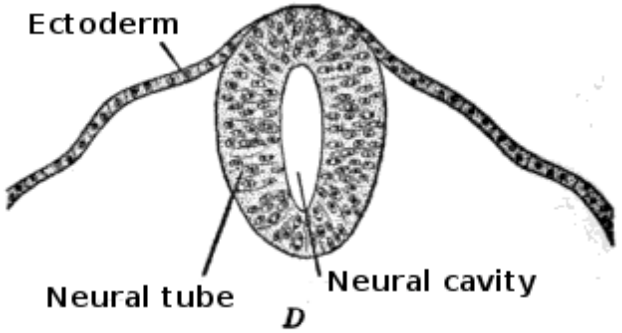
- It is important to validate the ontology against real data. But this is the intrinsic talk, you might say…..
- Data often tells you that your ontology has errors.

# Iterative data-ontology evaluation

Ontology classes: Neural plate




Neural tube




Ontology stage class definitions:

5-9 somites->10-13 somites

10-13 somites->prim5

Neural tube

Instance data: 9 somites

**Neural plate or neural tube?**

# Ongoing intrinsic issues for development of biological ontologies

- Lacking tools to enable visualization/editing via other transitive relations
- Better ability to synchronize pieces of ontologies reused in another ontology
- Standards for documentation, annotation properties and layering/import approaches for documenting in ontologies themselves, mechanisms for linking out to trackers/lists, wikis
- Better ways to roll together text definitions from dependent classes

# Special Thanks

Matt Brush
Scott Hoffman

Pioneers in our group to make all ontologies usable by more people.