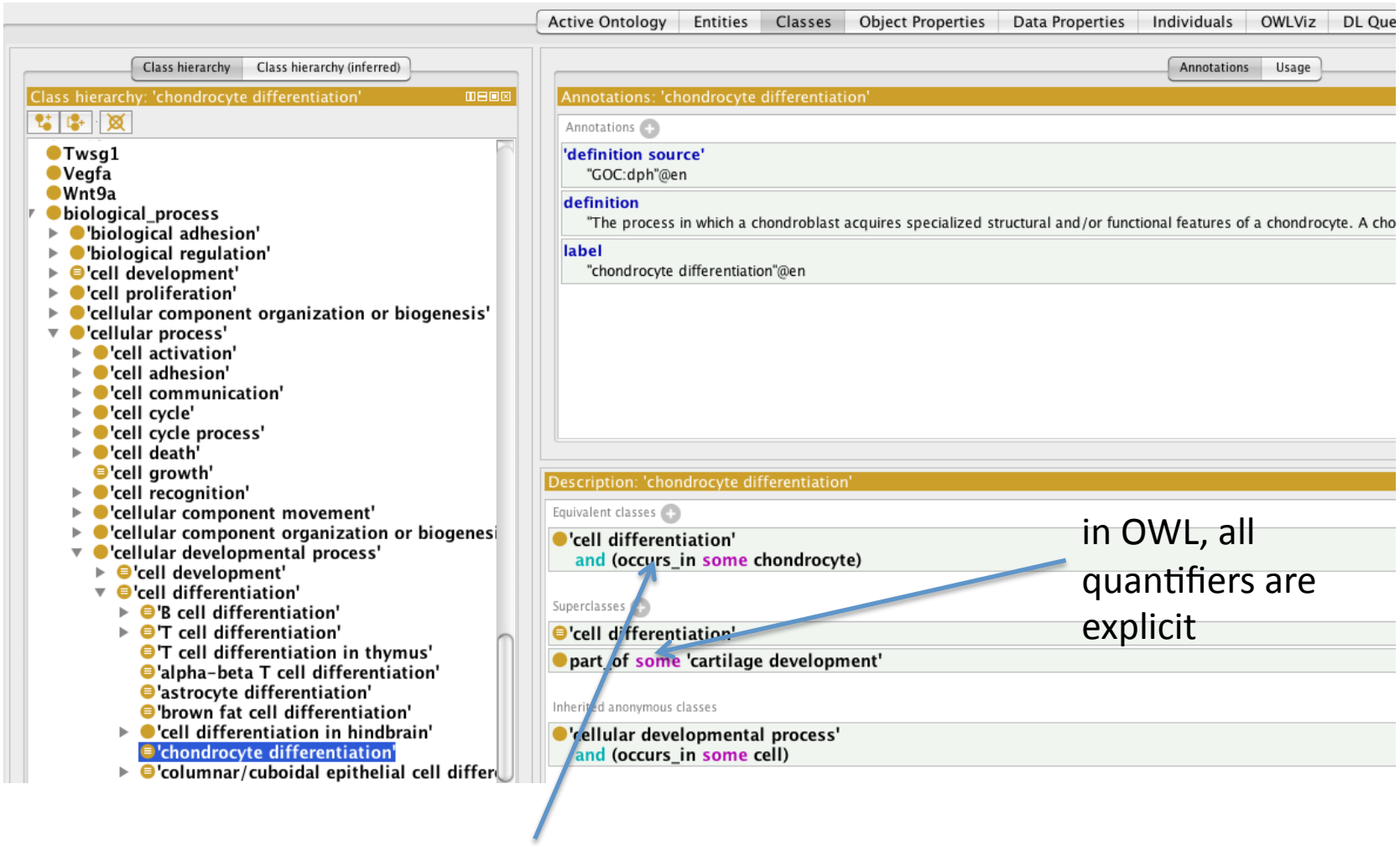
Annotation Extension (col 16)

- In use by MGI
  - cell types
- Ready to roll:
  - GOA
  - Pombe
- Relations used in c16
  - annotators prefer fewer, more generic relations

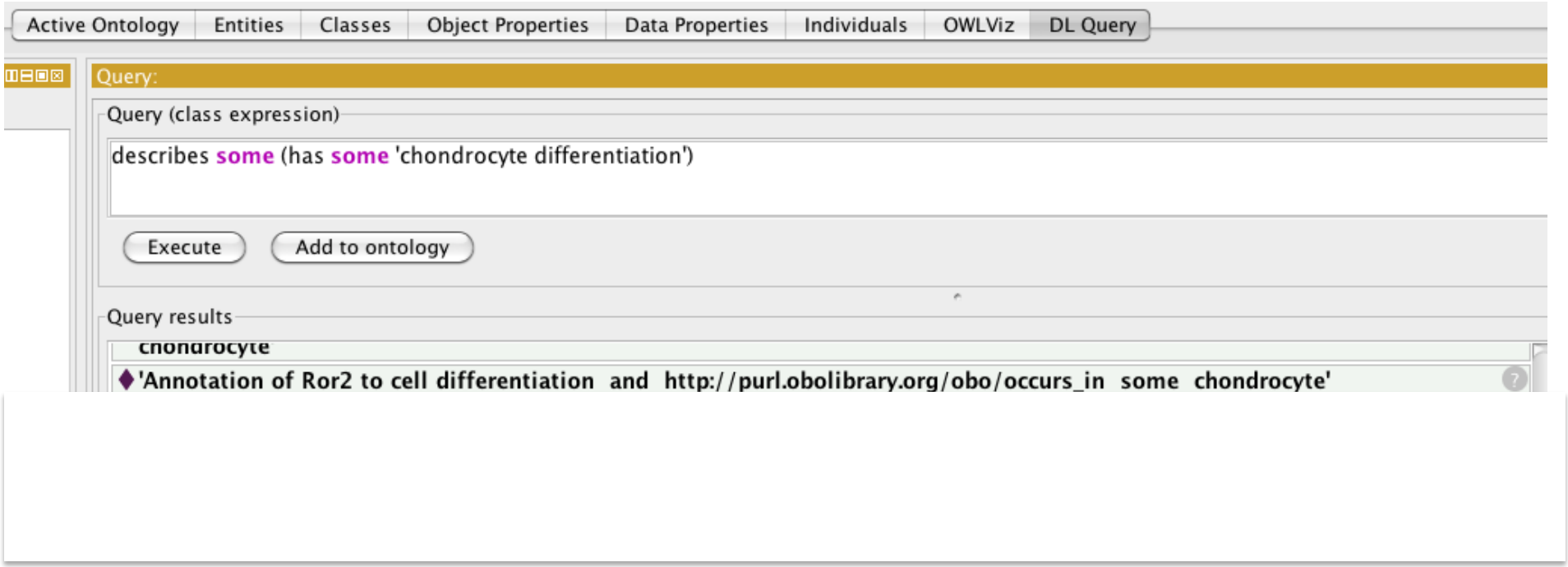
Case study: interpreting MGI annotations

- All c16 info currently comes from MGI
- MGI **don't** internally capture relation/role for c16
  - mapped to "occurs_in" for GAF.20
- Mixture of practice
  - e.g.
    - differentiation + chondrocyte
    - "chondrocyte differentiation"
    - "chondrocyte differentiation" + chondrocyte
  - we would like to infer these are mutually equivalent
  - we would also like to infer that differentiation + "columnar chondrocyte" is a subclass of "chondrocyte differentiation"
  - done using OWL reasoners in web GAF Tools



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

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



Missing relations in MGI structured notes

- MGI **don't** capture relation/role for c16
  - this is required for reasoning
- I previously recommended using "occurs_in" as a default uniform relation
  - this results in some nonsense…
    - e.g. **mammary gland development** that *occurs in* an **epithelial cell**
  - change to generic *has_primary_participant* relation and use lookup table to determine meaning

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

Algorithm to determine meaning for generic relation

- Allow generic '*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

Interpretation problems

- MGI MGI:96552 Il3 GO:0070668 MGI:MGI:3605581|PMID:15947484 IDA P interleukin 3 BPA| Csfmu|HCGF|Il-3|MCGF|PSF protein taxon:10090 20090605 MGI occurs_in(CL:0000084) VEGA:OTTMUSP00000005831

- GO:0070668 =
  - positive regulation of mast cell proliferation
- CL:0000084 =
  - T cell

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

Recommendation for cell types in c16

- Can we fix MGI internal representation?
- Core set:
  - occurs_in
  - has_primary_participant
    - use wisely in conjunction with lookup table
    - more work required
    - usually better to request a pre-coordinated term
      - TG template
- Extend to anatomy
  - clade-specific AOs or Uberon

Next steps for cell types in c16

- Querying in AmiGO2 and QuickGO
  - demo from Seth
- Automatic deepening of annotations
  - integrate into web GAF tools
  - requires further work on bp_xl_cl (Terry will be working on this)
- Dynamic grouping classes in term enrichment

Gene products and complexes in c16 for BPs

- What are the different roles a gene product or complex can play in a process?
  - active participant
    - this is the implicit role of the annotated gene product or complex
  - 'target'
    - "target" can mean different things in different contexts
    - Original proposal:
      - has_input
      - has_output
      - has_mediator

has_target

- What is the role of the target?

Base term	relation	role of target
kinase activity	has_target	phosphorylated
(regulation of) nucleic acid binding transcription factor activity	has_target	transcribed gene or gene product
(regulation of) protein binding transcription factor activity	has_target	transcribed gene or gene product
binding	has_target	binding partner
establishment of localization	has_target	cargo
establishment of localization	results_in_transport_from	start (pre-coordination recommended)
establishment of localization	results_in_transport_to	end (pre-coordination recommended)

Discussion

- 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-aware software is used in the curator interface and to infer annotations
- (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

Full-on post-composition

- c16 extends existing GAF model
  - can be used for nested expressions, but awkward
- Proposal:
  - develop LEGO application and tool chain in parallel
    - Use TG as a starting point
    - Underlying model is OWL
    - Demonstrate term enrichment with nested post-composition
  - Break backwards compatiblity with GAFs
    - can provide lossy translation

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

c16

- Syntax:
  - Relation(Entity)

- Base set of relations still under discussion
- Any used in go_ext / TG are fine:
  - part_of
  - occurs_in (May 2011)

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

- Yes – *most* of the time
  - E.g. development + lung
  - Or: binding + gene product
  - Or: regulation of transcription + gene product
- 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