# Relations in pre and post composition

# Why not keep it simple?

- Can we not describe everything in biology by **bundling terms together**?
  - E.g.
    - development + lung
  - The meaning of this is fairly unambiguous, right?
    - So can't we do everything like this?

# No!

- Counterexample:
  - transport + nurse cell + oocyte + germline ring canal + Dcp1
- What's going on here?
  - transport of nurse cell to germline ring canal through a oocyte?
    - nope
  - transport of something from an oocyte to a nurse cell through a germline ring canal?
    - closer….

# We need to know the **role*** of each individual participant

- Example:
  - cargo
  - transporter (typically implicit, c1+2 in GAF)
  - start location
  - end location
  - conduit

***Note: I'm using the word in the plain everyday English sense, rather than a specific philosophical meaning, e.g. as in BFO**

# Current implementation (pre-composition)

```
[Term]
id: GO:0007300 ! ovarian nurse cell to oocyte transport
intersection_of: GO:0006810 ! transport
intersection_of: results_in_transport_from CL:0000026 ! nurse cell
intersection_of: results_in_transport_to CL:0000023 ! oocyte
intersection_of: results_in_transport_through GO:0045172 ! germline ring
canal
```

We use (somewhat wordy) relations to designate the role the participant plays.

- results_in_transport_from – start location
- results_in_transport_to – end location
- results_in_transport_through - conduit

We can call them what we want so long as we use them consistently

Christopher J. Mungall, Michael Bada, Tanya Z. Berardini, Jennifer Deegan, Amelia Ireland, Midori A. Harris, David P. Hill, and Jane Lomax. Cross-Product Extensions of the Gene Ontology. Journal of Biomedical Informatics 2010

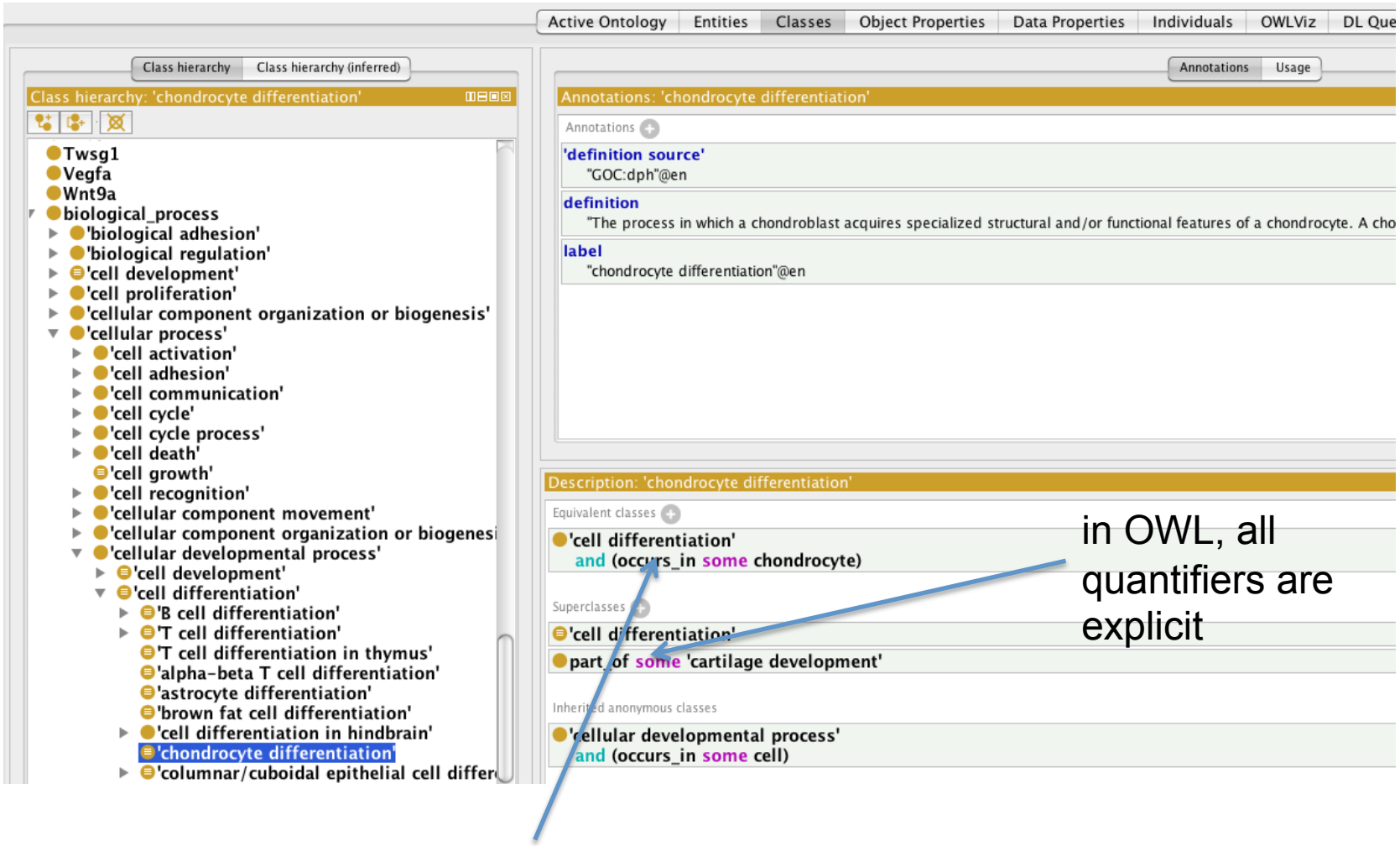
http://wiki.geneontology.org/index.php/XP:biological_process_xp_cell

# Fundamental principle

- If we use post-composition *instead* of pre-composed terms *without* explicitly designating the roles the participants play we **lose information** that would have been unambiguous in a pre-composed term

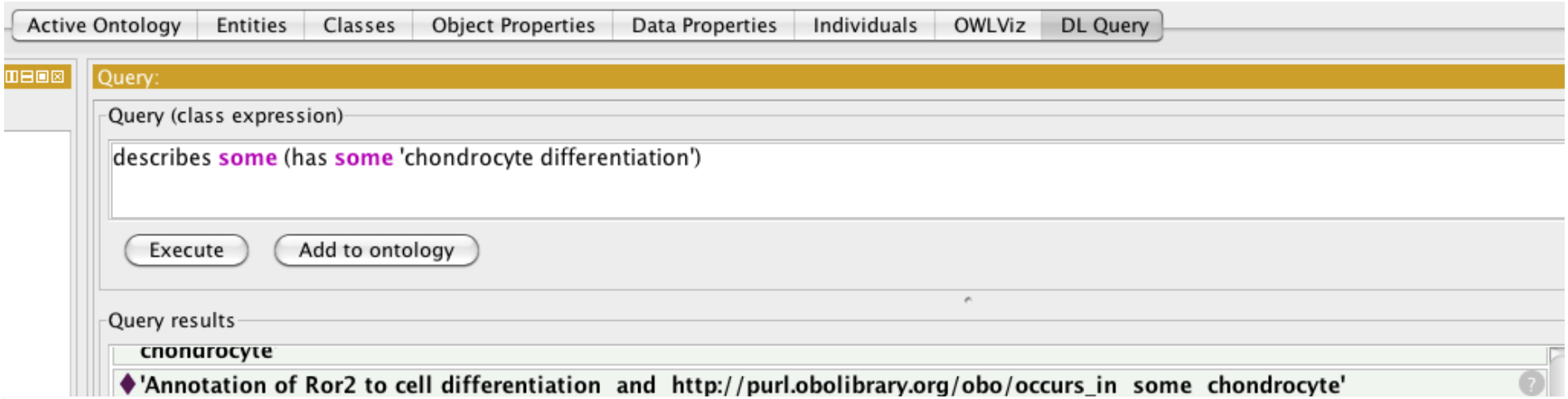
# Case study: interpreting MGI annotations

- MGI don't capture relation/role for c16
- I recommended using "occurs_in" as a default relation
  - this results in some complete nonsense…
    - e.g. mammary gland development that occurs in an epithelial cell
- Is this good enough?
  - defn of good enough: can a reasoner infer the most specific existing pre-coordinated term from post-composed annotation?
  - **this is testable!**




remember, we changed all these relations to be consistent with MGI c16

# Query for pre-coordinated term returns non-explicitly annotated pre-coordinated terms




TODO: test for false positives

# But isn't the role obvious most of the time?

- Yes – *most* of the time
  - E.g. development + lung
  - Can I not just say this?
- I agree the following is ugly, and looks to be redundant:

```
[Term]
id: GO:0030324 ! lung development
intersection_of: GO:0032502 ! developmental process
intersection_of: OBO_REL:results_in_complete_development_of UBERON:0002048 !
lung
```

# But isn't the role obvious most of the time?

- Yes – *most* of the time
  - E.g. development + lung
- Compromise:
  - We will provide a **table** with a **default** relation/role for a certain subset of GO terms
  - Annotators should always look up this table when making c16 annotations to make sure what they think they are saying is consistent with everyone else
    - (or done automatically with software)
  - This compromise is necessary to accommodate existing MGI structured notes

| GO ID | GO Term | Default Role for CL in c16 |
|---|---|---|
| GO:0055082 | cellular chemical homeostasis | location |
| GO:0001708 | cell fate specification | target state |
| GO:0001709 | cell fate determination | target state |
| GO:0045165 | cell fate commitment | target state |
| GO:0048468 | cell development | target state |
| GO:0030154 | cell differentiation | target state |
| GO:0001775 | cell activation | activated state |
| GO:0045058 | T cell selection | selected state |
| GO:0016049 | cell growth | grower |
| GO:0032940 | secretion by cell | secretor |
| GO:0012501 | programmed cell death | dier (ok, need a better name here) |
| GO:0001906 | cell killing | dier |
| GO:0002507 | tolerance induction | acted upon |
| GO:0050896 | response to stimulus | acted upon |
| GO:0046907 | intracellular transport | location |
| GO:0060326 | cell chemotaxis | cargo |
| GO:0051301 | cell division | parent state |
| GO:0052127 | movement on or near host | cargo |
| GO:0051674 | localization of cell | cargo |
| GO:0048469 | cell maturation | target state |
| GO:0043697 | cell dedifferentiation | initial state |
| GO:0031130 | creation of an inductive signal | signal originator |
| GO:0007267 | cell-cell signaling | signal transmitter |
| GO:0044237 | cellular metabolic process | location |
| GO:0007155 | cell adhesion | adherer |
| GO:0007166 | cell surface receptor linked signaling pathway | location |
| GO:0007166 | cell surface receptor linked signaling pathway | signal receiver |
| GO:0022403 | cell cycle phase | location |
| GO:0048870 | cell motility | cargo |

# obo representation

http://www.geneontology.org/scratch/xps/go_templates.obo

```
[Term]
id: GO:0055082 ! cellular chemical homeostasis
relationship: primary_cell_participant_role role:location

[Term]
id: GO:0001708 ! cell fate specification
relationship: primary_cell_participant_role role:target_state

[Term]
id: GO:0001709 ! cell fate determination
relationship: primary_cell_participant_role role:target_state

[Term]
id: GO:0045165 ! cell fate commitment
relationship: primary_cell_participant_role role:target_state

[Term]
id: GO:0048468 ! cell development
relationship: primary_cell_participant_role role:target_state

[Term]
id: GO:0001775 ! cell activation
relationship: primary_cell_participant_role role:activated_state

[Term]
id: GO:0016049 ! cell growth
relationship: primary_cell_participant_role role:grower
```

…

# Algorithm to determine meaning

- Allow '*has_primary_participant*' relation
  - this would be the default for all MGI c16s
- For any `has_primary_participant(CL:xxxx)` in c16 follow is_a* path up from term used in annotation to root
  - *we may have to allow across regulates too
  - but NOT part_of
- Any role encountered in lookup table applies
  - Each role corresponds to a relation
- Sometimes multiple roles can apply
  - E.g. following degranulation up leads to cell activation (default role: activated) and secretion by cell (default role: secretor)
    - THIS IS FINE: entity in c16 carries out both these roles

# Processes involving two or more cells or cell types

- E.g. axon guidance
  - here we designate ‘source’ as the default role
  - if you want to say which cell type the axon is being guided towards (target/destination), you MUST explicitly designate this

# Remaining problems interpreting MGI c16s

- The above procedure should allow us to use the majority of MGI c16s
- Issues remaining
  - c5: ‘visual learning’ c16: ‘pyramidal cell’
  - intended meaning: “pyramidal cell process involved in visual learning”
    - formally:
      - process and has_participant some pyramidal_cell and part_of some visual_learning

# Beyond cell types: gross anatomy, cell components

- c5: ‘mammary gland development’
- c16: ‘epithelial cell’

# Gene products and complexes as targets

- Similar principles apply
- most of the time the default role (aka '**target**') is obvious and can be specified in advance
  - E.g. binding
- The gene product in c1+2 always plays the role of 'agent' or 'active participant'
  - extend GAF to allow other roles?

# Multiple molecular participants

- signal transduction

# Conclusions/opinions

- Post-composition is not a cure for all ills
  - experience* shows that frequent problems are: ambiguity, creativity..
- Post-composition should never **replace** pre-composition **unless**:
  - The semantics of post-composition are formally specified
  - The relations used are closely coordinated with the logical definitions used in the main ontology
  - Modern ontology software is used in the interface
- (even loosely specified) post-composition is fine to **enhance** pre-composition
  - danger: annotators will elect to use loose post-composition instead of existing pre-coordinated terms