# Semantic similarity

- Definition
  - A *profile* is a set of phenotypes associated with a genotype or an evolutionary lineage
- Motivation: enabling use of PhenoscapeKB for discovery of similar profiles between
  - genotypes (e.g. across models)
  - lineages
  - genotypes and lineages (across knowledge domains)
- Objectives: computational machinery for finding similar profiles
  - Biologically informative similarity measure(s)
  - Enabling search in, or among, large collections of profiles

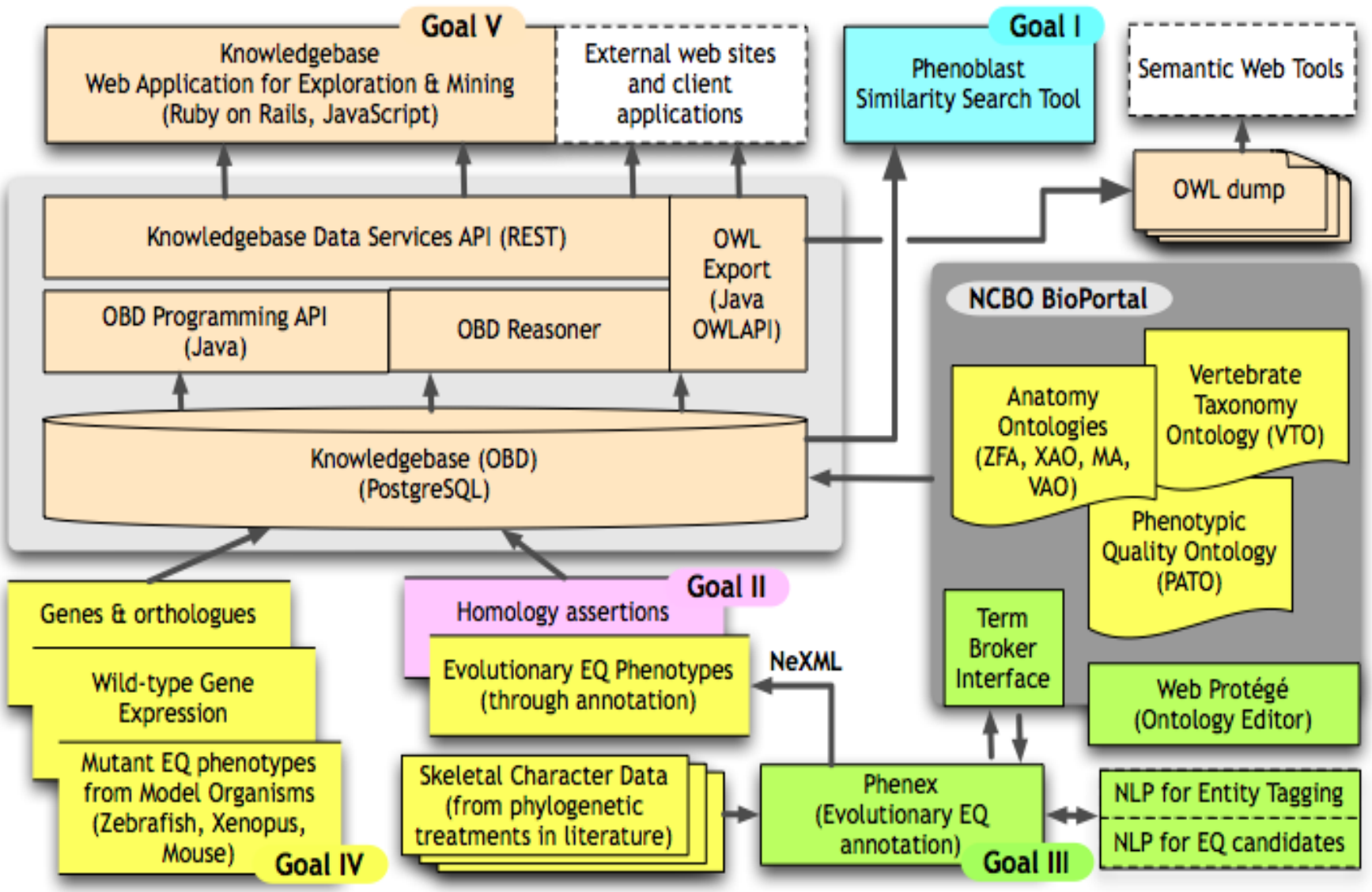
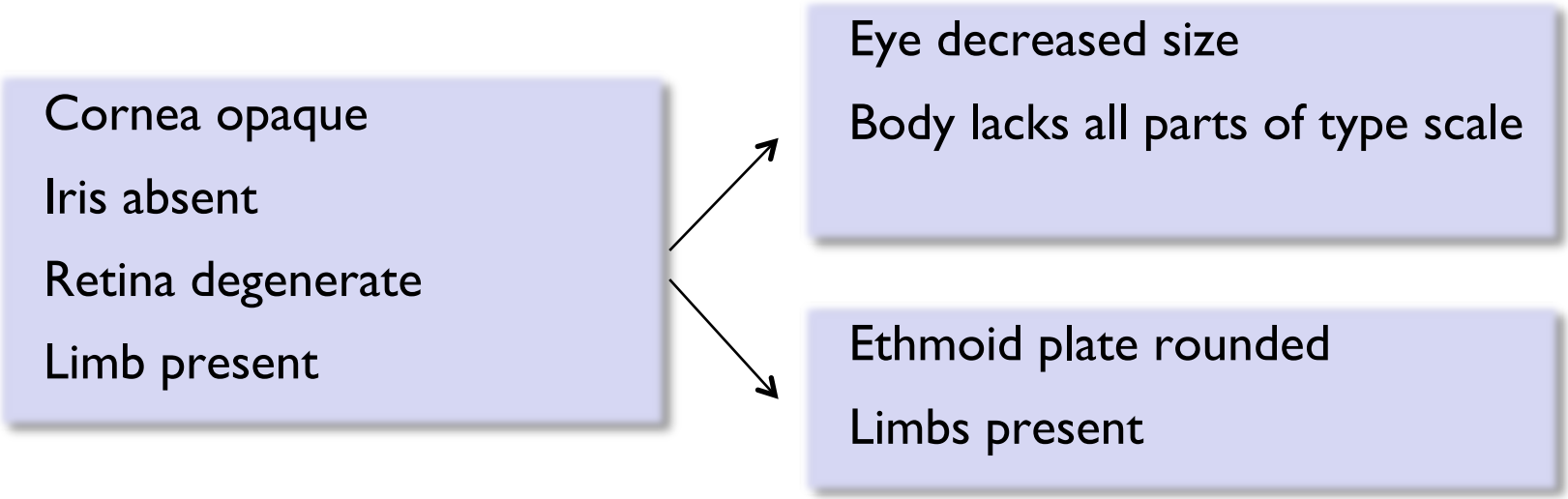


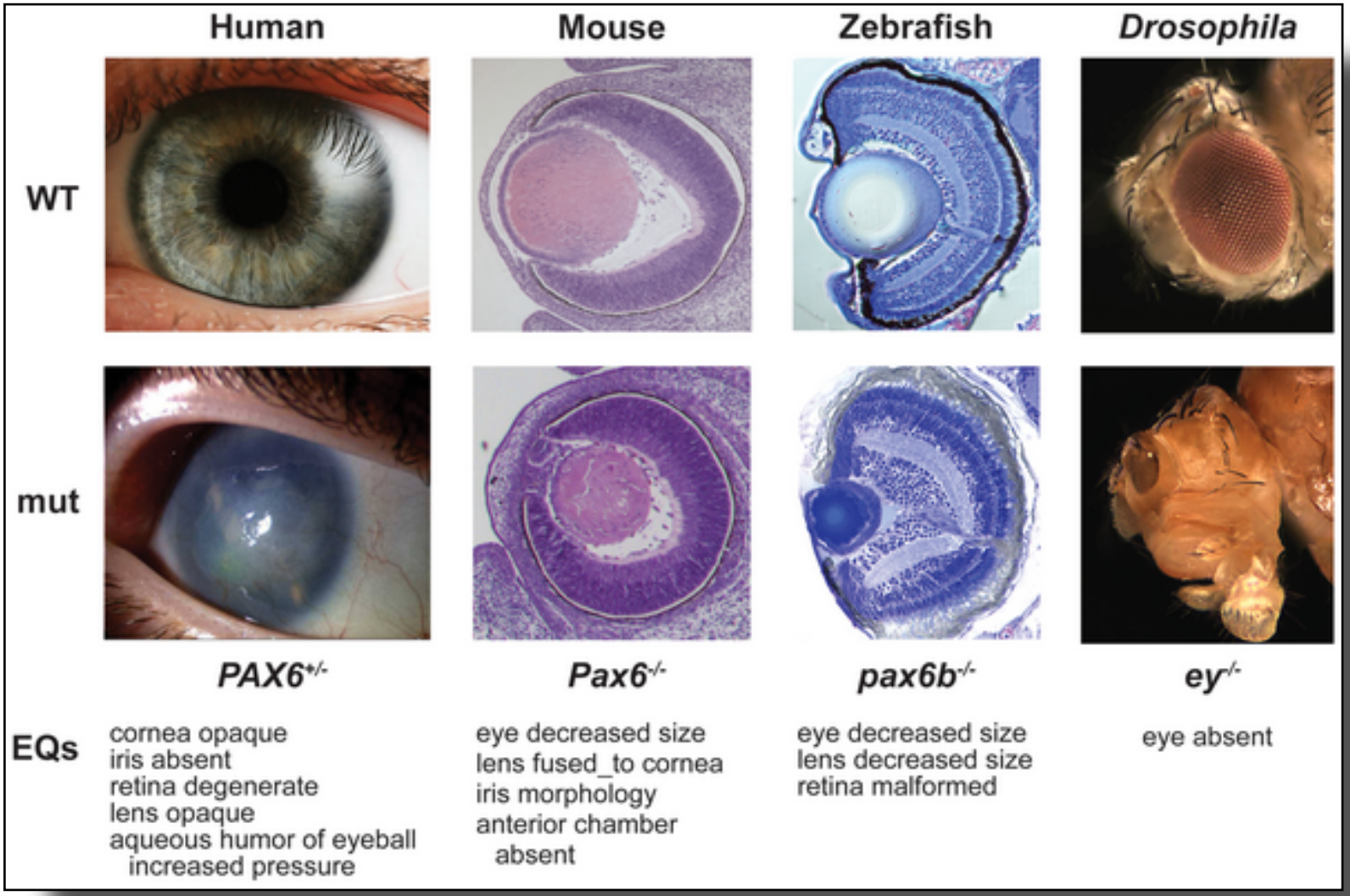

**Fig. 3.** Diagram of the architecture, with components color-coded by goals I-V.

# Accounting for ontology structure, information content, and messy biology




- Terms need not necessarily match lexically
- There need not be a one-to-one match between phenotypes, for biological and methodological reasons
- Similarity between common phenotypes is less informative
- A match in quality alone is not meaningful

# Profiles




Washington et al. 2009

# Semantic similarity

- Participants
  - Balhoff, Blake, Mungall, Lapp, Mabee, Midford, Vision, postdoc TBD
- Progress to date
  - Building on prior work by Washington, Mungall, et al
  - Some work on Phenoscape I dataset
  - Collected literature, started postdoc recruitment

# Immediate future plans

- Deploy measures implemented in OBD on Phenoscape I dataset
- Explore ideas for (hopefully faster) set overlap measures, mindful of
  - Biological interpretation of similarity ranking and/or probability
  - Algorithm complexity/scalability
  - Ability to precompute input data structure
- Recruit postdoc
- Develop evaluation process, test datasets and demonstration project (e.g. visualization of profile clusters)

# Challenges

- We know of deficiencies with existing measures
  - But there is no guarantee methods can be developed that will give better rankings, scale, be portable, etc.
- We need to keep up with, and be open to potentially adapting, ideas being developed in parallel elsewhere
  - Similarly motivated work within GO and PATO communities
  - Superficially different research in semantic web community

# Capstone

- How often are genes known to be involved in fin-limb transition retrieved by the system?
  - Known genes involved in limb growth and patterning: *Bmps, Fgfs, Gdf5, Sox9*
  - Raises both phylogenetic and serial homology issues
  - Focuses annotation: Important to generate a large haystack within which to search for needles
- Suggest we flesh out capstone plans some this week…
  - Analyses for capstone may constrain current work in important ways, needs to be better defined
  - If it includes analysis of non fin-limb phenotypes (e.g. reduction of the hyomandicula from jaw to ear) or different datatypes (e.g. expression), it may affect annotation effort