PCS for Phylogenetic Systematic Literature Curation

Background

The Phenoscape project (http://www.phenoscape.org) seeks to extract systematic character data from the evolutionary literature. Systematic characters consist of two or more character states contrasting some aspect of phenotype, such as anatomy or behavior, of the taxa under study. The original use of these data is most often to recover phylogenetic relationships in the absence of, or together with, molecular data. However, such data have considerable reuse value for studying patterns of phenotypic evolution in a comparative context, *e.g.* on a given phylogeny, and linking data on natural phenotypic diversity with data on the roles of genes in the development of phenotypes in model organisms.

Phenex (Balhoff et al, 2010) is an interactive platform-independent desktop application designed to facilitate effective and consistent annotation of such data. It has been used in the Phenoscape project for several years and proven to be effective and user-friendly in supporting a manual annotation workflow. In this workflow (Dahdul et al, 2010), curators are required to parse the free text and search a set of ontologies to select the appropriate terms to annotate the characters using the Entity–Quality (EQ) model (Mabee et al, 2007) (Table 1). If there are terms that are not covered by existing ontologies, the curators must wait for new terms to be added before they can complete the curation task. These time-consuming steps in the curation workflow prevent the efficient scaling of curation to increased phenotypic diversity.

CharaParser (Cui, in press) is a text mining system that semi-automatically identifies candidate entity and quality terms in free-text character narratives and generates candidate EQ expressions for review by the human curator. The Phenoscape NLP work group (i.e., this team) is currently working to integrate CharaPaser into Phenex to produce a complete system with improved efficiency while retaining the rich and user-friendly features of the original Phenex system. For the time being, we shall call the complete system **Phenoscape Curation System (PCS)**.

Table 1: Examples of Systematic Character Narratives and EQ Statements

Systematic Character	**EQ using terms**		**EQ using term IDs**	
	Entity	**Quality**	**Entity**	**Quality**
First dorsal-fin rays (1) deeply branched (2) unbranched	dorsal fin lepidotrichium	branched	TAO:0001418	PATO:0000402
	dorsal fin lepidotrichium dorsal fin lepidotrichium	unbranched	TAO:0001418	PATO:0000414
Inner dentary tooth row (1) absent (2) present	dentary tooth row* [inner dentary tooth row]	count* [absent]	TAO:0001952	PATO:0000070
	dentary tooth row* [inner dentary tooth row]	count* [present]	TAO:0001952	PATO:0000070
Dentary (1) lower surface of dentary posterior to symphysis without any conspicuous notch	ventral margin and (part_of only dentary) and (posterior_to only mandibular symphysis)	shape* [not notched]	BSPO:0000684⊓ (OBO_REL:part_of ∀ TAO:0000191) ⊓ (BSPO:0000099 ∀TAO:0001851)	PATO:0000052
(2) a notch along lower border of the dentary just posterior to the convoluted symphysis	ventral margin and (part_of only dentary) and (posterior_to only mandibular symphysis)	notched	BSPO:0000684⊓ (OBO_REL:part_of ∀ TAO:0000191) ⊓ (BSPO:0000099 ∀ TAO:0001851)	PATO:0001495

The input to PCS is a set of articles from the phylogenetic systematic literature. Within each article is a semi-structured narrative of multiple systematic characters similar to those shown in Table 1. The output is a list of EQ statements that represent the original systematic characters. An EQ statement associates an entity term drawn from an organism-specific anatomy or process ontology such as Teleost Anatomy Ontology (TAO: http://obofoundry.org/cgi-bin/detail.cgi?id=teleost_anatomy) and Gene Ontology Biological Process(http://obofoundry.org/cgi-bin/detail.cgi?id=biological_process), with a quality term from the generic Phenotype and Trait Ontology (PATO: http://obofoundry.org/wiki/index.php/PATO:Main_Page). Table 1 shows three examples of original systematic character narratives (source: Buckup, 1998) and their corresponding EQ statements created by human curators. Note, limited by the coverage of existing ontologies, a precise E or Q term may not be located. In these cases, a broader term has to be used in order to link the narrative to an ontology (terms with * in Table 1 are among those cases, terms in "[]" are the more precise and more desirable terms). Note also that there may be one or more EQ statements for each systematic character.

PCS System Schematic Diagram

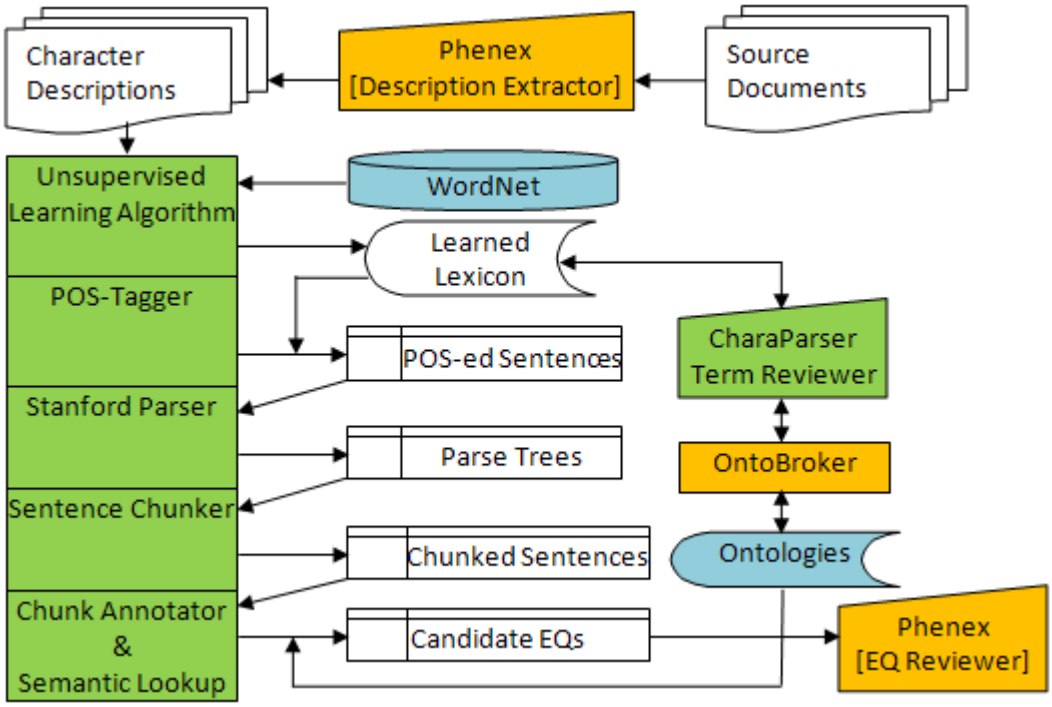

Figure 1. PCS System Schematic Diagram. Orange-colored components are part of Phenex, green-colored components are part of CharaParser, blue-colored parts are existing knowledge resources, and the remaining are the source or the intermediate results produced by the system. Trapezoidal components indicate modules that are interactive with users.

Figure 1 depicts the components of the PCS and related workflow, which involves the following steps:

1. Research assistants extract character matrices and narratives, together with other useful information (such as the identity of specimens examined), from source documents and record them in NeXML format (http://www.nexml.org) using Phenex.

2. CharaParser runs an unsupervised learning algorithm to collect entity terms and quality terms from character narratives.

3. Collected terms are reviewed by a human curator using CharaParser (Figure 2).

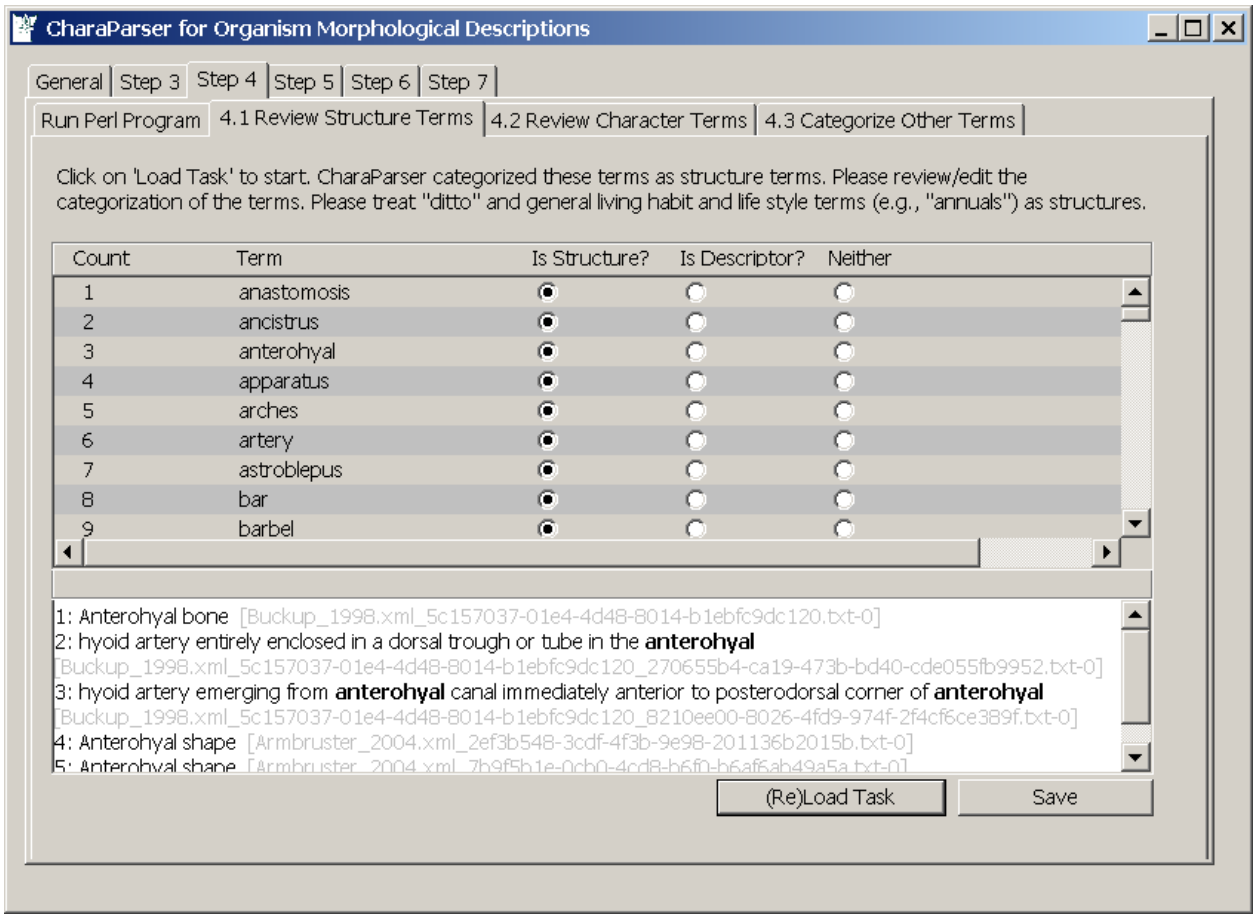

Figure 2. CharaParser Term Reviewer Module. Here entity (i.e. “structure”) terms identified by CharaParser are listed for the curator to review. The curator may assign a term to a different category (e.g., descriptor or neither) if that term is not an entity term. The original context in which a term appears may be displayed when the curator clicks anywhere on the row the term sits. On “4.2” and “4.3” tabs, the curator may review quality (i.e., “character”) terms or identify additional entity/quality terms the system failed to collect.

4. Entity and quality terms that are not in ontologies are submitted to target ontologies by using an OntoBroker service that immediately provides provisional IDs.

5. Using the ontologies including provisional terms, CharaParser executes a series of algorithms to produce candidate EQ statements.

6. Candidate EQ statements are reviewed by a human curator using the Phenex interface (Figure 3). Phenex supports ontology lookup by the curator in case a wrong term ID is applied by the automated process (Figure 4).

7. User feedback from the EQ Review step is captured and used on the fly to correct errors generated by the automated process.

The OntoBroker and user feedback modules are currently not functional yet. PCS may be used in an interactive mode where an individual document is curated or a batch mode where hundreds of documents are curated.

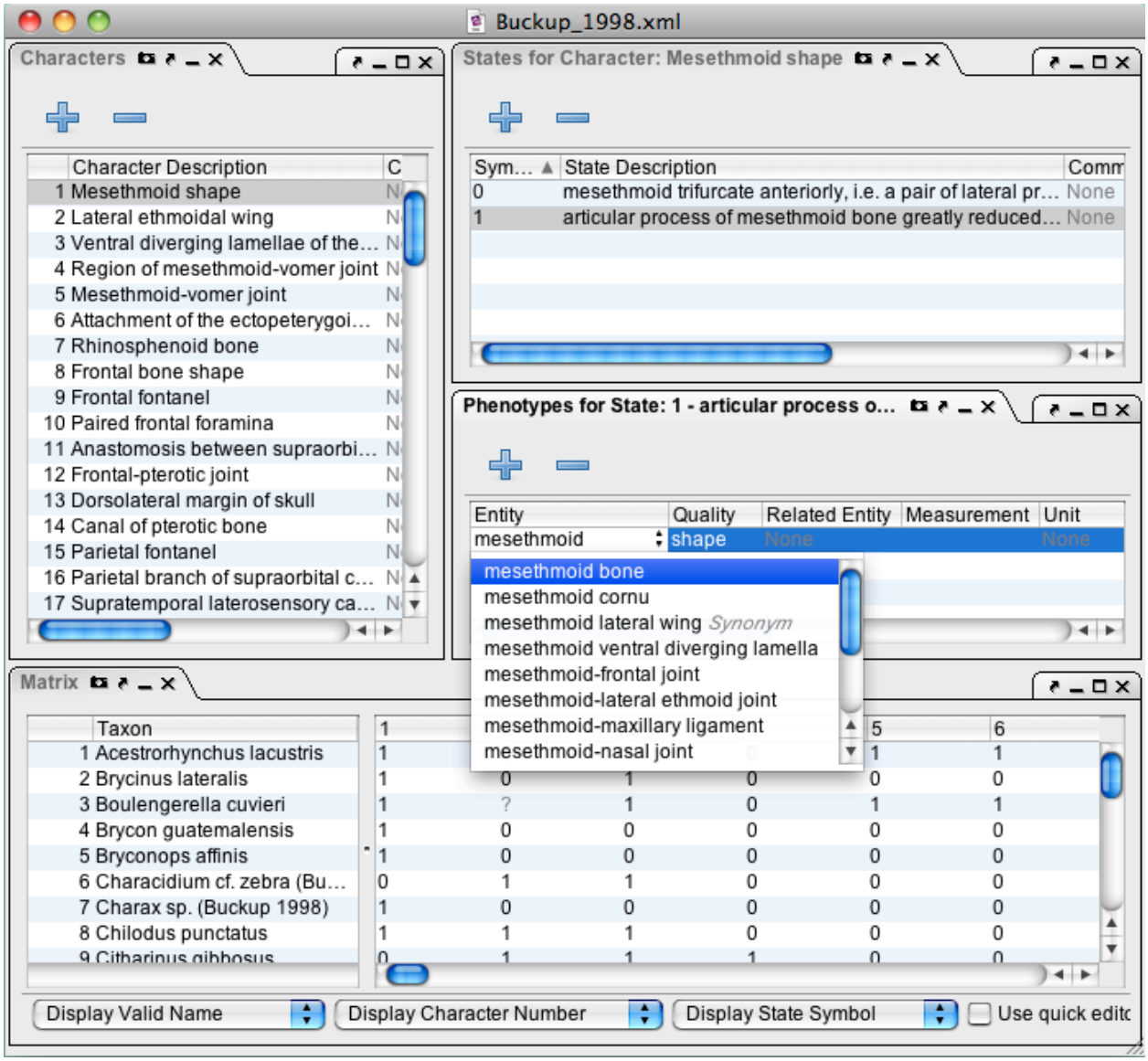

Figure 3. Phenex EQ Statement Reviewer Interface. This interface shows the systematic character information in a source document Buckup, 1998. The information is shown both in the “Characters” and “States for Character” sections as textual narratives and in “Matrix” section as a taxon-by-character matrix. These provide the contextual information about a character and are used by the curator to evaluate the EQ statement (shown in “phenotypes for state” section).

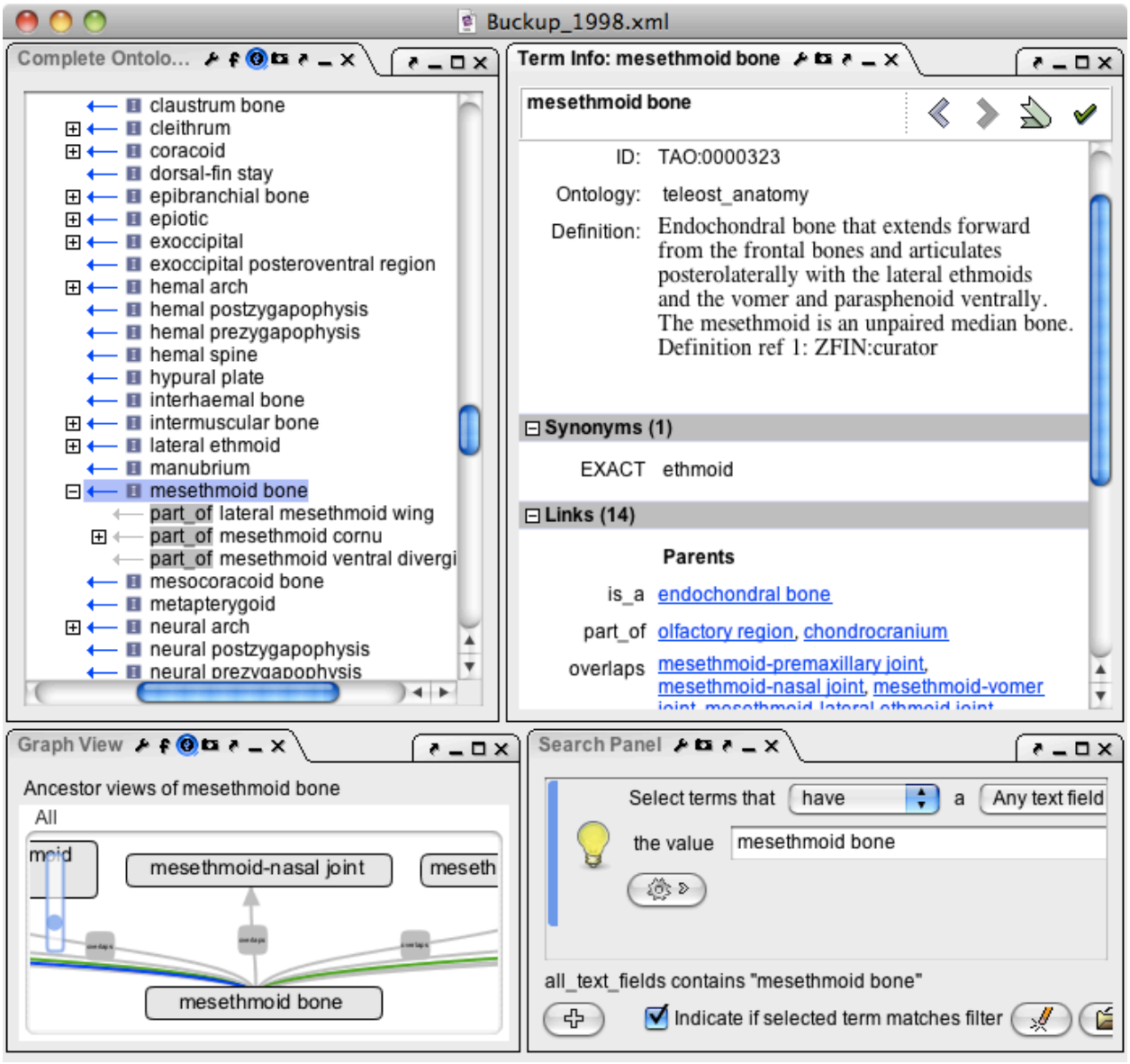

Figure 4. Phenex Ontology Look-up Interface. This interface allows the curator to search for a matching term in an ontology, for example, TAO. Besides the search panel, the interface can be configured to show the tree structure of a complete ontology, graph views of a term, and the detailed definition information of a term.

System Adaptability and Interactivity

The adaptability and interactivity of PCS is due to the complementary features of Phenex and CharaParser. While both these tools are desktop applications, we will be happy to install the system for any curators to use. The GUI of Phenex has been optimized for evolutionary biologists who are accustomed to working with lists of taxa, character narratives, and taxon-by-character matrices. It can be configured to load terms from any OBO ontology so that it can be applied to data curation for any taxonomic groups as long as the appropriate anatomy, taxonomy, and phenotype ontologies exist. Similarly, CharaParser uses unsupervised learning methods to adapt a general-purpose syntactic parser for

semantic markup of morphological narratives of any taxonomic groups. CharaParser has been evaluated and used to perform fine-grained semantic markup of morphological narratives of plants, ants, fish, and invertebrate fossils. The completely integrated PCS (integration is currently underway) will have a look and feel consistent with the current Phenex GUI and will be useful for curating semi-structured systematic character narratives of any taxonomic groups.

Performance

The text mining component of PCS (i.e., CharaParser) has been benchmarked and evaluated under a slightly different setting (Cui, in press). Phenex has been used to produce more than 50,000 manually curated EQ statements from 55 publications. New documents and EQ statements are generated manually every day. These curated EQs can be found at http://kb.phenoscape.org and will be our basis for benchmarking PCS. We will provide precision and recall measurements as requested by March 1, 2012.

Proposed task for BioCreative Track III:

Input: A set of systematic character narratives of fish or dinosaur and a set of ontologies (PATO, TAO, BSPO: http://obofoundry.org/cgi-bin/detail.cgi?id=spatial).

Sample systematic character narratives are in Table 1 and more examples can be found at http://kb.phenoscape.org/taxon_annotations (click on "source" in the Results table).

All ontologies are accessible via OBO foundry at http://www.obofoundry.org/.

Output: EQ statements using terms and EQ statements using term IDs. These EQ statements reflect the semantic content of the input. Table 1 shows both types of output.

The task will be performed in three modes: fully manual, using Phenex, or using PCS. In each of the modes, the output EQ statements will be recorded in a table format like that of Table 1.

Task Mode: Manual: the curator creates EQ statements based on the systematic characters without software assistance.

Task Mode: Using Phenex: curator create EQ statements using Phenex alone.

Task Mode: Using PSO: curator create EQ statements using PSO (i.e. Phenex + CharaParser).

REFERENCES

Balhoff, J. P., W. M. Dahdul, C. R. Kothari, H. Lapp, J. G. Lundberg, P. M. Mabee, P. E. Midford, M. Westerfield, and T. J. Vision. 2010. Phenex: Ontological annotation of phenotypic diversity. PLoS ONE 5(5):e10500. doi:10.1371/journal.pone.0010500

Buckup, P. A. 1998. Relationships of the Characidiinae and phylogeny of characiform fishes (Teleostei: Ostariophysi). Pages 123-144 in Phylogeny and Classification of Neotropical Fishes (L. R. Malabarba, R. E. Reis, R. P. Vari, Z. M. S. Lucena, and C. A. S. Lucena, eds.). EDIPUCRS, Porto Alegre, Brazil.

Cui, H. In press. CharaParser for Fine-Grained Semantic Annotation of Organism Morphological Descriptions. *Journal of American Society of Information Science and Technology*.

Dahdul, W. M., J. P. Balhoff, J. Engeman, T. Grande, E. J. Hilton, C. R. Kothari, H. Lapp, J. G. Lundberg, P. E. Midford, T. J. Vision, M. Westerfield, and P. M. Mabee. 2010. Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. *PLoS ONE* 5(5):e10708. doi:10.1371/journal.pone.0010708

Mabee, P. M., Ashburner, M., Cronk, Q., Gkoutos, G. V., Haendel, M., Segerdell, E., Mungall, C. J., et al. (2007). Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol*, 22(7). doi:10.1016/j.tree.2007.03.013