Principles for Building Biomedical Ontologies

ISMB 2005

May 18, 2005

Introductions

- Suzanna Lewis:
  - Head of the BDGP bioinformatics group and a founder of the GO
- Barry Smith:
  - Research Director of the ECOR
- Michael Ashburner:
  - Professor of Genetics at the University of Cambridge; Founder and PI of FlyBase; and Founder and PI of the GO
- Mark Musen:
  - Head of Stanford Medical Informatics
- Rama Balakrishnan:
  - Scientific Content Editor at the SGD and for the GO
- David Hill:
  - Scientific Content Editor at the MGI and for the GO

Special thanks to

- Christopher J. Mungall
- Winston Hide

Outline for the Morning

- A definition of “*ontology*”
- Four sessions:
  - Organizational Management
  - Principles for Ontology Construction
  - Case Studies from the GO
  - Summation

Ontology (as a branch of philosophy)

- *The science of what is: of the kinds and structures of the objects, and their properties and relations in every area of reality.*
- In simple terms, it seeks the classification of entities.
- Defined by a scientific field's vocabulary and by the canonical formulations of its theories.
- Seeks to solve problems which arise in these domains.

In computer science, there is an information handling problem

- Different groups of data-gatherers develop their own idiosyncratic terms and concepts in terms of which they represent information.
- To put this information together, methods must be found to resolve terminological and conceptual incompatibilities.
- Again, and again, and again…

The Solution to this Tower of Babel problem

- A shared, common, backbone taxonomy of relevant entities, and the relationships between them, within an application domain
- This is referred to by information scientists as an *’Ontology’*.

Which means…

Instances are not included!

- It is the generalizations that are important
- Please keep this in mind, it is a crucial to understanding the tutorial

Motivation: to capture biology.

- Inferences and decisions we make are based upon what we know of the biological reality.
- An ontology is a computable representation of this underlying biological reality.
- Enables a computer to reason over the data in (some of) the ways that we do.

Principles for Building Biomedical Ontologies

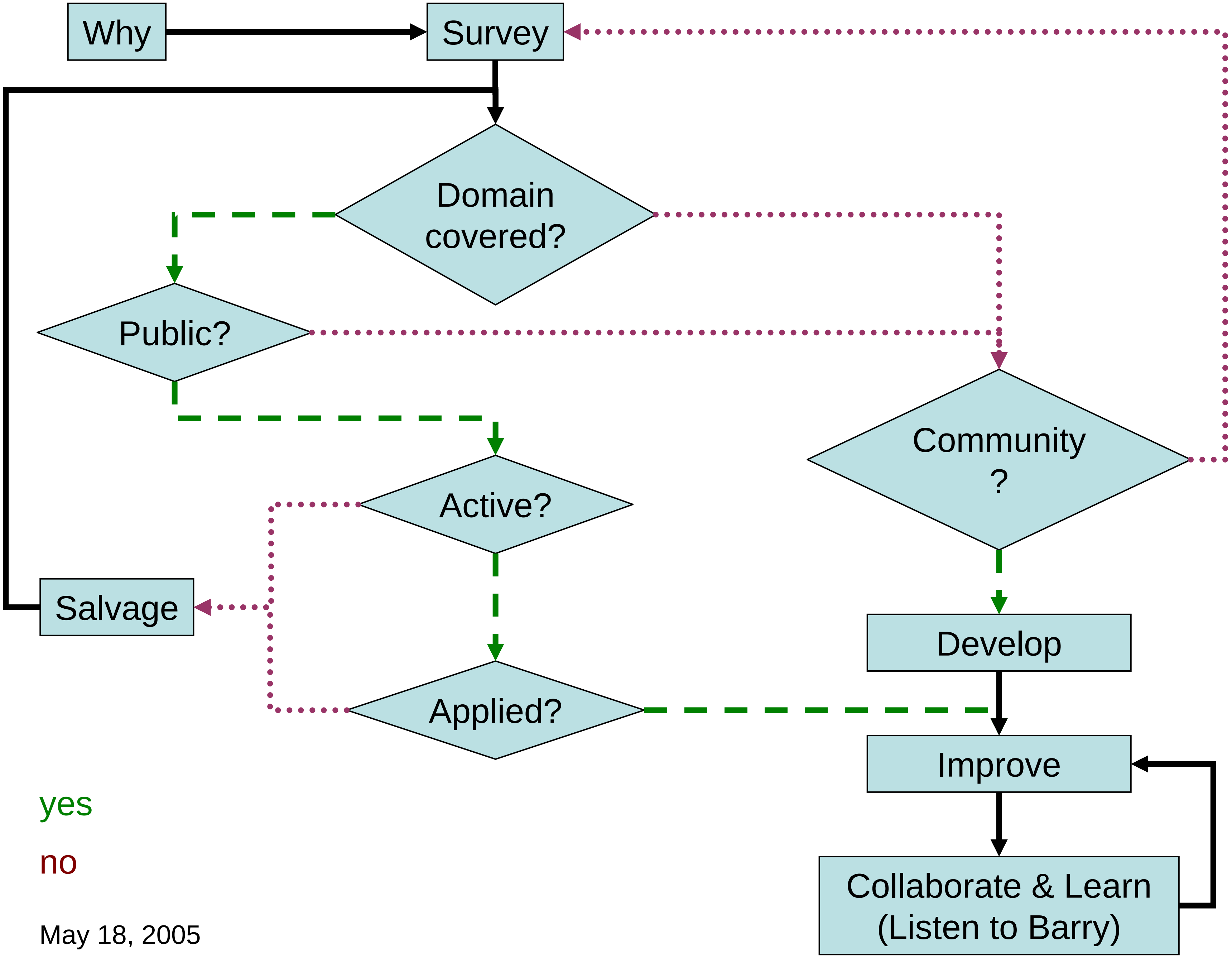
Michael Ashburner and Suzanna Lewis

http://obo.sourceforge.net

May 18, 2005

You need (want) an ontology

- What do you do?
- Where do you turn?
- Who are you going to call?



Why

Survey

Domain covered?

Public?

Active?

Applied?

Salvage

Community ?

Develop

Improve

Collaborate & Learn (Listen to Barry)

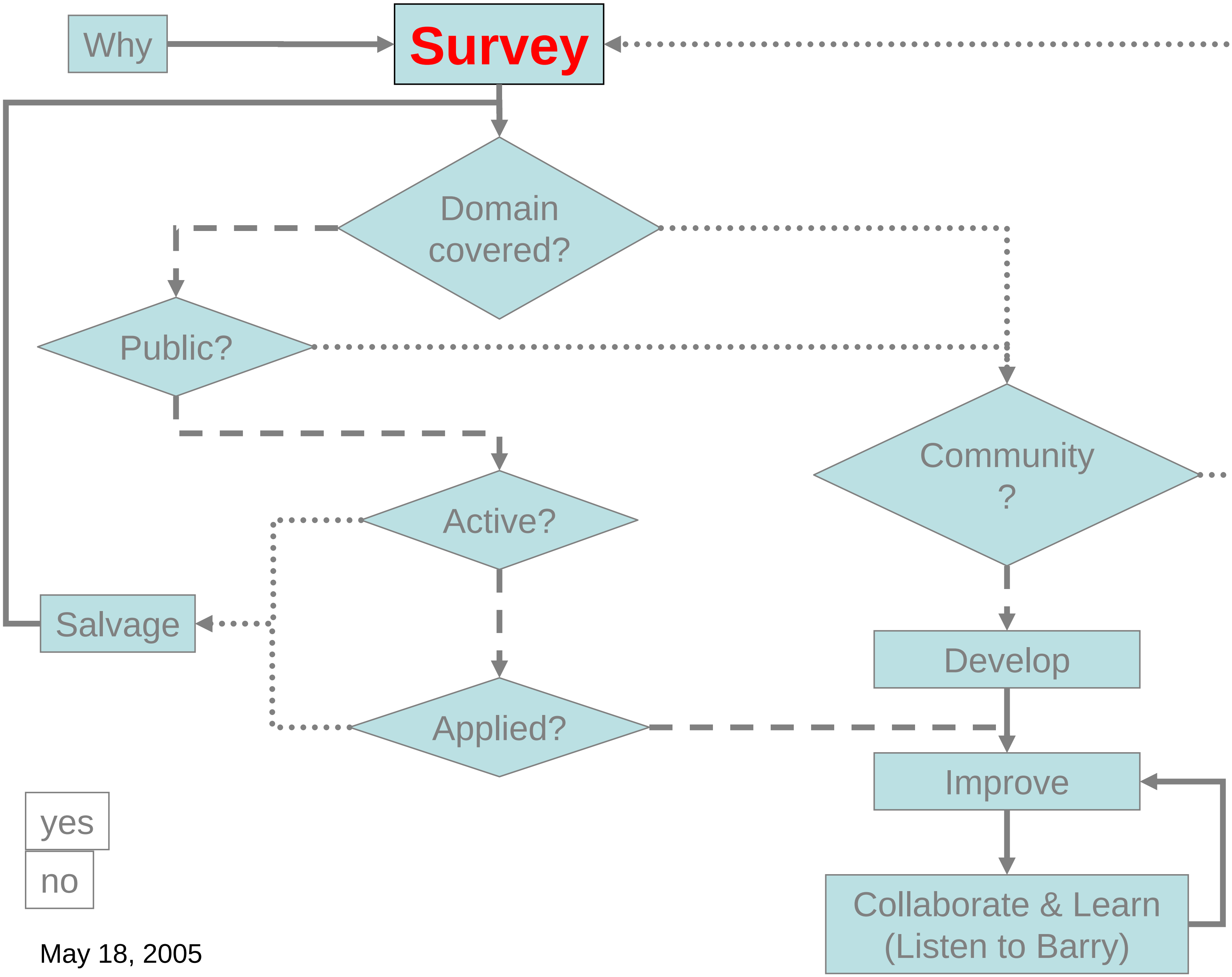
yes

no

May 18, 2005

Evaluating ontologies

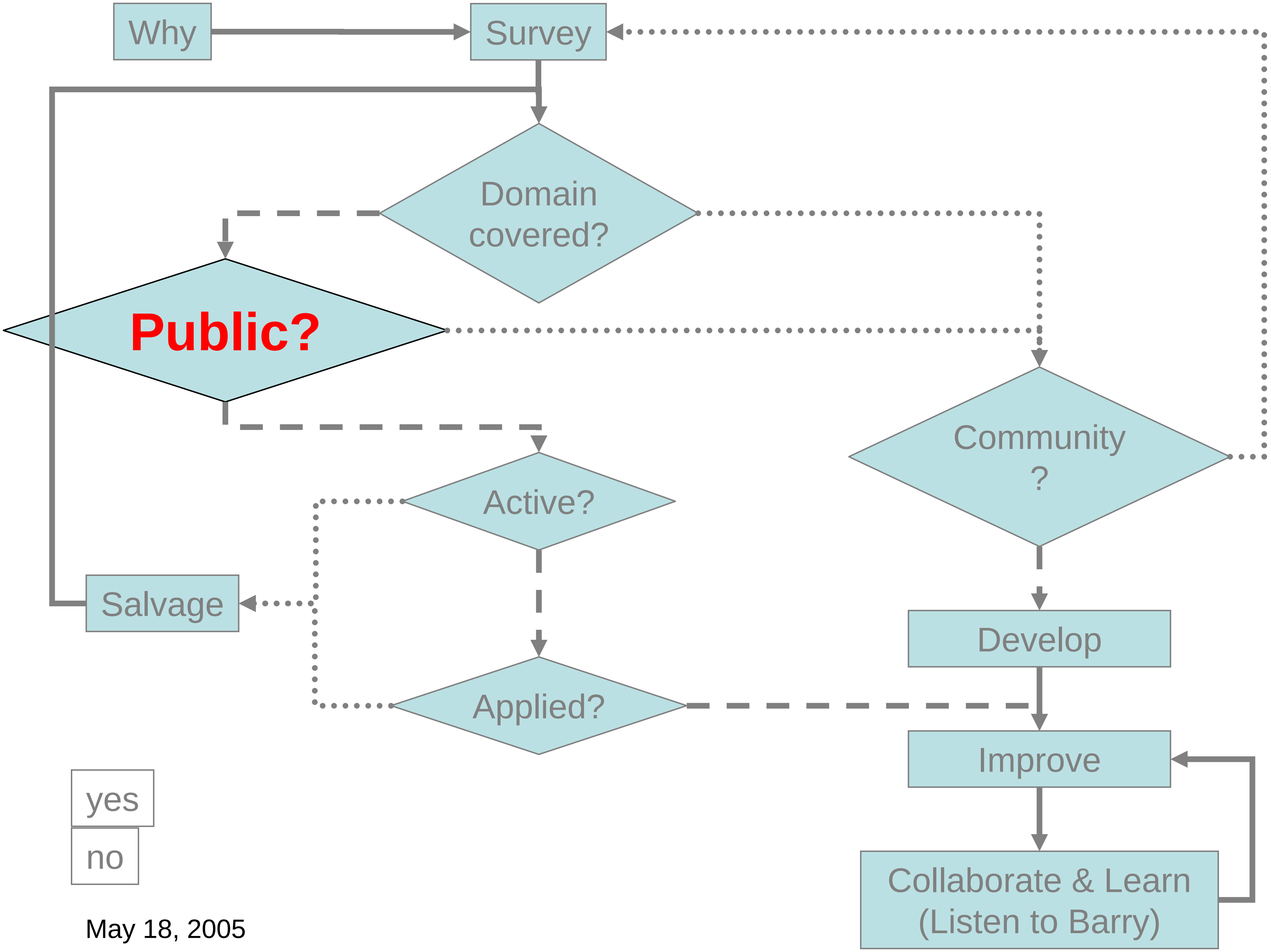
- Is there a community?
  - If not, need to rethink the question
- What domain does it cover?
- It is privately held?
- Is it active?
- Is it in applied use?


Why
Survey
Domain covered?
Public?
Community ?
Active?
Salvage
Develop
Applied?
Improve
Collaborate & Learn (Listen to Barry)
yes
no

May 18, 2005

Due diligence & background research

- Step 1: Learn what is out there
  - The most comprehensive list is on the OBO site. http://obo.sourceforge.net
- Assess ontologies critically and realistically.
- Do not reinvent. Collaborate.
- Start building—but not in isolation.



Why

Survey

Domain covered?

Public?

Community ?

Active?

Salvage

Develop

Applied?

Improve

Collaborate & Learn
(Listen to Barry)

yes

no

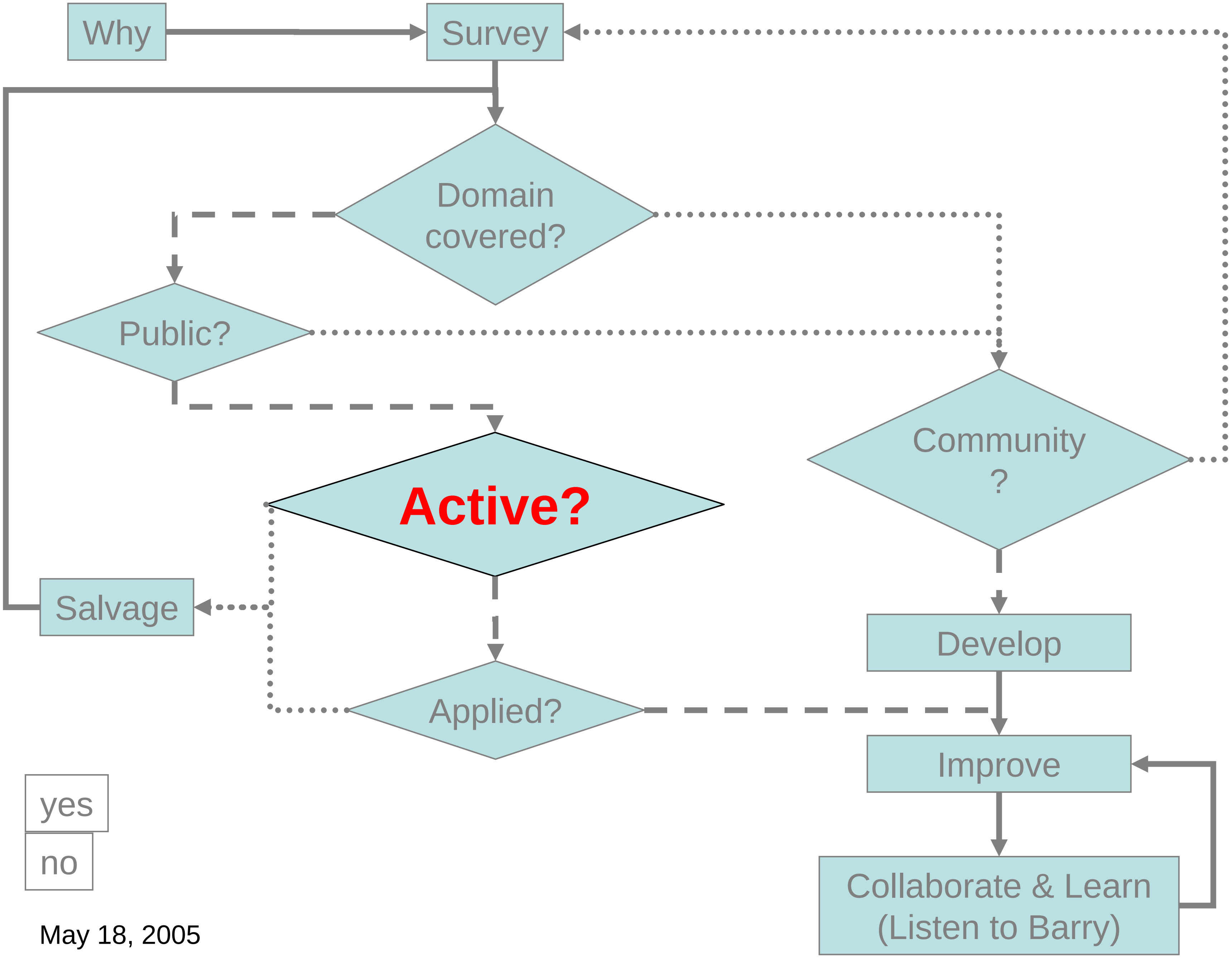
May 18, 2005

Ontologies must be shared

- Proprietary ontologies
  - Belief that ownership of the terminology gives the owners a competitive edge
  - For example, Incyte or Monsanto in the past

Ontologies must be shared

- Communities form scientific theories
  - that seek to explain all of the existing evidence
  - and can be used for prediction
- These communities are all directed to the same biological reality, but have their own perspective
- The computable representation must be shared
- Ontology development is inherently collaborative



Why

Survey

Domain covered?

Public?

Active?

Community ?

Salvage

Develop

Applied?

Improve

Collaborate & Learn (Listen to Barry)

yes

no

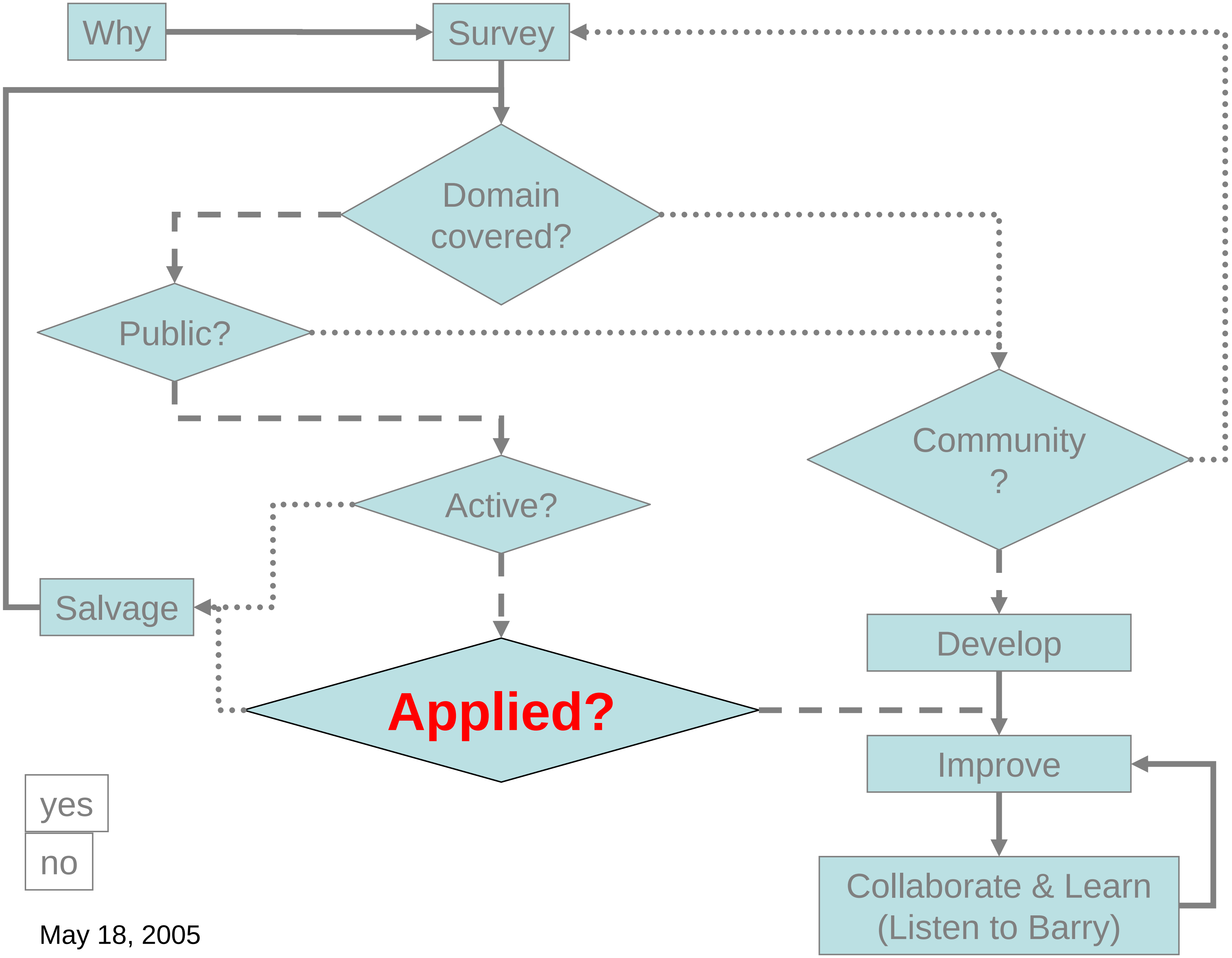
May 18, 2005

Pragmatic assessment of an ontology

- Is there access to help, e.g.:

  help-me@weird.ontology.inc ?

- Does a warm body answer help mail within a ‘reasonable’ time—say 2 working days ?



Why

Survey

Domain covered?

Public?

Community ?

Active?

Salvage

Develop

Applied?

Improve

Collaborate & Learn (Listen to Barry)

yes

no

May 18, 2005

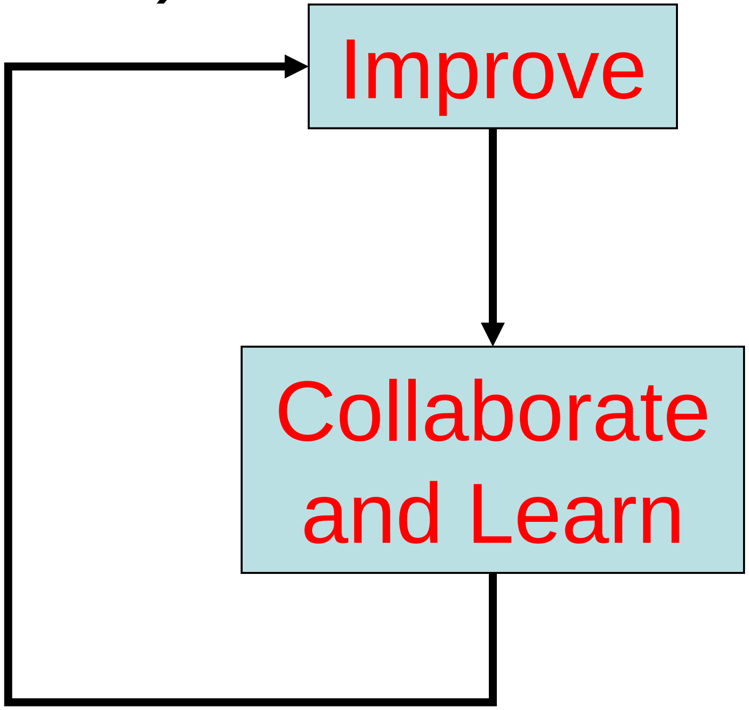
Where the rubber meets the road

- Every ontology improves when it is applied to actual instances of data
- It improves even more when these data are used to answer research questions
- There will be fewer problems in the ontology and more commitment to fixing remaining problems when important research data is involved that scientists depend upon
- Be very wary of ontologies that have never been applied

Work with that community

- To improve (if you found one)
- To develop (if you did not)

- How?



Improve

Collaborate and Learn

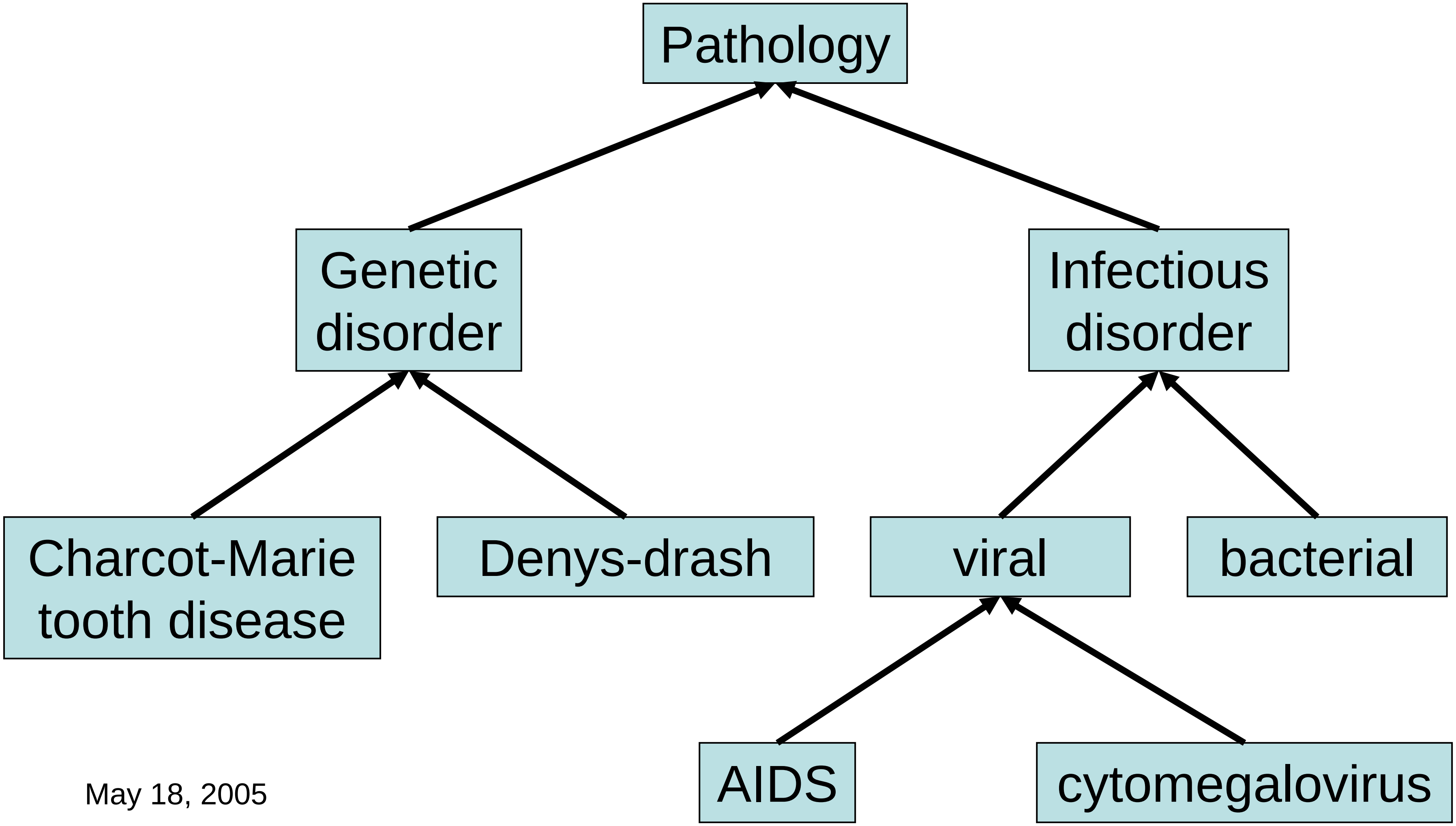
What do **YOU** call an ontology?

- Controlled vocabularies
  - A simple list of terms
- For example, EpoDB:
  - gene names and families, developmental stages, cell types, tissue types, experiment names, and chemical factors

What do YOU call an ontology?

- Pure subsumption hierarchies
  - single ‘*is_a*’ relationship
- For example, eVoc for attributes of cDNA libraries:
  - Anatomical system, cell type, development stage, experimental technique, microarray platform, pathology, pooling strategy, tissue preparation, treatment

eVOC *is_a* hierarchy



What is it **YOU** call an ontology?

- Data Model
  - BioPax: a specification for data exchange of biological (metabolic) processes
- Hybrids
  - Gene Ontology: Mix of subsumption (*is_a*), *part_of*, and *derives_from* relationships

What do **YOU** call an ontology?

- Suite
  - NCI Thesaurus
- Knowledgebases
  - PharmGKB
  - Reactome
  - IMGT (Immunogenetics]

A little sociology

Experience from building the GO

May 18, 2005

Community vs. Committee ?

- Members of a committee represent themselves.
  - Committees design camels
- Members of a community represent their community.
  - Communities design race horses

Design for purpose - not in abstract

- Who will use it?
  - If no one is interested, then go back to bed
- What will they use it for?
  - Define the domain
- Who will maintain it?
  - Be pragmatic and modest

GO takes the bottom-up approach

- Top-down is another strategy
- For example, the Foundational Model of Anatomy (FMA)
- Both require active involvement from community experts

Start with a concrete proposal —not a blank slate.

- But do not commit your ego to it.
- Distribute to a small group you respect:
  - With a shared commitment.
  - With broad domain knowledge.
  - Who will engage in vigorous debate without engaging their egos (or, at least not *too* much).
  - Who will do concrete work.

Step 1:

- Alpha0: the first proposal - broad in breadth but shallow in depth. By one person with broad domain knowledge.
  - Distribute to a small group (<6).
  - Get together for two days and engage in vigorous discussion. Be open and frank. Argue, but do not be dogmatic.
- Reiterate over a period of months. Do as much as possible face-to-face, rather than by phone/email. Meet for 2 days every 3 months or so.

Step 2:

- Distribute Alpha1 to your group.
- All now test this Alpha1 in real life.
- Do not worry that (at this stage) you do not have tools - hack it.

Step 3:

- Reconvene as a group for two days.
- Share experiences from implementation:
  - Can your Alpha1 be implemented in a useful way ?
  - What are the conceptual problems ?
  - What are the structural problems ?

Step 4:

- Establish a mechanism for change.
  - Use CVS or Subversion.
  - Limit the number of editors with write permission (ideally to one person).
- Release a Beta1.
- Seriously implement Beta1 in real life.
- Build the ontology in depth.

Step 5:

- After about 6 months reconvene and evaluate.
- Is the ontology suited to its purpose ?
- Is it, in practice, usable ?
- Are we happy about its broad structure and content ?

Step 6:

- Go public.
  - Release ontology to community.
  - Release the products of its instantiation.
  - Invite broad community input and establish a mechanism for this (e.g. SourceForge).

Step 7:

- Proselytize.
  - Publish in a high profile journal.
  - Engage new user groups.
- Emphasize openness.
- Write a grant.

Step 8:

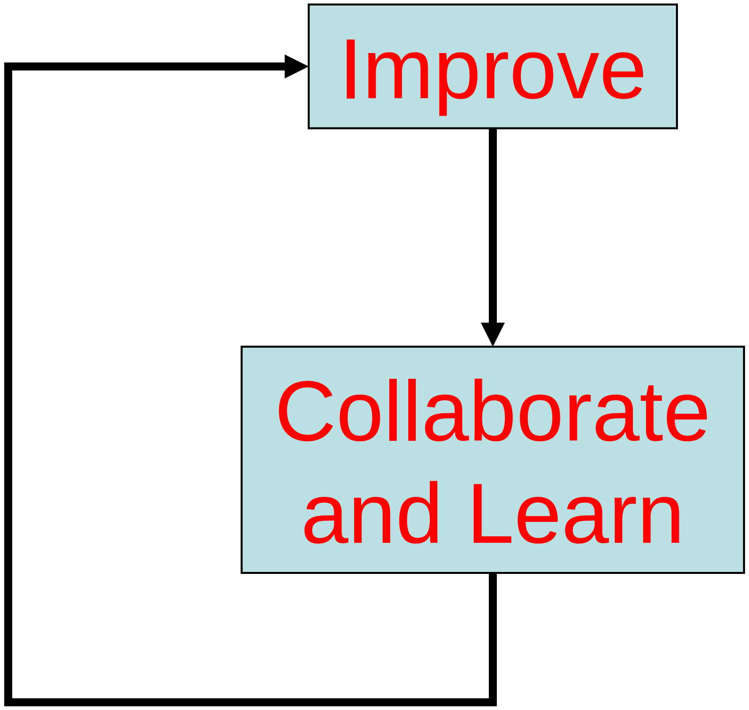
Have fun!

Take-home message

- Don’t reinvent—*Use* the power of combination and collaboration

Improvements come in two forms

- Getting it right
  - It is impossible to get it right the 1st (or 2nd, or 3rd, …) time.
- What we know about reality is continually growing



Principles for Building Biomedical Ontologies

Barry Smith

http://ifomis.de

May 18, 2005

Ontologies as Controlled Vocabularies

- expressing discoveries in the life sciences in a uniform way

- providing a uniform framework for managing annotation data deriving from different sources and with varying types and degrees of evidence

Overview

- Following basic rules helps make better ontologies
- We will work through some examples of ontologies which do and not follow basic rules
- We will work through the principles-based treatment of relations in ontologies, to show how ontologies can become more reliable and more powerful

Why do we need rules for good ontology?

- Ontologies must be intelligible both to humans (for annotation) and to machines (for reasoning and error-checking)
- Unintuitive rules for classification lead to entry errors (problematic links)
- Facilitate training of curators
- Overcome obstacles to alignment with other ontology and terminology systems
- Enhance harvesting of content through automatic reasoning systems

SNOMED-CT Top Level

- Substance
- Body Structure
- Specimen
- Context-Dependent Categories*
- Attribute
- Finding*
- Staging and Scales
- Organism
- Physical Object
- Events
- Environments and Geographic Locations
- Qualifier Value
- Special Concept*
- Pharmaceutical and Biological Products
- Social Context
- Disease
- Procedure
- Physical Force

Examples of Rules

- Don’t confuse entities with concepts
- Don’t confuse entities with ways of getting to know entities
- Don’t confuse entities with ways of talking about entities
- Don’t confuse entities with artifacts of your database representation ...
- An ontology should not change when the programming language changes

First Rule: Univocity

- Terms (including those describing relations) should have the same meanings on every occasion of use.
- In other words, they should refer to the same kinds of entities in reality

Example of univocity problem in case of *part_of* relation

(Old) Gene Ontology:

- ‘part_of’ = ‘may be part of’
  - flagellum part_of cell
- ‘part_of’ = ‘is at times part of’
  - replication fork part_of the nucleoplasm
- ‘part_of’ = ‘is included as a sub-list in’

Second Rule: Positivity

- Complements of classes are not themselves classes.
- Terms such as ‘non-mammal’ or ‘non-membrane’ do not designate genuine classes.

Third Rule: Objectivity

- Which classes exist is not a function of our biological knowledge.

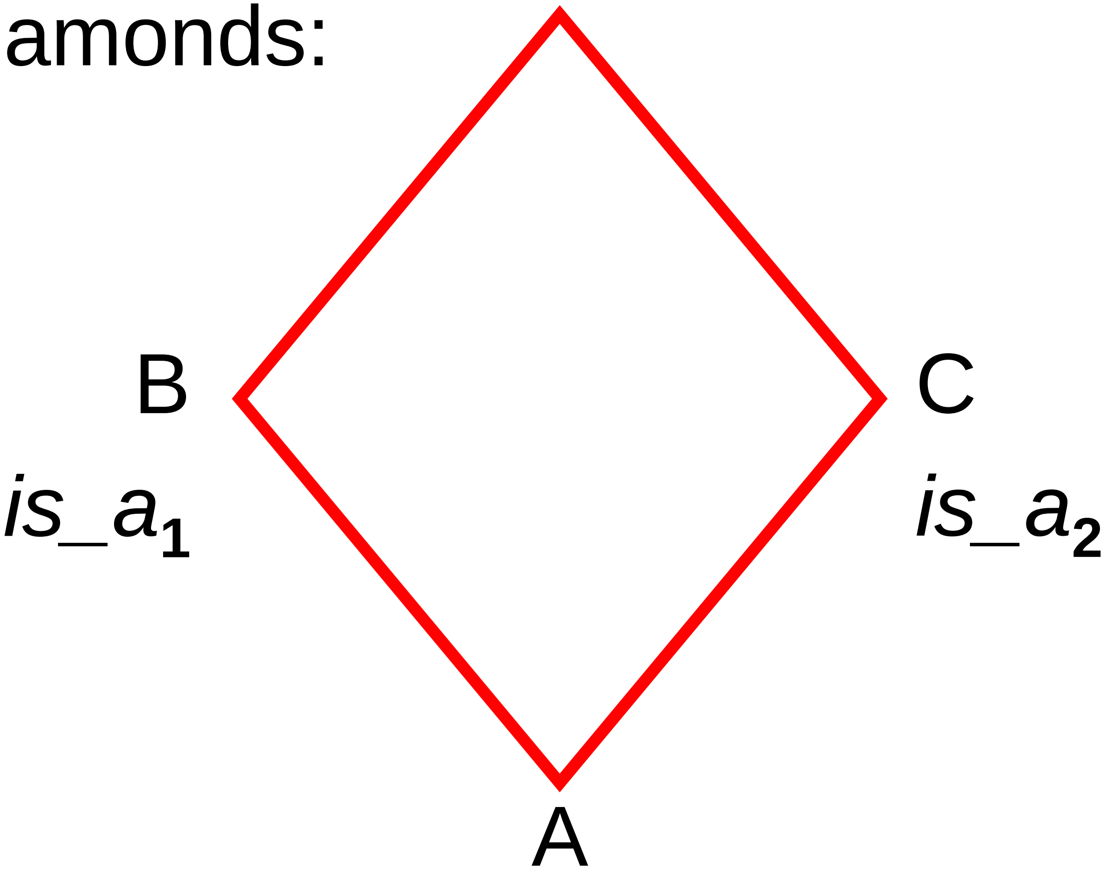
- Terms such as ‘unknown’ or ‘unclassified’ or ‘unlocalized’ do not designate biological natural kinds.

Fourth Rule: Single Inheritance

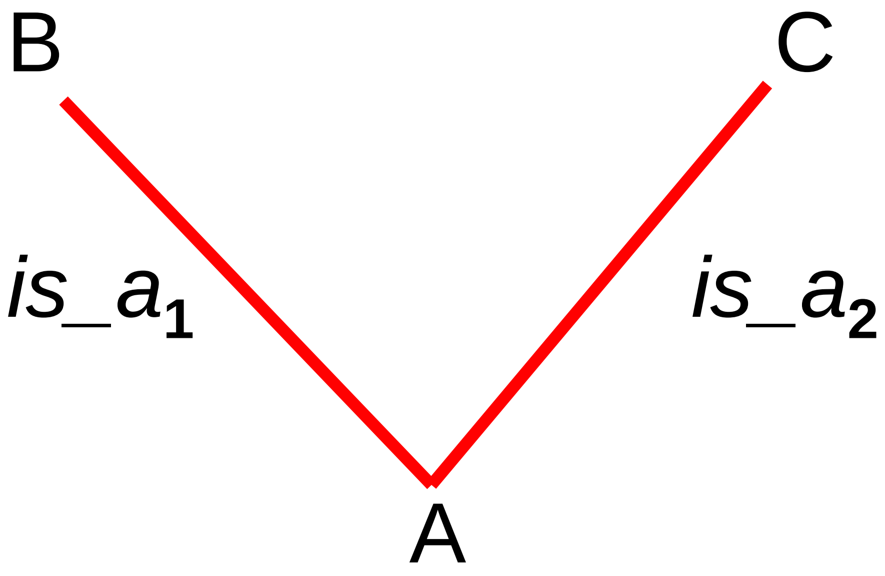
No class in a classificatory hierarchy should have more than one *is_a* parent on the immediate higher level

Rule of Single Inheritance

- no diamonds:



Problems with multiple inheritance



B

C

is_a$_1$

is_a$_2$

A

‘*is_a*’ no longer univocal

‘*is_a*’ is pressed into service to mean a variety of different things

- shortfalls from single inheritance are often clues to incorrect entry of terms and relations
- the resulting ambiguities make the rules for correct entry difficult to communicate to human curators

is_a Overloading

- serves as obstacle to integration with neighboring ontologies
- The success of ontology alignment depends crucially on the degree to which basic ontological relations such as *is_a* and *part_of* can be relied on as having the same meanings in the different ontologies to be aligned.

Use of multiple inheritance

- The resultant mélange makes coherent integration across ontologies achievable (at best) only under the guidance of human beings with relevant biological knowledge
- How much should reasoning systems be forced to rely on human guidance?

Fifth Rule: Intelligibility of Definitions

- The terms used in a definition should be simpler (more intelligible) than the term to be defined
- otherwise the definition provides no assistance
  - to human understanding
  - for machine processing

To the degree that the above rules are not satisfied, error checking and ontology alignment will be achievable, at best, only with human intervention and via force majeure

Some rules are Rules of Thumb

- The world of biomedical research is a world of difficult trade-offs
- The benefits of formal (logical and ontological) rigor need to be balanced
  - Against the constraints of computer tractability,
  - Against the needs of biomedical practitioners.
- BUT alignment and integration of biomedical information resources will be achieved only to the degree that such resources conform to these standard principles of classification and definition

Current Best Practice:

The Foundational Model of Anatomy

- Follows formal rules for definitions laid down by Aristotle.
- A definition is the specification of the essence (nature, invariant structure) shared by all the members of a class or natural kind.

The Aristotelian Methodology

- Topmost nodes are the undefinable primitives.
- The definition of a class lower down in the hierarchy is provided by specifying the parent of the class together with the relevant *differentia*.
- *Differentia* tells us what marks out instances of the defined class within the wider parent class as in
  - human == *rational* animal.

FMA Examples

- **Cell**
  - *is an* **anatomical structure** **[topmost node]**
  - that *consists of* **cytoplasm** *surrounded by* a **plasma membrane** with or without a **cell nucleus** **[differentia]**

The FMA regimentation

- Brings the advantage that each definition reflects the position in the hierarchy to which a defined term belongs.
- The position of a term within the hierarchy enriches its own definition by incorporating automatically the definitions of all the terms above it.
- The entire information content of the FMA’s term hierarchy can be translated very cleanly into a computer representation

Definitions should be intelligible to both machines and humans

- Machines can cope with the full formal representation
- Humans need to use modularity
- **Plasma membrane**
  - *is a* **cell part [immediate parent]**
  - that *surrounds* the **cytoplasm [differentia]**

Terms and relations should have clear definitions

- These tell us how the ontology relates to the world of biological instances, meaning the actual particulars in reality:
    - actual cells, actual portions of cytoplasm, and so on…

Sixth Rule: Basis in Reality

- When building or maintaining an ontology, always think carefully at how classes (types, kinds, species) relate to instances in reality

Axioms governing instances

- Every class has at least one instance

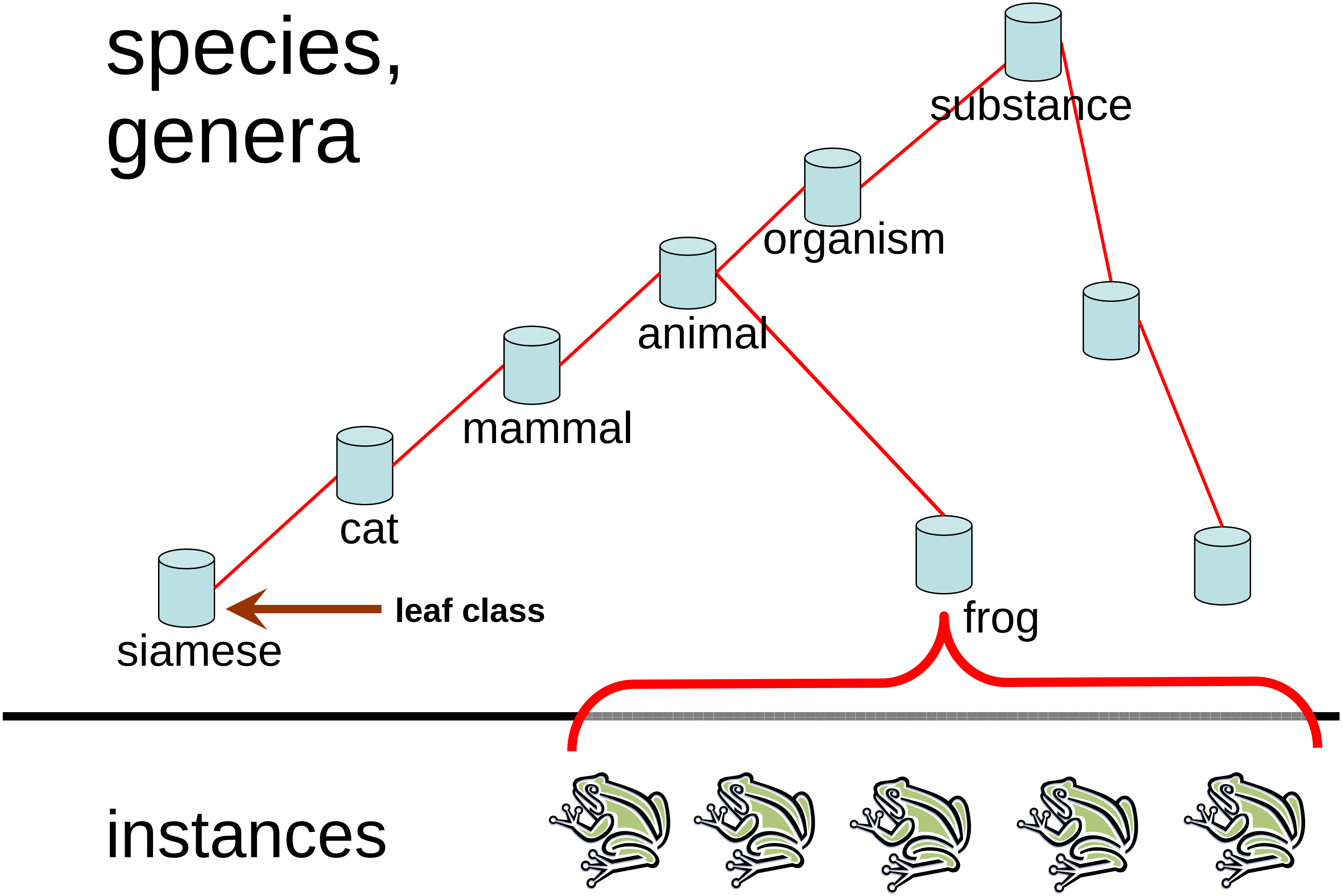
- Every genus (parent class) has an instantiated species (differentia + genus)

- Each species (child class) has a smaller class of instances than its genus (parent class)

Axioms governing Instances

- Distinct classes on the same level never share instances
- Distinct leaf classes within a classification never share instances

species, genera

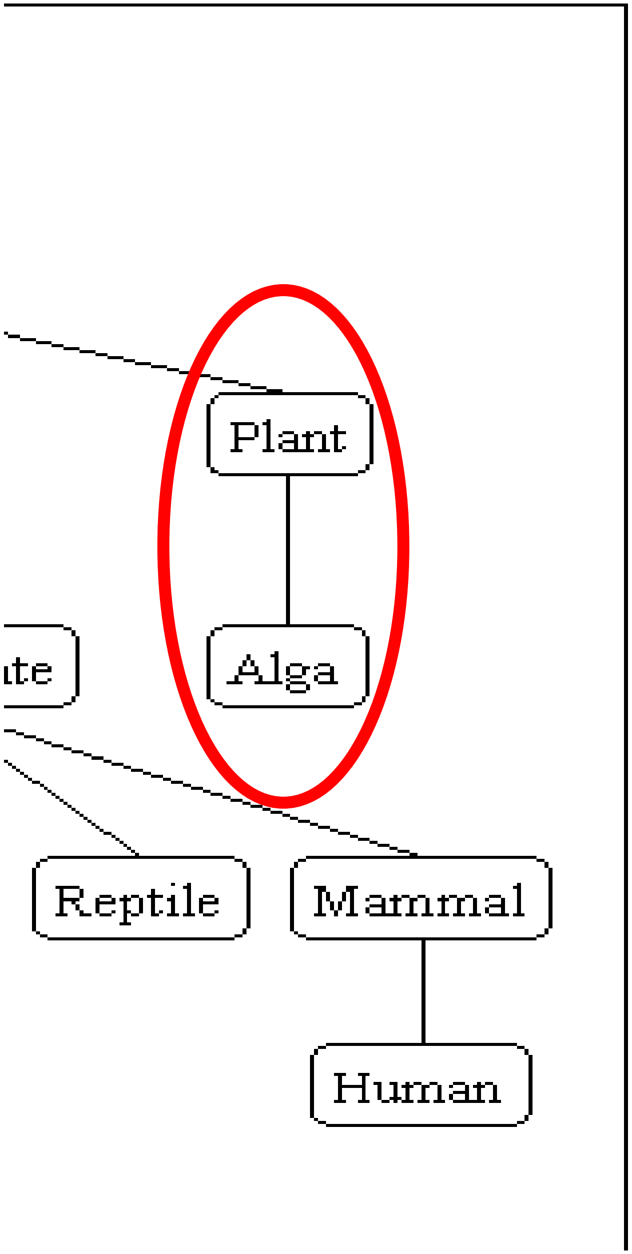


instances

May 18, 2005

Axioms

- Every genus (parent class) has at least two children



- UMLS Semantic Network

May 18, 2005

Interoperability

- Ontologies should work together
  - ways should be found to avoid redundancy in ontology building and to support reuse
  - ontologies should be capable of being used by other ontologies (cumulation)

Main obstacle to integration

- Current ontologies do not deal well with
  - Time and
  - Space and
  - Instances (particulars)
- Our definitions should link the terms in the ontology to instances in spatio-temporal reality

The problem of ontology alignment

SNOMED

MeSH

UMLS

NCIT

HL7-RIM …

- **Still remain too much at the level of TERMINOLOGY**
- **Not based on a common set of rules**
- **Not based on a common set of relations**

None of these have clearly defined relations

An example of an unclear definition *A is_a B*

- ‘A’ is more specific in meaning than ‘B’

- unicorn *is_a* one-horned mammal
- HL7-RIM: Individual Allele *is_a* Act of Observation
- cancer documentation *is_a* cancer
- disease prevention *is_a* disease

Benefits of well-defined relationships

- If the relations in an ontology are well-defined, then reasoning can cascade from one relational assertion (*A* R_1 *B*) to the next (*B* R_2 *C*). Relations used in ontologies thus far have not been well defined in this sense.

- *Find all DNA binding proteins* should also find all transcription factor proteins because
  - *Transcription factor is_a DNA binding protein*

How to define *A is_a B*

A is_a B =def.

1. *A* and *B* are names of universals (natural kinds, types) in reality
2. all instances of *A* are as a matter of biological science also instances of B

A standard definition of *part_of*

A part_of B =def

A composes (with one or more other physical units) some larger whole *B*

This confuses relations between meanings or concepts with relations entities in reality

Biomedical ontology integration / interoperability

- Will never be achieved through integration of meanings or concepts
- The problem is precisely that different user communities use *different concepts*
- ***What’s really needed is to have well-defined commonly used relationships***

Idea:

- Move from associative relations between meanings to strictly defined relations between the entities themselves.
- The relations can then be used computationally in the way required

Key idea:
To define ontological relations

- For example: *part_of, develops_from*
- Definitions will enable computation
- It is not enough to look just at classes or types.
  - We need also to take account of *instances* and *time*

Kinds of relations

- Between classes:
  - *is_a*, *part_of*, ...
- Between an instance and a class
  - this explosion **instance_of** the class explosion
- Between instances:
  - Mary’s heart **part_of** Mary

Key

- In the following discussion:
- Classes are in upper case
  - ‘*A*’ is the class
- Instances are in lower case
  - ‘*a*’ is a particular instance

Seventh Rule: Distinguish Universals and Instances

- A good ontology must distinguish clearly between
  - **universals (types, kinds, classes)**

  and
  - **instances (tokens, individuals, particulars)**

Don’t forget instances when defining relations

- *part_of* as a relation between classes versus ***part_of*** as a relation between instances
- *nucleus part_of cell*
- your heart ***part_of*** you

Part_of as a relation between classes is more problematic than is standardly supposed

- testis *part_of* human being ?
- heart *part_of* human being ?
- human being *has_part* human testis ?

Analogous distinctions are required for nearly all foundational relations of ontologies and semantic networks:

- *A causes B*
- *A is_located in B*
- *A is_adjacent_to B*

Reference to instances is necessary in defining mereotopological relations such as spatial occupation and spatial adjacency

Why distinguish universals from instances?

- What holds on the level of instances may not hold on the level of universals

- *nucleus adjacent_to cytoplasm*
- **Not:** *cytoplasm adjacent_to nucleus*
- *seminal vesicle adjacent_to urinary bladder*
- **Not:** *urinary bladder adjacent_to seminal vesicle*

part_of

- *part_of* must be time-indexed for spatial universals
- *A part_of B* is defined as:

  Given any instance *a* and any time *t*,

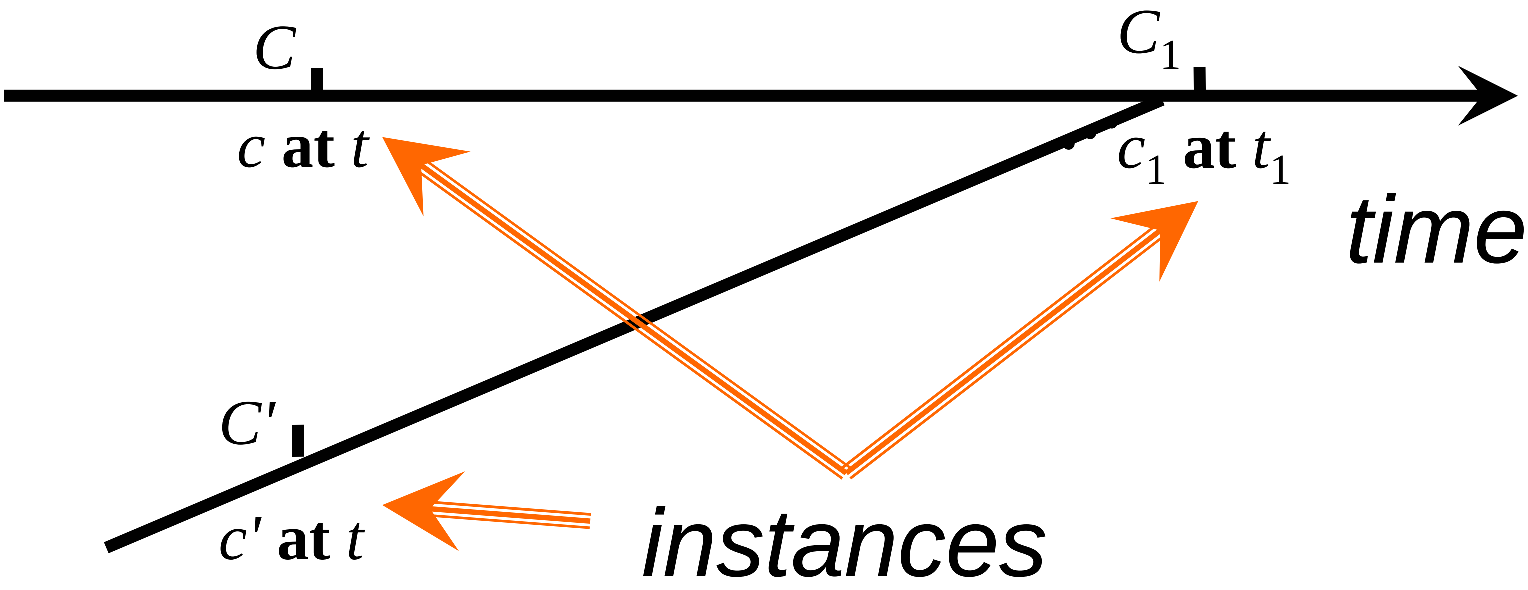
  If *a* is an instance of the universal *A* at *t*,

  then there is some instance *b* of the universal *B*

  such that

  a is an instance-level **part_of** *b* at *t*

derives_from

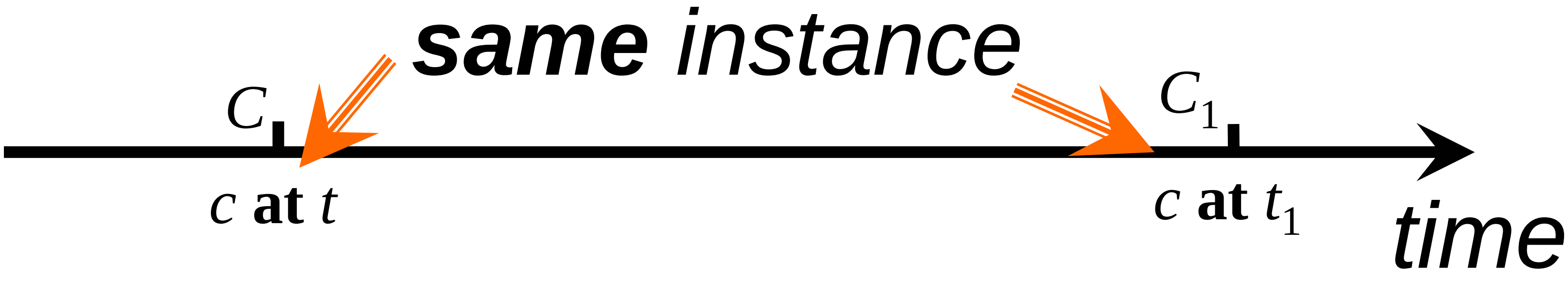


ovum

zygote derives_from

sperm

transformation_of



***same** instance*

C C_1

c **at** t

c **at** t_1

time

pre-RNA → mature RNA

child → adult

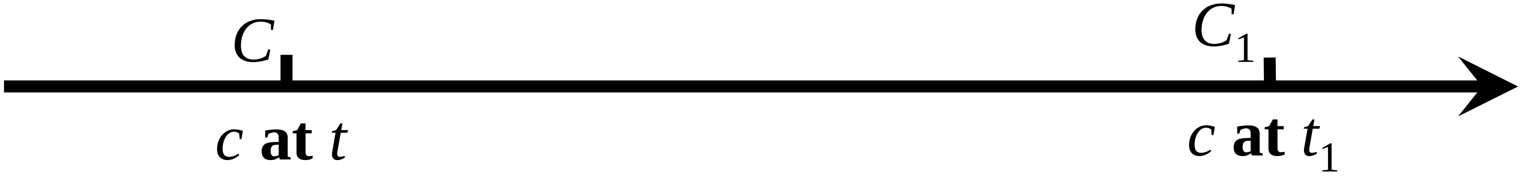
transformation_of

- C_2 transformation_of C_1 is defined as

  Given any instance c of C_2

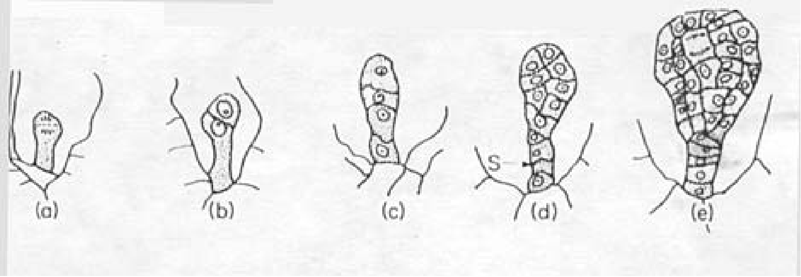
  c was at some earlier time an instance of C_1

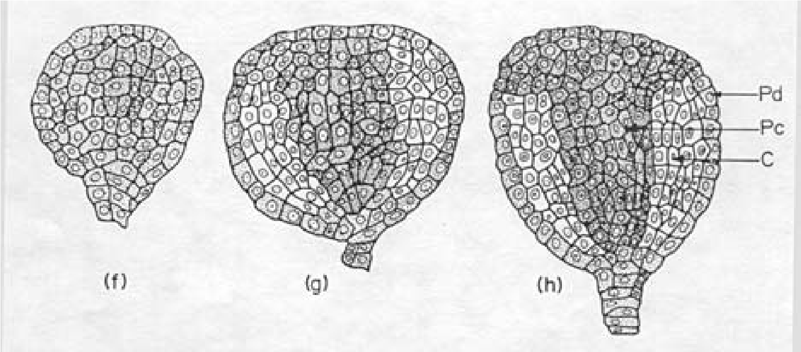
embryological development



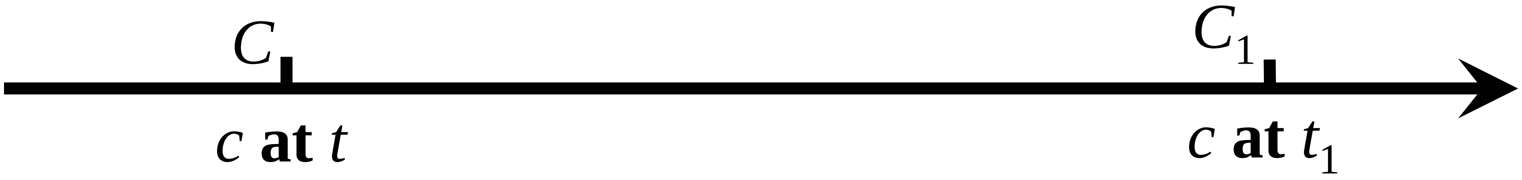
C C_1

c **at** t c **at** t_1





tumor development

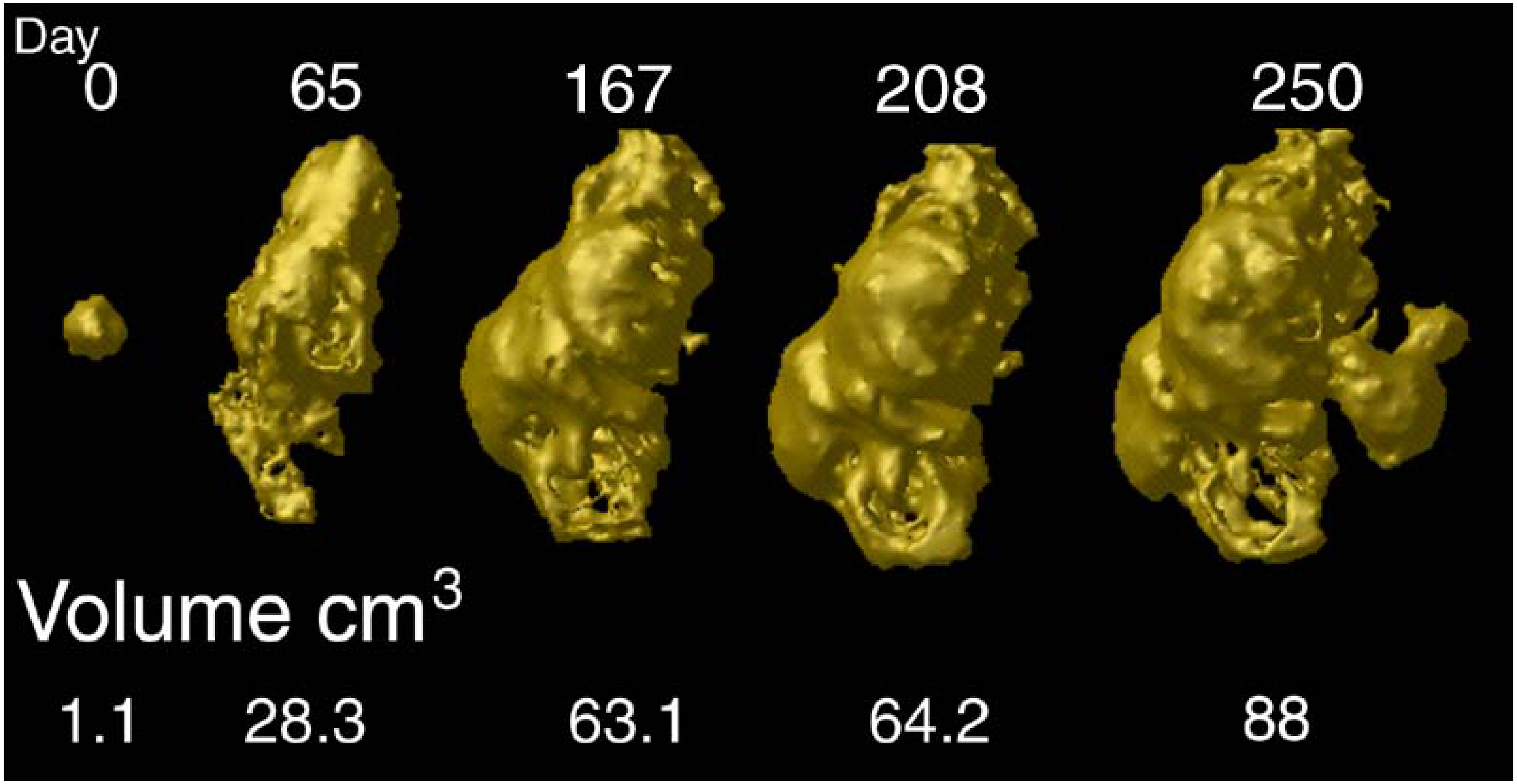


C

C_1

c **at** t

c **at** t_1



May 18, 2005

Definitions of the **all-some** form

allow cascading inferences

If A R_1 B and B R_2 C, then we know that

every A stands in R_1 to *some* B, but we know also that, whichever B this is, it can be plugged into the R_2 relation, because R_2 is defined for *every* B.

Not only relations

- We can apply the same methodology to other top-level categories in ontology, e.g.
  - anatomical structure
  - process
  - function (regulation, inhibition, suppression, co-factor ...)
  - boundary, interior (contact, separation, continuity)
  - tissue, membrane, sequence, cell

Relations to describe topology of nucleic sequence features

- Based on the formal relationships between pairs of intervals in a 1-dimensional space.
- Uses the coincidence of edges and interiors
- Enables questions regarding the equality, overlap, disjointedness, containment and coverage of genomic features.
- Conventional operations in genomics are simplified
- Software no longer needs to know what kind of feature particular instances are

For features A & B	An end of A intersects an end of B	Interior of A intersects interior of B	An end of A intersects interior of B	Interior of A intersects an end of B
A is ***disjoint*** *from B*	False	False	False	False
A ***meets*** *B*	True	False	False	False
A ***overlaps*** *B*	False	True	True	True
A is ***inside*** *B*	False	True	True	False
A ***contains*** *B*	False	True	False	True
A ***covers*** *B*	True	True	False	True
A is ***covered_by*** *B*	True	True	True	False
A equals B	True	True	False	False

disjoint

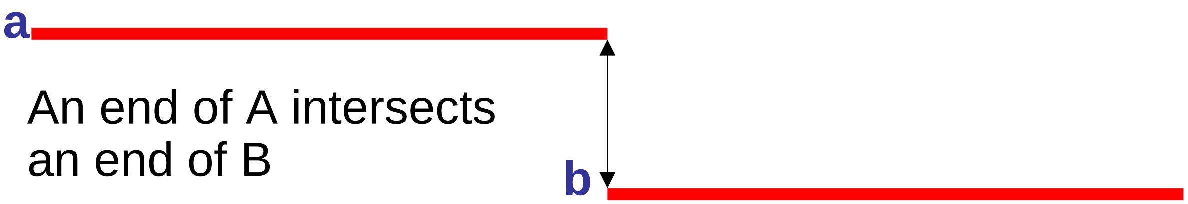
An end of A does NOT intersect an end of B

Interior of A does NOT intersect interior of B

An end of A does NOT intersect interior of B

Interior of A does NOT intersect an end of B

meets



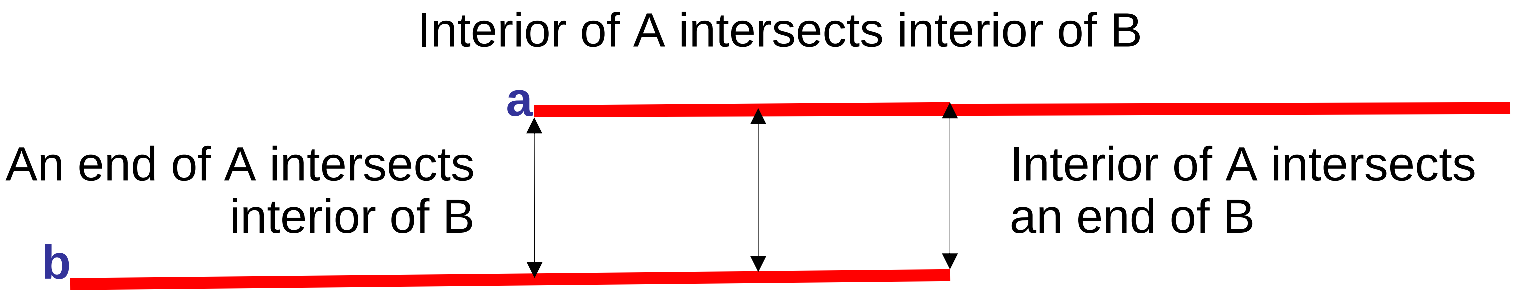
An end of A intersects
an end of B

An end of A does NOT intersect interior of B

Interior of A does NOT intersect an end of B

Interior of A does NOT intersect interior of B

overlaps



Interior of A intersects interior of B

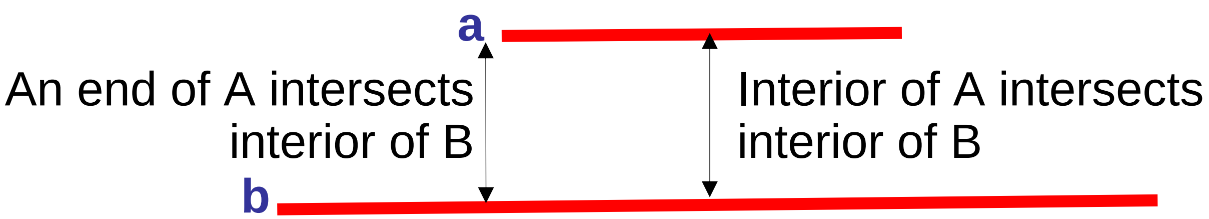
An end of A intersects interior of B

Interior of A intersects an end of B

An end of A does NOT intersect an end of B

May 18, 2005

inside



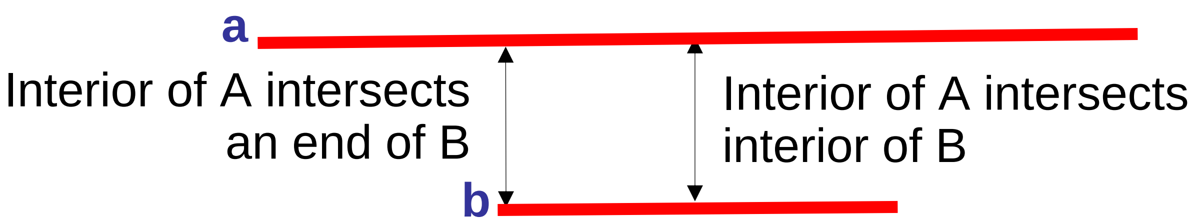
An end of A intersects interior of B

Interior of A intersects interior of B

Interior of A does NOT intersect an end of B

An end of A does NOT intersect an end of B

contains



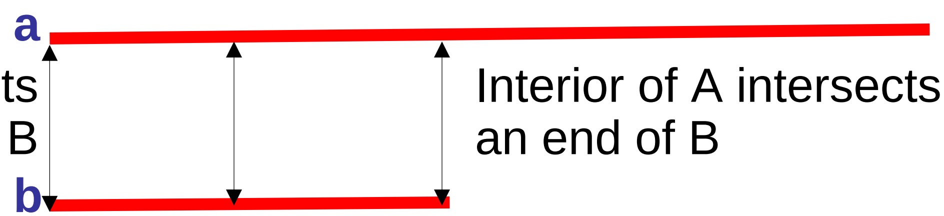
Interior of A intersects an end of B

Interior of A intersects interior of B

An end of A does NOT intersect an end of B

An end of A does NOT intersect interior of B

covers



Interior of A intersects interior of B

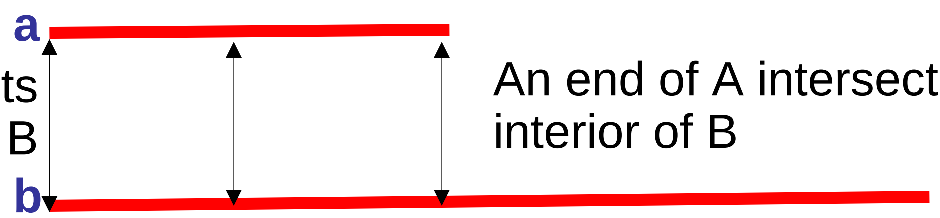
An end of A intersects
an end of B

Interior of A intersects
an end of B

An end of A does NOT intersect interior of B

covered_by

Interior of A intersects interior of B



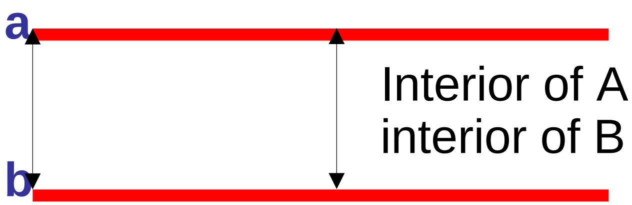
An end of A intersects an end of B

An end of A intersects interior of B

Interior of A does NOT intersect an end of B

May 18, 2005

equals



An end of A intersects an end of B

Interior of A intersects interior of B

An end of A does NOT intersect an interior of B

Interior of A does NOT intersect an end of B

The Rules

1. **Univocity:** Terms should have the same meanings on every occasion of use
2. **Positivity:** Terms such as ‘non-mammal’ or ‘non-membrane’ do not designate genuine classes.
3. **Objectivity:** Terms such as ‘unknown’ or ‘unclassified’ or ‘unlocalized’ do not designate biological natural kinds.
4. **Single Inheritance:** No class in a classification hierarchy should have more than one *is_a* parent on the immediate higher level
5. **Intelligibility of Definitions:** The terms used in a definition should be simpler (more intelligible) than the term to be defined
6. **Basis in Reality:** When building or maintaining an ontology, always think carefully at how classes relate to instances in reality
7. **Distinguish Universals and Instances**

What we have argued for:

- A methodology which enforces clear, coherent definitions
- This promotes quality assurance
  - intent is not hard-coded into software
  - Meaning of relationships is defined, not inferred
- Guarantees automatic reasoning across ontologies and across data at different granularities

Principles for Building Biomedical Ontologies

Rama Balakrishnan and David Hill

http://www.geneontology.org

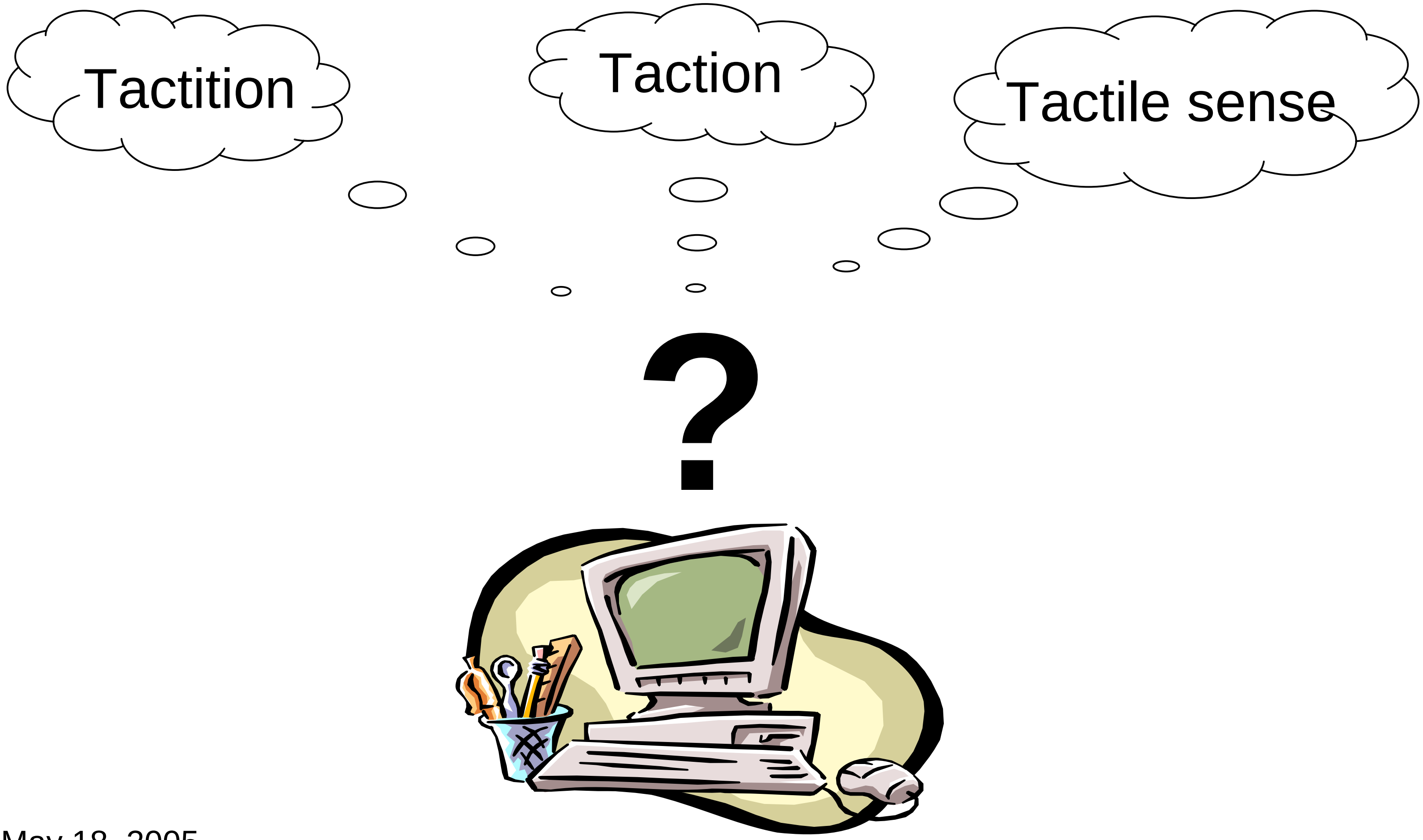
May 18, 2005

How has GO dealt with some specific aspects of ontology development?

- Univocity
- Positivity
- Objectivity
- Definitions
  - Formal definitions
  - Written definitions
- Ontology Alignment

The Challenge of Univocity:

People call the same thing by different names



May 18, 2005

Univocity: GO uses 1 term and many characterized synonyms

Tactition

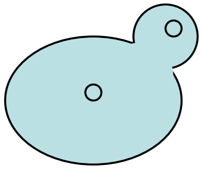
Taction

Tactile sense

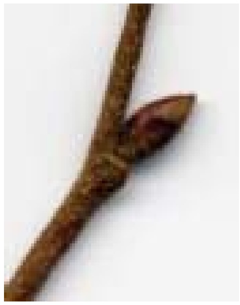
perception of touch ; GO:0050975

The Challenge of Univocity: People use the same words to describe different things

= bud initiation

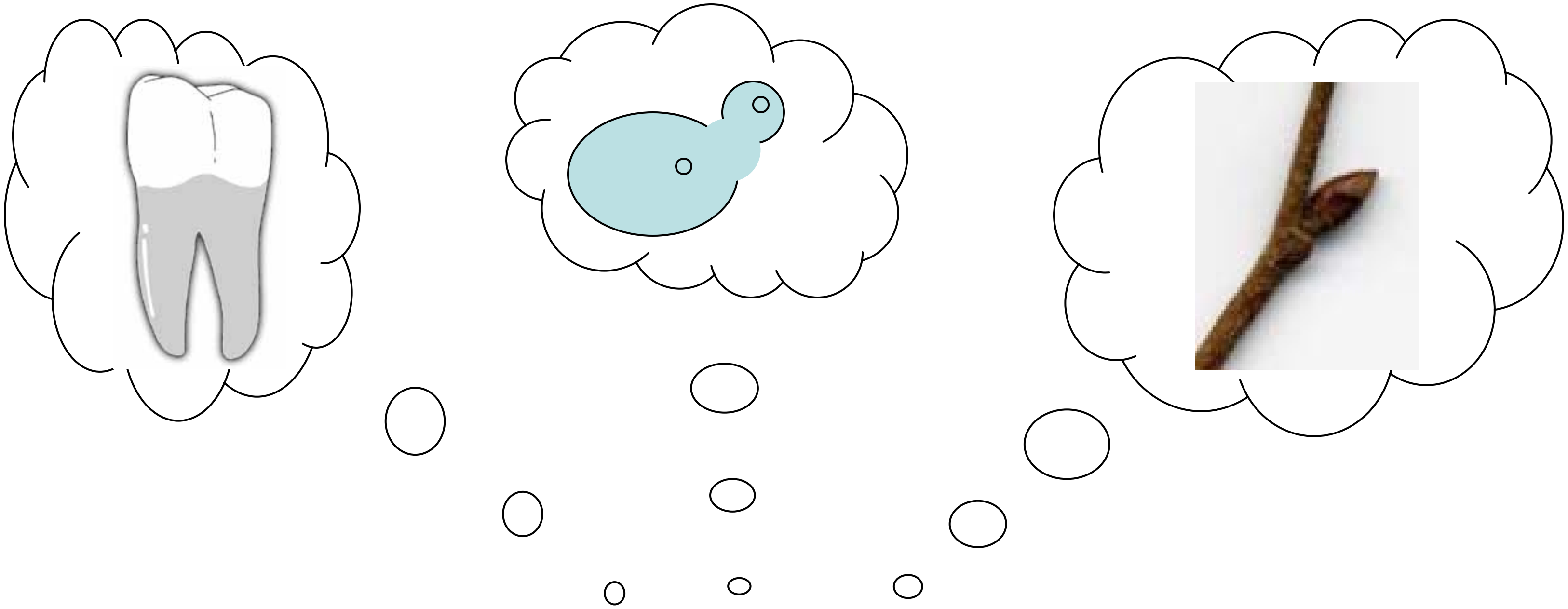


= bud initiation



= bud initiation

May 18, 2005



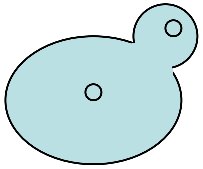
Bud initiation? How is
a computer to know?

May 18, 2005

Univocity: GO adds “sensu” descriptors to discriminate among organisms

= bud initiation

sensu *Metazoa*



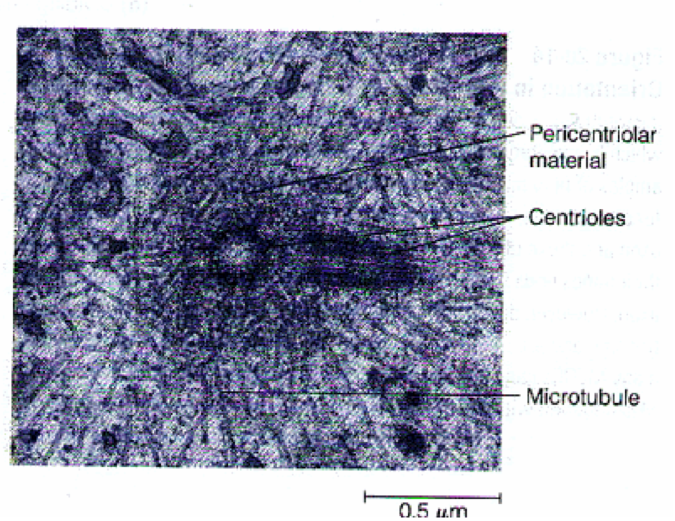
= bud initiation

sensu *Saccharomyces*

= bud initiation

sensu *Viridiplantae*

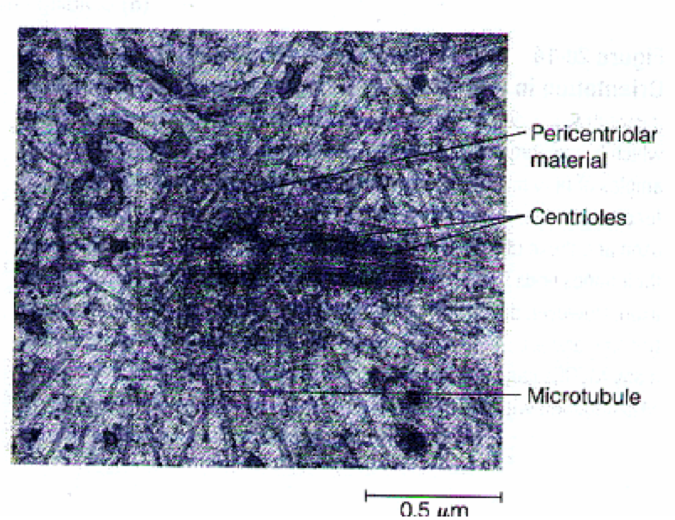
The Challenge of Positivity



Some organelles are membrane-bound.
A centrosome is not a membrane bound organelle,
but it still may be considered an organelle.

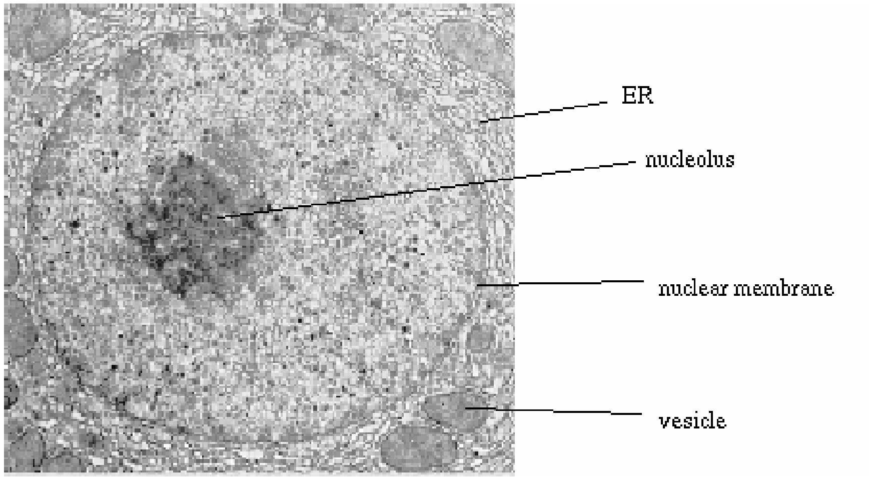
May 18, 2005

The Challenge of Positivity: Sometimes absence is a distinction in a Biologist's mind



non-membrane-bound organelle

GO:0043228



membrane-bound organelle

GO:0043227

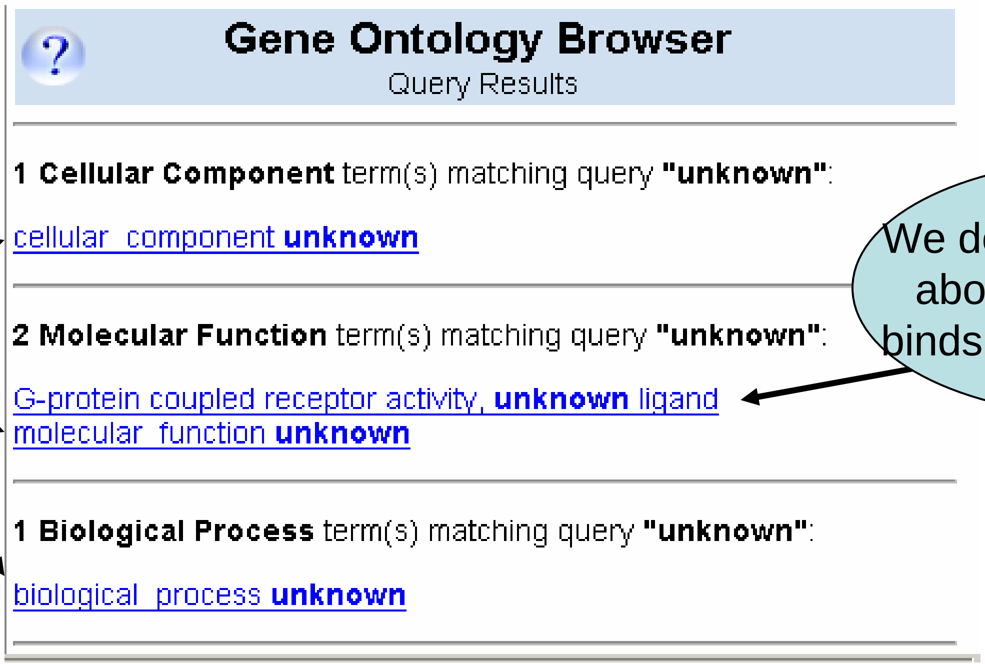
Positivity

- Note the logical difference between
  - “*non-membrane-bound organelle*” and
  - “*not a membrane-bound organelle*”

- The latter includes everything that is not a membrane bound organelle!

The Challenge of Objectivity: Database users want to know if we don’t know anything (Exhaustiveness with respect to knowledge)

We don’t know anything about a gene product with respect to these



Gene Ontology Browser

Query Results

1 Cellular Component term(s) matching query **"unknown"**:

cellular_component **unknown**

2 Molecular Function term(s) matching query **"unknown"**:

G-protein coupled receptor activity, **unknown** ligand
molecular_function **unknown**

1 Biological Process term(s) matching query **"unknown"**:

biological_process **unknown**

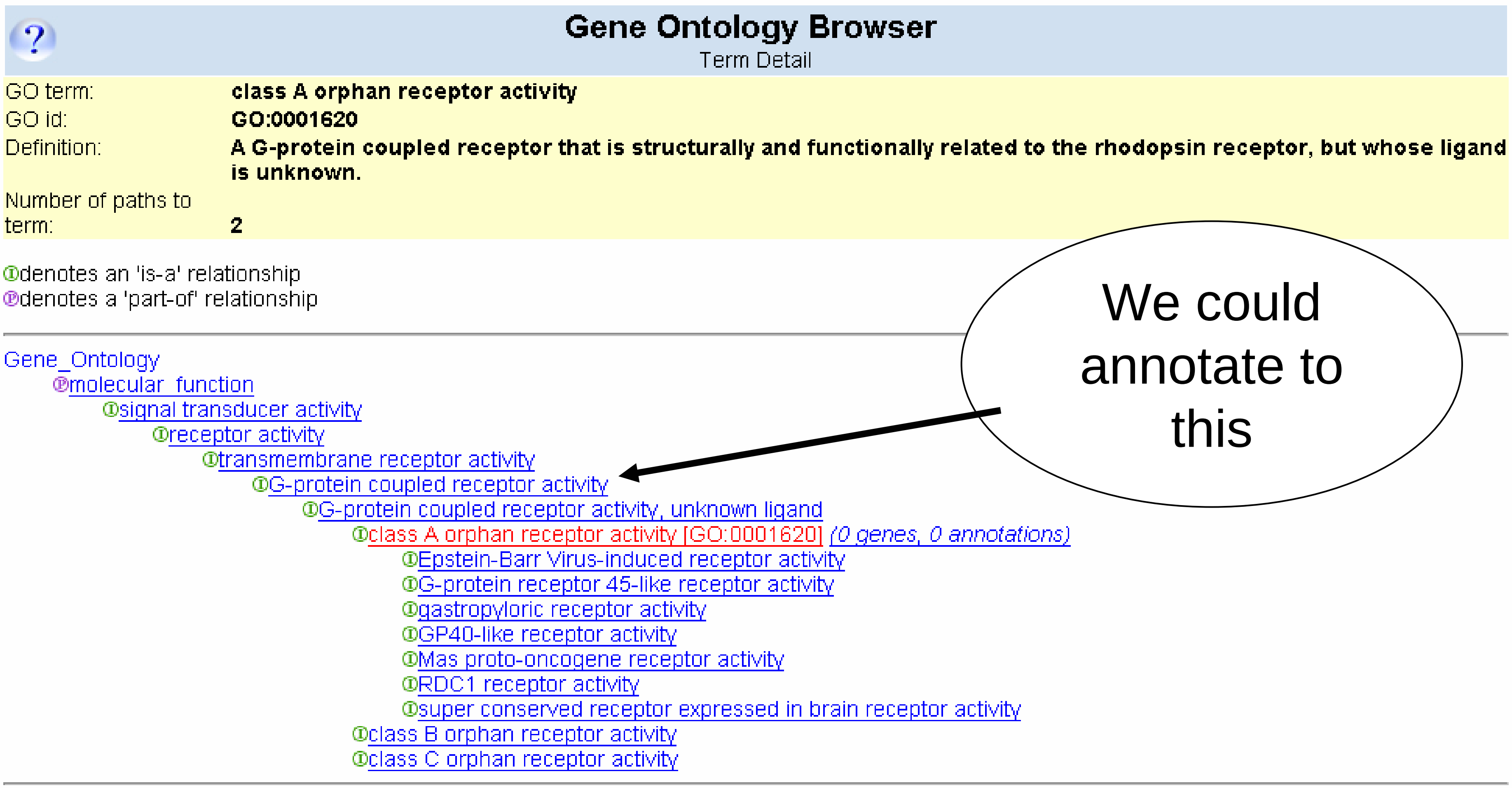
We don’t know anything about the ligand that binds this type of GPCR

May 18, 2005

Objectivity

- How can we use GO to annotate gene products when we know that we don’t have any information about them?
  - Currently GO has terms in each ontology to describe unknown
  - An alternative might be to annotate genes to root nodes and use an evidence code to describe that we have no data.
- Similar strategies could be used for things like receptors where the ligand is unknown.

GPCRs with unknown ligands



Gene Ontology Browser

Term Detail

GO term:	**class A orphan receptor activity**
GO id:	**GO:0001620**
Definition:	**A G-protein coupled receptor that is structurally and functionally related to the rhodopsin receptor, but whose ligand is unknown.**
Number of paths to term:	**2**

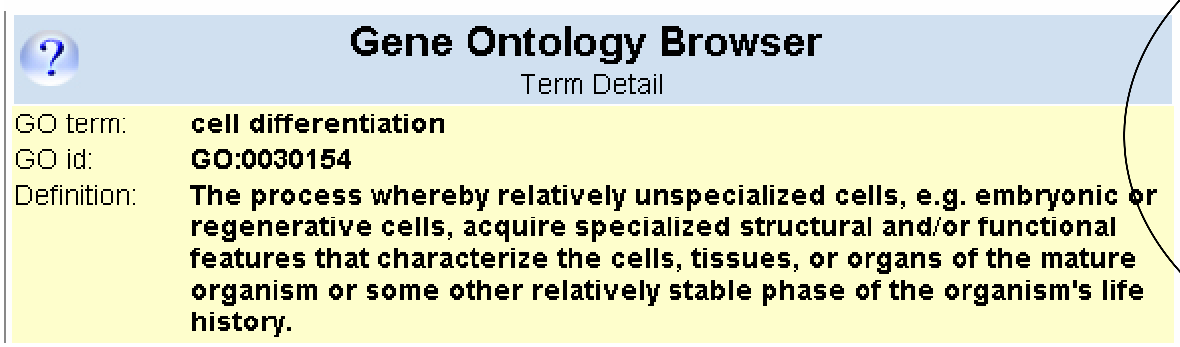
Ⓘdenotes an 'is-a' relationship
Ⓟdenotes a 'part-of' relationship

- Gene_Ontology
  - Ⓟmolecular_function
    - Ⓘsignal transducer activity
      - Ⓘreceptor activity
        - Ⓘtransmembrane receptor activity
          - ⒾG-protein coupled receptor activity
            - ⒾG-protein coupled receptor activity, unknown ligand
              - Ⓘclass A orphan receptor activity [GO:0001620] *(0 genes, 0 annotations)*
                - ⒾEpstein-Barr Virus-induced receptor activity
                - ⒾG-protein receptor 45-like receptor activity
                - Ⓘgastropyloric receptor activity
                - ⒾGP40-like receptor activity
                - ⒾMas proto-oncogene receptor activity
                - ⒾRDC1 receptor activity
                - Ⓘsuper conserved receptor expressed in brain receptor activity
              - Ⓘclass B orphan receptor activity
              - Ⓘclass C orphan receptor activity

We could annotate to this

May 18, 2005

GO Definitions

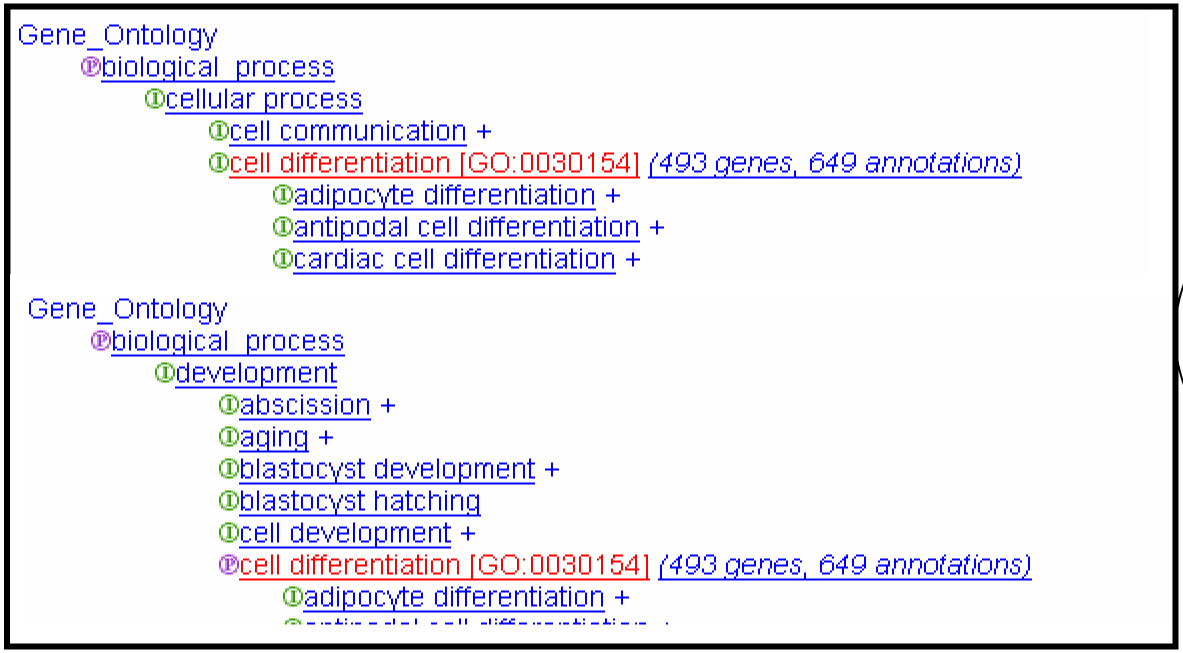


Gene Ontology Browser

Term Detail

GO term:	**cell differentiation**
GO id:	**GO:0030154**
Definition:	**The process whereby relatively unspecialized cells, e.g. embryonic or regenerative cells, acquire specialized structural and/or functional features that characterize the cells, tissues, or organs of the mature organism or some other relatively stable phase of the organism's life history.**

A definition written by a biologist:
necessary & sufficient conditions
written definition
(not computable)



- Gene_Ontology
  - Ⓟbiological_process
    - Ⓘcellular process
      - Ⓘcell communication +
      - Ⓘcell differentiation [GO:0030154] *(493 genes, 649 annotations)*
        - Ⓘadipocyte differentiation +
        - Ⓘantipodal cell differentiation +
        - Ⓘcardiac cell differentiation +

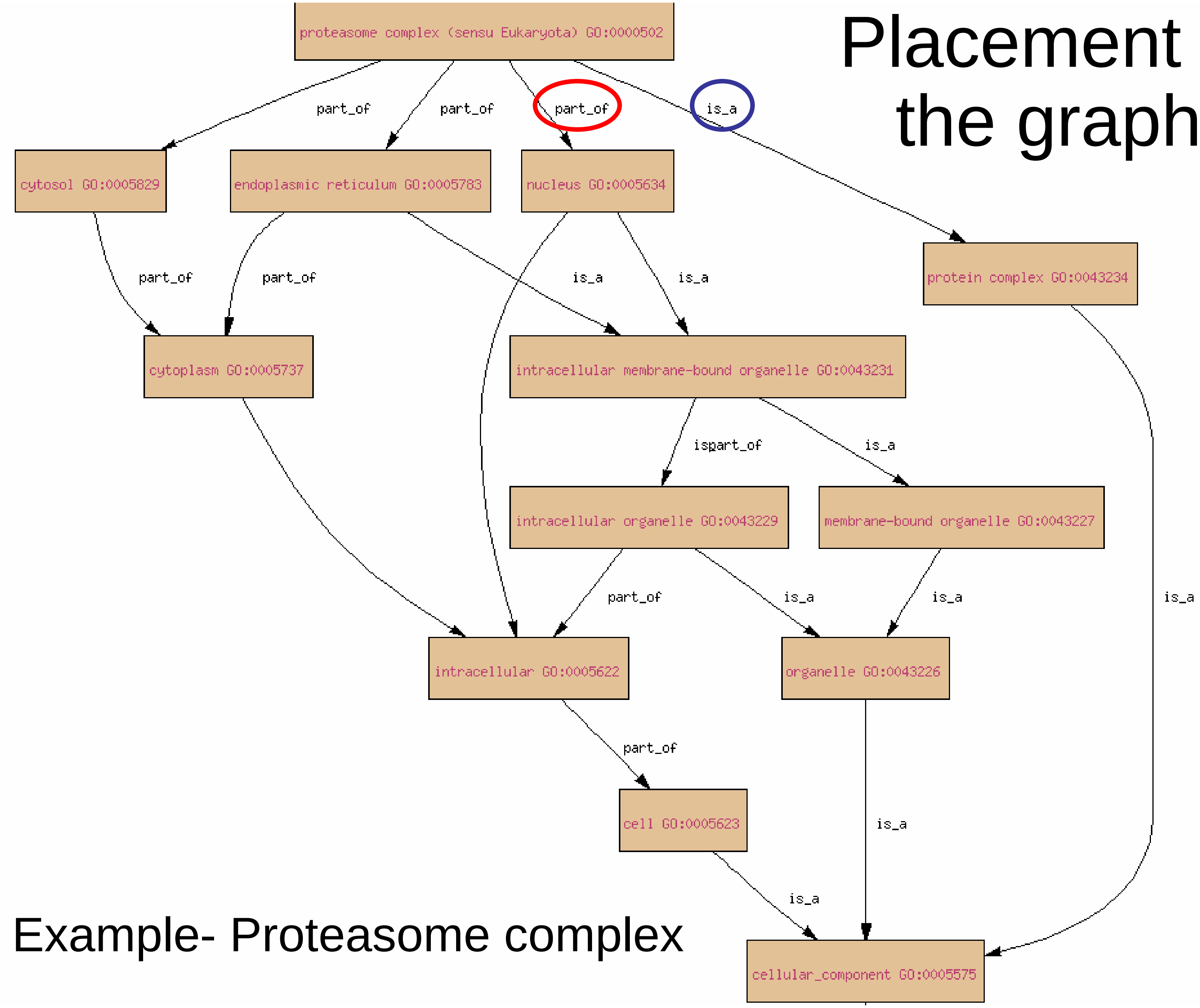
- Gene_Ontology
  - Ⓟbiological_process
    - Ⓘdevelopment
      - Ⓘabscission +
      - Ⓘaging +
      - Ⓘblastocyst development +
      - Ⓘblastocyst hatching
      - Ⓘcell development +
      - Ⓟcell differentiation [GO:0030154] *(493 genes, 649 annotations)*
        - Ⓘadipocyte differentiation +
        - Ⓘantipodal cell differentiation +

Graph structure:
necessary conditions
formal
(computable)

Relationships and definitions

- The set of *necessary conditions* is determined by the graph
    - This can be considered a *partial* definition
- Important considerations:
    - Placement in the graph- selecting parents
    - Appropriate relationships to different parents
    - True path violation

Placement in the graph

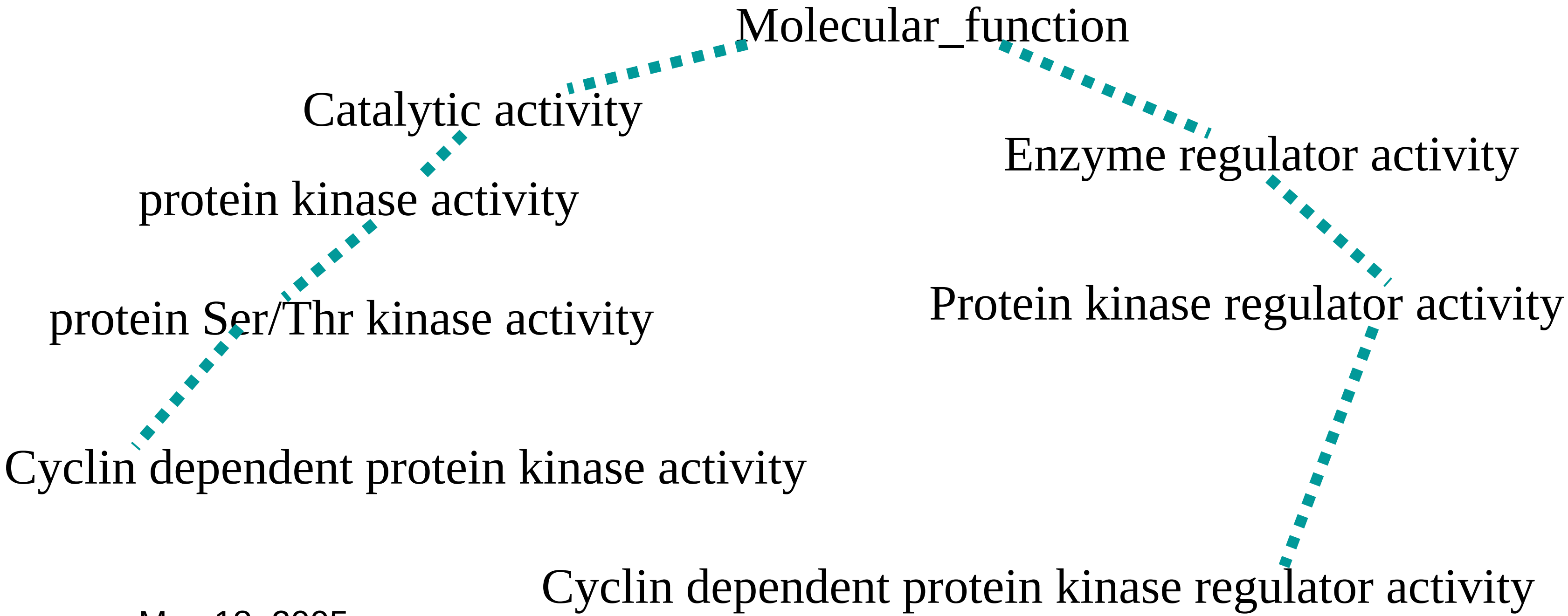


proteasome complex (sensu Eukaryota) GO:0000502
part_of
part_of
part_of
is_a
cytosol GO:0005829
endoplasmic reticulum GO:0005783
nucleus GO:0005634
protein complex GO:0043234
part_of
part_of
is_a
is_a
cytoplasm GO:0005737
intracellular membrane-bound organelle GO:0043231
ispart_of
is_a
intracellular organelle GO:0043229
membrane-bound organelle GO:0043227
part_of
is_a
is_a
is_a
intracellular GO:0005622
organelle GO:0043226
part_of
cell GO:0005623
is_a
is_a
cellular_component GO:0005575

- Example- Proteasome complex

The importance of relationships

- Cyclin dependent protein kinase
  - Complex has a catalytic and a regulatory subunit
  - How do we represent these activities (function) in the ontology?
  - Do we need a new relationship type (regulates)?



Molecular_function

Catalytic activity

Enzyme regulator activity

protein kinase activity

protein Ser/Thr kinase activity

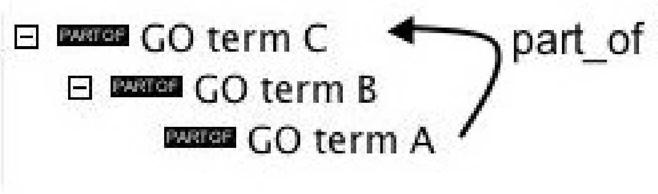
Protein kinase regulator activity

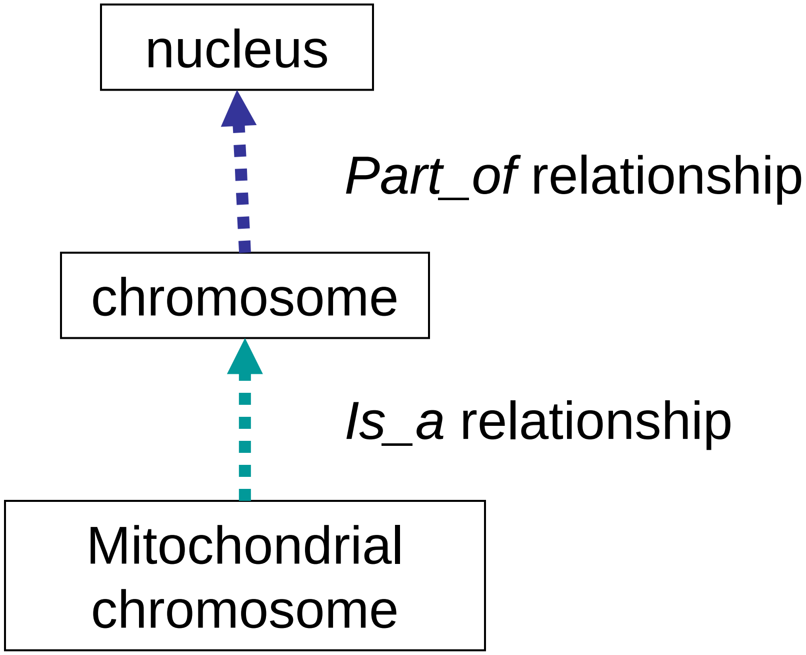
Cyclin dependent protein kinase activity

Cyclin dependent protein kinase regulator activity

True path violation
What is it?

..”the pathway from a child term all the way up to its top-level parent(s) must always be true".





nucleus

Part_of relationship

chromosome

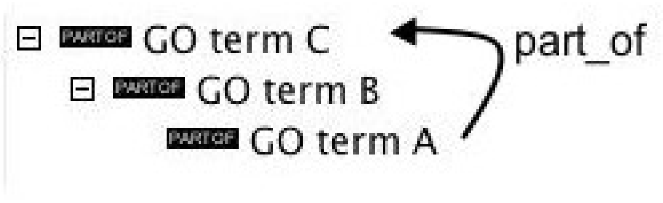
Is_a relationship

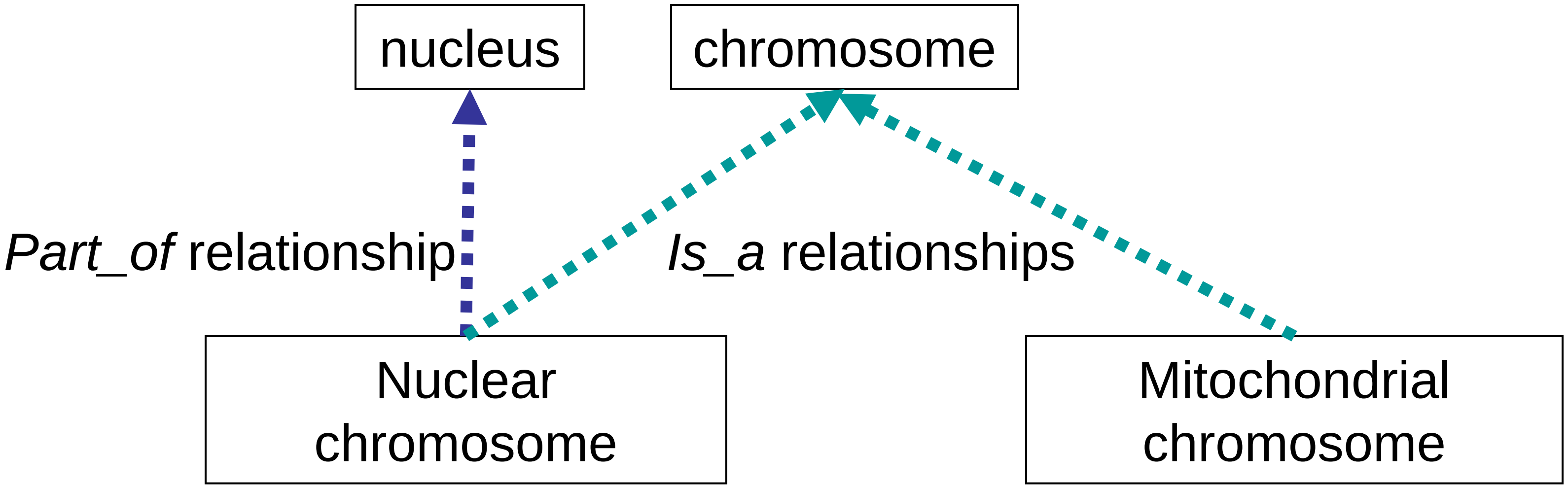
Mitochondrial chromosome

May 18, 2005

True path violation
What is it?

..”the pathway from a child term all the way up to its top-level parent(s) must always be true".





nucleus

chromosome

Part_of relationship

Is_a relationships

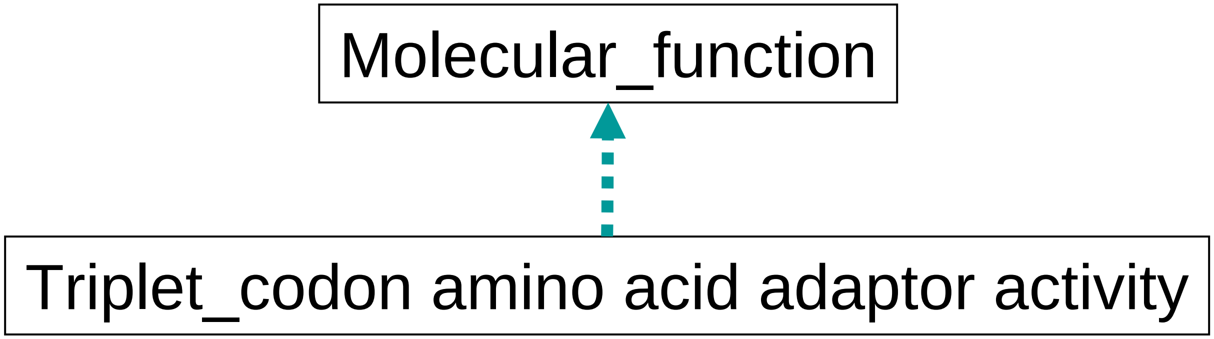
Nuclear chromosome

Mitochondrial chromosome

The Importance of synonyms for utility:
How do we represent the function of tRNA?

Biologically, what does the tRNA do?

Identifies the codon and inserts the amino acid in the growing polypeptide



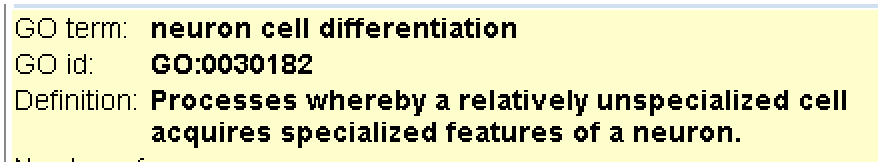
Molecular_function

Triplet_codon amino acid adaptor activity

GO Definition: Mediates the insertion of an amino acid at the correct point in the sequence of a nascent polypeptide chain during protein synthesis.

Synonym: tRNA

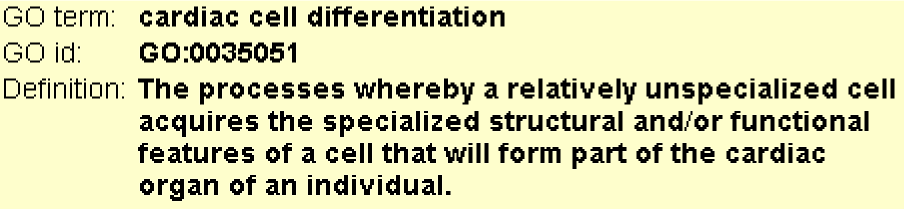
GO textual definitions: Related GO terms have similarly structured (normalized) definitions



GO term: **neuron cell differentiation**

GO id: **GO:0030182**

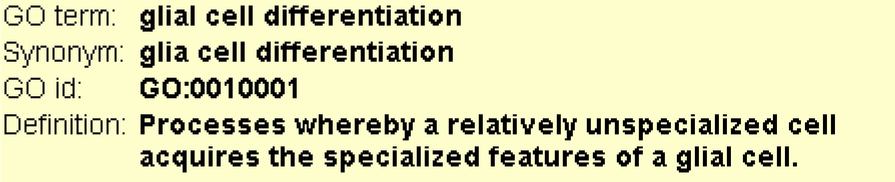
Definition: **Processes whereby a relatively unspecialized cell acquires specialized features of a neuron.**



GO term: **cardiac cell differentiation**

GO id: **GO:0035051**

Definition: **The processes whereby a relatively unspecialized cell acquires the specialized structural and/or functional features of a cell that will form part of the cardiac organ of an individual.**

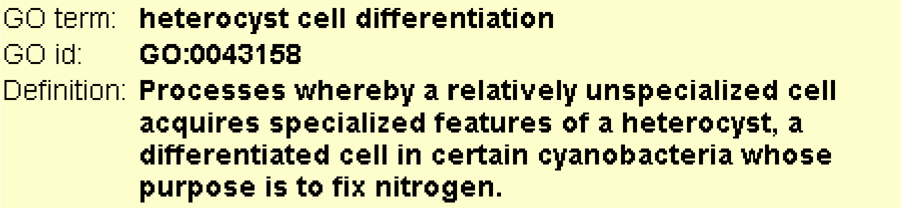


GO term: **glial cell differentiation**

Synonym: **glia cell differentiation**

GO id: **GO:0010001**

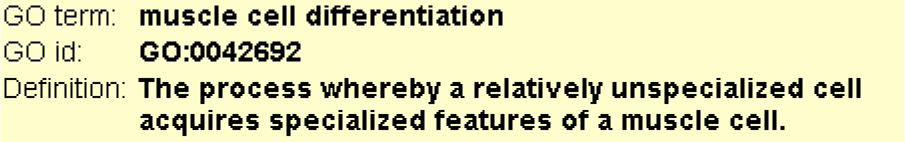
Definition: **Processes whereby a relatively unspecialized cell acquires the specialized features of a glial cell.**



GO term: **heterocyst cell differentiation**

GO id: **GO:0043158**

Definition: **Processes whereby a relatively unspecialized cell acquires specialized features of a heterocyst, a differentiated cell in certain cyanobacteria whose purpose is to fix nitrogen.**



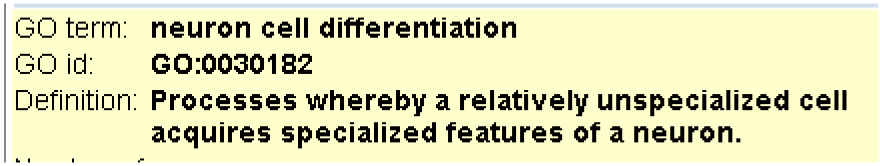
GO term: **muscle cell differentiation**

GO id: **GO:0042692**

Definition: **The process whereby a relatively unspecialized cell acquires specialized features of a muscle cell.**

May 18, 2005

Structured definitions contain both **genus** and **differentiae**



GO term: **neuron cell differentiation**
GO id: **GO:0030182**
Definition: **Processes whereby a relatively unspecialized cell acquires specialized features of a neuron.**

Essence = Genus + Differentiae

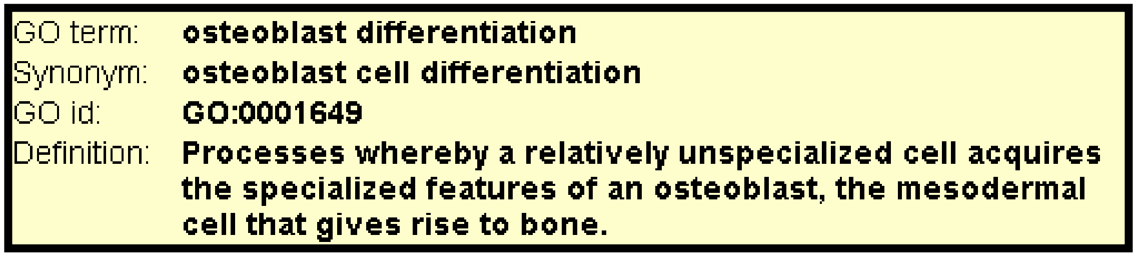
neuron cell differentiation =
Genus: **differentiation** (processes whereby a relatively unspecialized cell acquires the specialized features of..)
Differentiae: *acquires features of* a **neuron**

Ontology alignment

One of the current goals of GO is to align:

Cell Types in GO	with	**Cell Types in the Cell Ontology**
cone cell fate commitment	↔	retinal_cone_cell
keratinocyte differentiation	↔	keratinocyte
adipocyte differentiation	↔	fat_cell
dendritic cell activation	↔	dendritic_cell
lymphocyte proliferation	↔	lymphocyte
T-cell homeostasis	↔	T_lymphocyte
garland cell differentiation	↔	garland_cell
heterocyst cell differentiation	↔	heterocyst

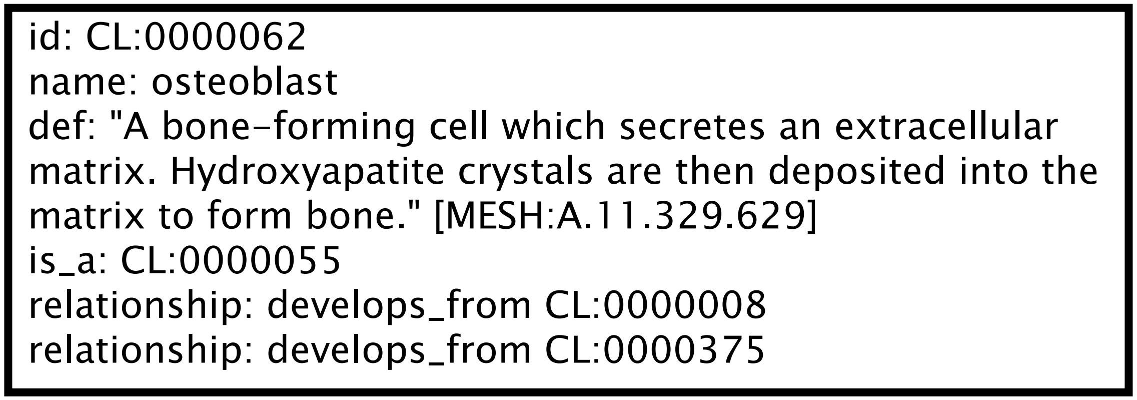
Alignment of the Two Ontologies will permit the generation of consistent and complete definitions



GO term:	**osteoblast differentiation**
Synonym:	**osteoblast cell differentiation**
GO id:	**GO:0001649**
Definition:	**Processes whereby a relatively unspecialized cell acquires the specialized features of an osteoblast, the mesodermal cell that gives rise to bone.**

GO

+



```
id: CL:0000062
name: osteoblast
def: "A bone-forming cell which secretes an extracellular
matrix. Hydroxyapatite crystals are then deposited into the
matrix to form bone." [MESH:A.11.329.629]
is_a: CL:0000055
relationship: develops_from CL:0000008
relationship: develops_from CL:0000375
```

Cell type

=

Osteoblast differentiation: Processes whereby an osteoprogenitor cell or a cranial neural crest cell acquires the specialized features of an osteoblast, a bone-forming cell which secretes extracellular matrix.

New Definition

Alignment of the Two Ontologies will permit the generation of consistent and complete definitions

id: GO:0001649

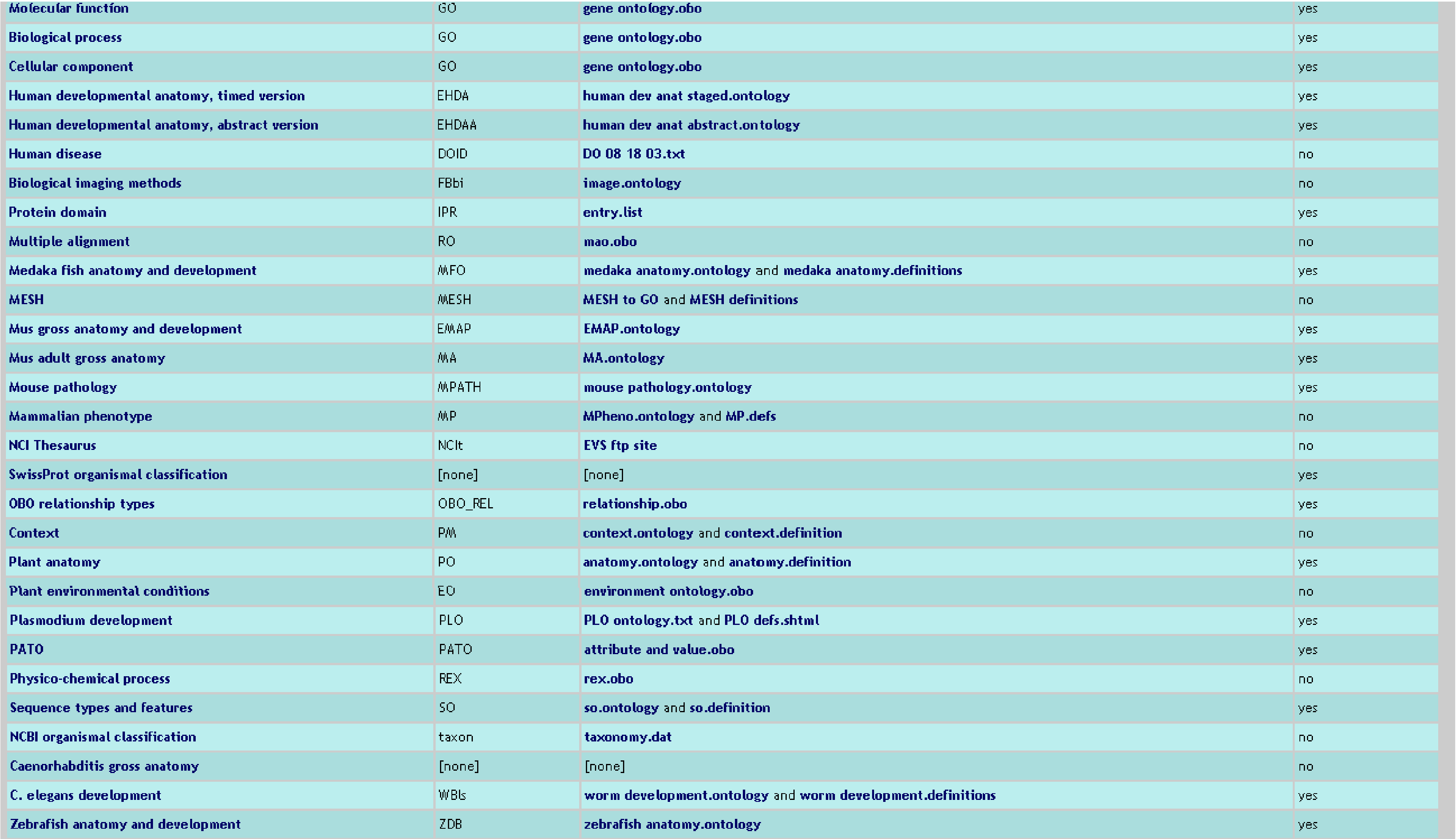
name: osteoblast differentiation
synonym: osteoblast cell differentiation
genus: differentiation GO:0030154 (differentiation)
differentium: *acquires_features_of* CL:0000062 (osteoblast)
definition (text): Processes whereby a relatively unspecialized cell acquires the specialized features of an osteoblast, the mesodermal cell that gives rise to bone

Formal definitions with necessary and sufficient conditions, in both human readable and computer readable forms

Other Ontologies that can be aligned with GO

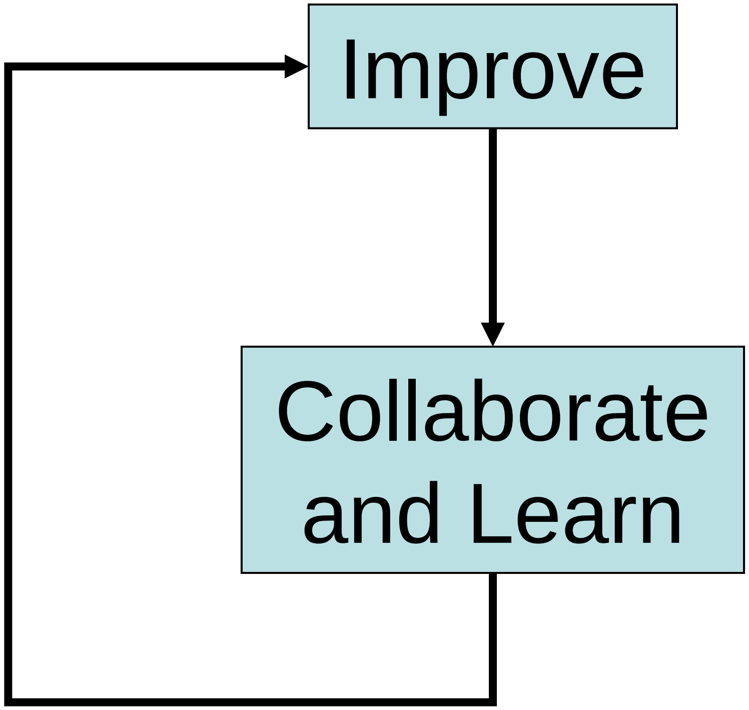
- Chemical ontologies
  - 3,4-dihydroxy-2-butanone-4-phosphate synthase activity
- Anatomy ontologies
  - metanephros development
- GO itself
  - mitochondrial inner membrane peptidase activity

But Eventually…



Molecular function	GO	gene ontology.obo	yes
Biological process	GO	gene ontology.obo	yes
Cellular component	GO	gene ontology.obo	yes
Human developmental anatomy, timed version	EHDA	human dev anat staged.ontology	yes
Human developmental anatomy, abstract version	EHDAA	human dev anat abstract.ontology	yes
Human disease	DOID	DO 08 18 03.txt	no
Biological imaging methods	FBbi	image.ontology	no
Protein domain	IPR	entry.list	yes
Multiple alignment	RO	mao.obo	no
Medaka fish anatomy and development	MFO	medaka anatomy.ontology and medaka anatomy.definitions	yes
MESH	MESH	MESH to GO and MESH definitions	no
Mus gross anatomy and development	EMAP	EMAP.ontology	yes
Mus adult gross anatomy	MA	MA.ontology	yes
Mouse pathology	MPATH	mouse pathology.ontology	yes
Mammalian phenotype	MP	MPheno.ontology and MP.defs	no
NCI Thesaurus	NCIt	EVS ftp site	no
SwissProt organismal classification	[none]	[none]	yes
OBO relationship types	OBO_REL	relationship.obo	yes
Context	PM	context.ontology and context.definition	no
Plant anatomy	PO	anatomy.ontology and anatomy.definition	yes
Plant environmental conditions	EO	environment ontology.obo	no
Plasmodium development	PLO	PLO ontology.txt and PLO defs.shtml	yes
PATO	PATO	attribute and value.obo	yes
Physico-chemical process	REX	rex.obo	no
Sequence types and features	SO	so.ontology and so.definition	yes
NCBI organismal classification	taxon	taxonomy.dat	no
Caenorhabditis gross anatomy	[none]	[none]	no
C. elegans development	WBls	worm development.ontology and worm development.definitions	yes
Zebrafish anatomy and development	ZDB	zebrafish anatomy.ontology	yes

Building Ontology



May 18, 2005