Linking semantic phenotypes to character matrices and specimens

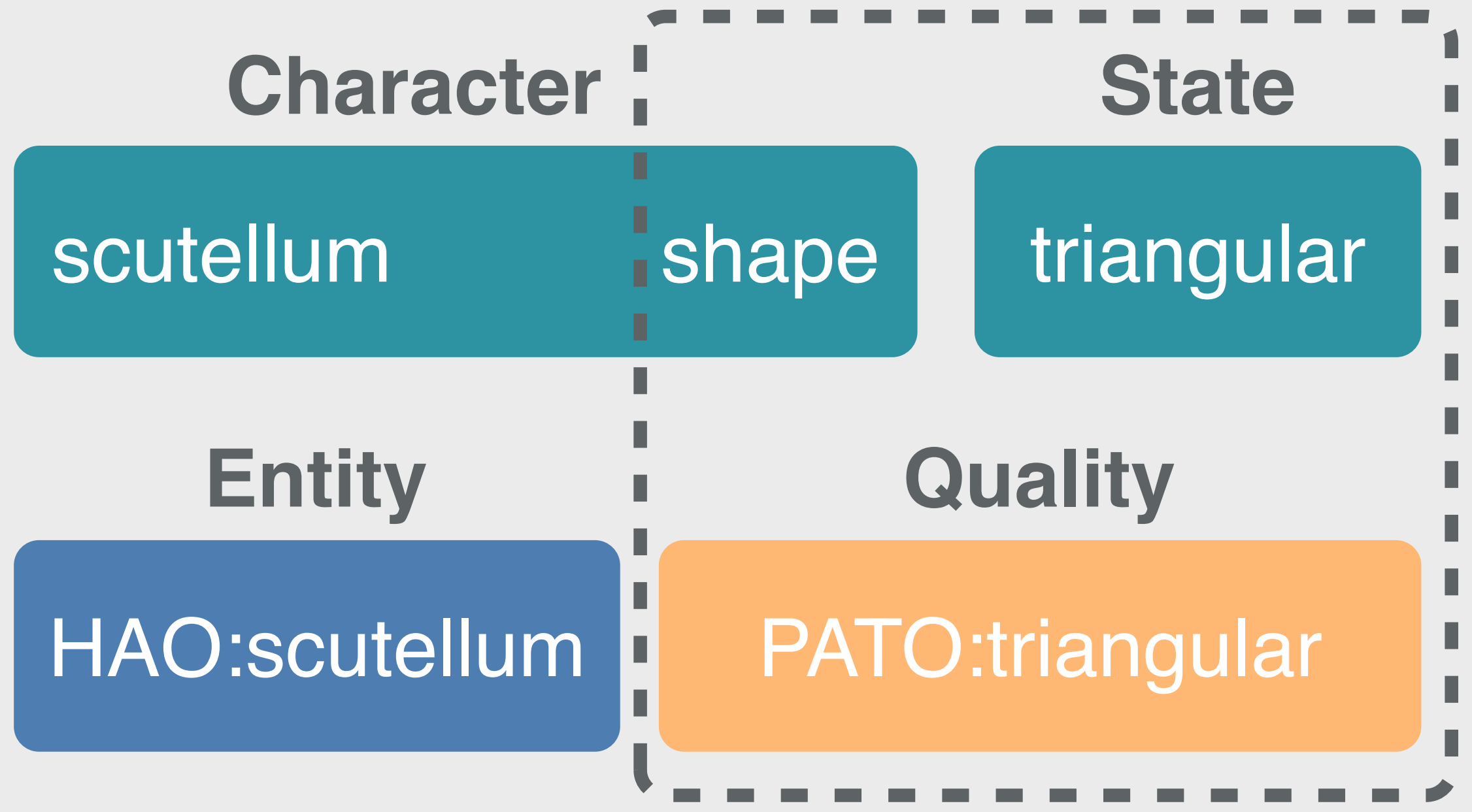
James P. Balhoff[1,2], Matthew J. Yoder[3], Andrew R. Deans[3]

[1]National Evolutionary Synthesis Center, Durham, NC; [2]University of North Carolina, Chapel Hill, NC; [3]North Carolina State University, Raleigh, NC

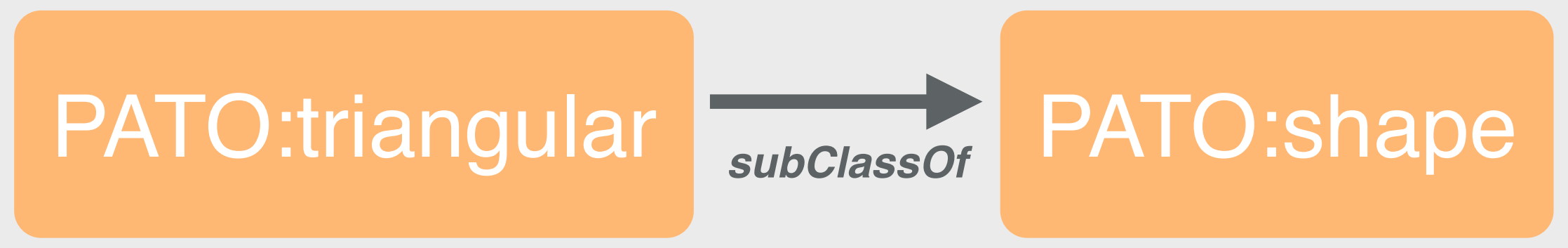
Background

Phenotype descriptions documented in the large body of published systematic biology literature are traditionally reported in a free-text format. As a consequence, they are largely inaccessible to computational methods for large-scale integrative analysis, including even seemingly basic steps such as linking them to biological knowledge maintained in databases for genetics, development, and other domains. Ontologies have become a foundational technology for establishing shared semantics, and, more generally, for capturing and computing with biological knowledge. Using the Web Ontology Language (OWL), we present a coherent semantic model combining free-text character matrix data, specimen metadata, and Entity–Quality phenotype descriptions.Our OWL framework provides consistent semantics for phenotypic descriptions across both the published literature annotated by the Phenoscape project as well as new semantics-based taxonomic descriptions being developed by the Hymenoptera Anatomy Ontology project.

Entity–Quality model

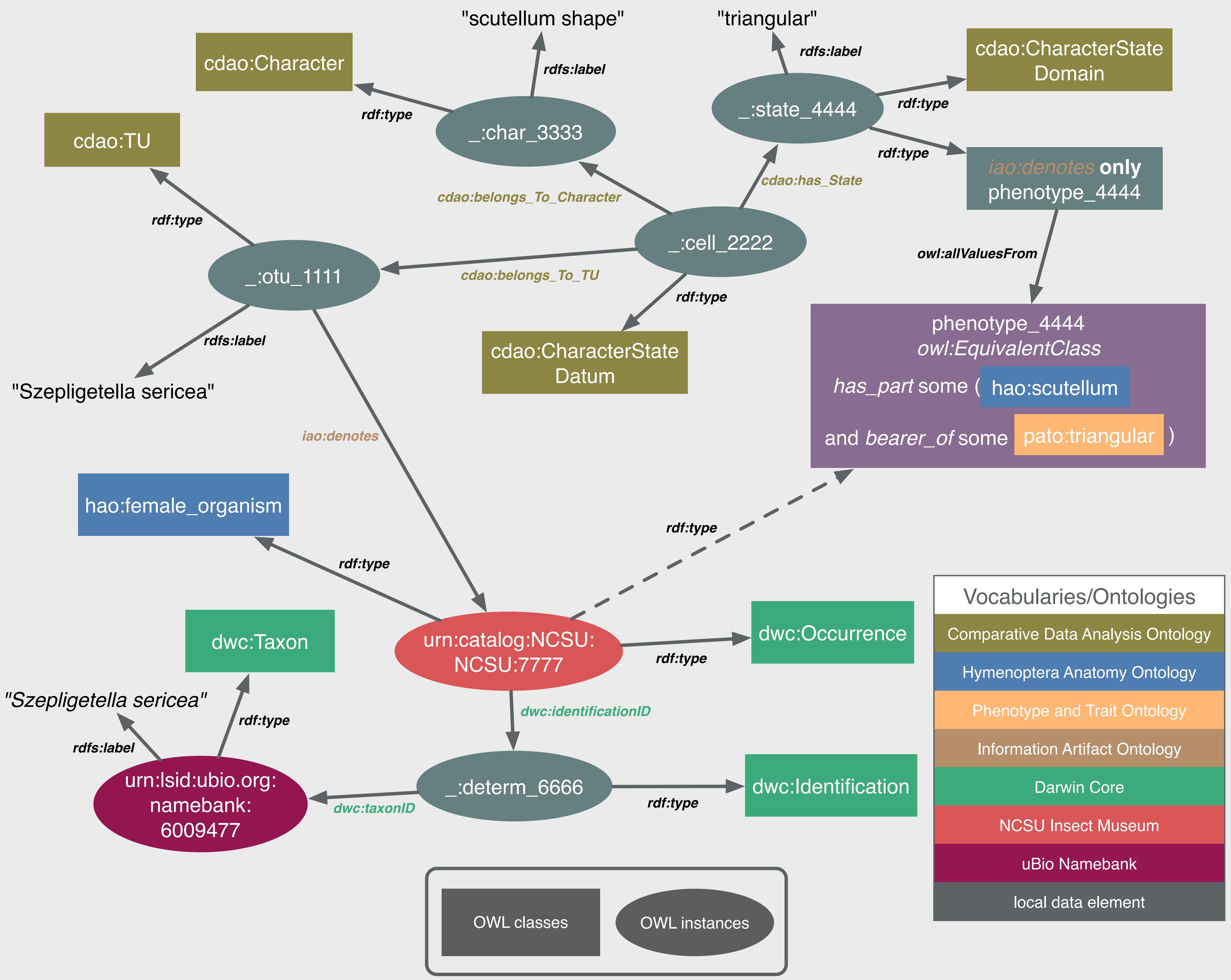


Free-text phenotypic character–character state descriptions, as found in evolutionary matrix data, can be mapped to the Entity–Quality semantic model. EQ associates an entity term drawn from an organism-specific anatomical ontology with a quality term from the generic Phenotype and Trait Ontology (PATO).



The varying attribute “shape” is redundant in the EQ model because this information is provided via the semantic structure of the PATO ontology.

Semantic model of characters, phenotypes, and specimens

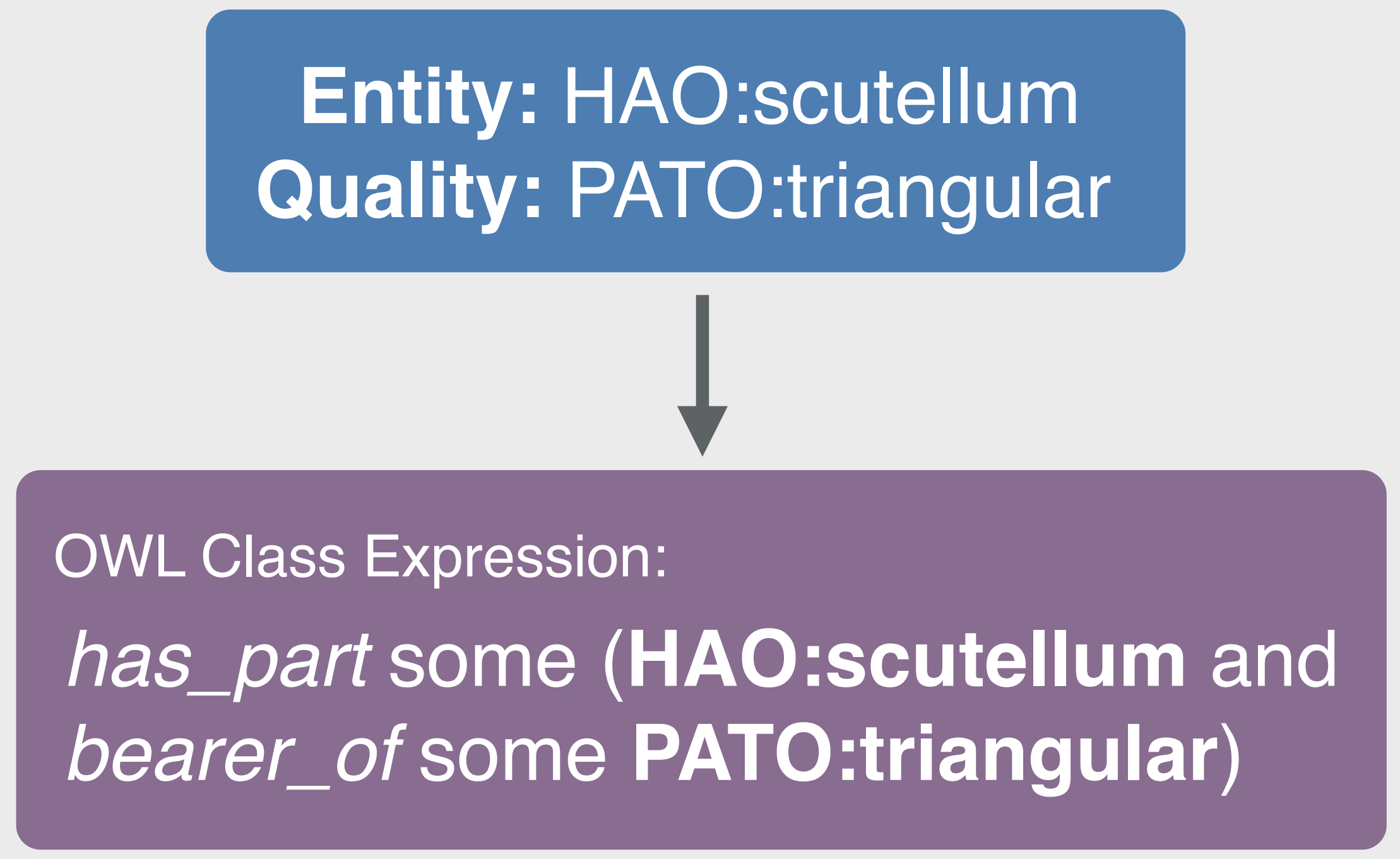


An OWL/RDF model depicting a single character matrix cell (_:cell_2222), represented using the **Comparative Data Analysis Ontology (CDAO)**, upper half, linked to a single museum specimen (urn:catalog:NCSU:NCSU:7777) described with the **Darwin Core** vocabulary, lower half. An Entity–Quality representation of the phenotype denoted by the given character state has been composed using terms from the **Hymenoptera Anatomy Ontology (HAO)** and the **Phenotype and Trait Ontology (PATO)**. The ***denotes*** property, from the **Information Artifact Ontology (IAO)**, is used to bridge observational data artifacts (CDAO data elements) to direct descriptions of organisms (as EQ phenotypes).

By applying an OWL 2 DL reasoner to the character matrix model, we can infer phenotypic characteristics of associated specimens (dashed arrow) using an asserted property chain:

inverse(*cdao:has_State*) **o** *cdao:belongs_to_TU* **o** *iao:denotes* → *iao:denotes*

EQ in OWL



In order to link the EQ conceptual model to other data, we must provide it with explicit semantics. An EQ phenotype is represented as an OWL class expression providing an axiomatic definition of a set of organisms possessing the described feature.

Conclusions

- Free-text character matrix data, EQ phenotype descriptions, and Darwin Core specimen metadata can be combined in one coherent OWL model.
- Following the Information Artifact Ontology, models of data can be bridged to models of organisms via properties such as *denotes*.
- An OWL specification for Darwin Core terms would be welcome and would ensure consistent usage of Darwin Core within semantic models.

Acknowledgments

We thank NSF DBI-0850223 to Andy Deans, and also the National Evolutionary Synthesis Center (NESCent) NSF EF-0423641. This work builds on close collaboration with the Phenoscape project, in particular Hilmar Lapp.

See also:
- http://hymao.org/
- http://www.phenoscape.org/
- http://kb.phenoscape.org/