Phenoscape: Semantic analysis of organismal traits and genes yields insights in evolutionary biology

Paula M. Mabee[1], Wasila M. Dahdul[1], James P. Balhoff [2], Hilmar Lapp [3], Prashanti Manda[4], Josef Uyeda[5], Todd Vision[6], Monte Westerfield[7]

[1]Department of Biology, University of South Dakota, Vermillion, South Dakota, USA
[2]Renaissance Computing Institute, University of North Carolina, Chapel Hill, North Carolina, USA
[3]Center for Genomic and Computational Biology, Duke University, Durham, North Carolina, USA
[4]Department of Computer Science, University of North Carolina at Greensboro, North Carolina, USA
[5]Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia, USA
[6]Department of Biology, University of North Carolina at Chapel Hill, North Carolina, USA
[7]Institute of Neuroscience, University of Oregon, Eugene, Oregon, USA

Corresponding Author:
Paula M. Mabee[1]
Email address: paula.mabee@usd.edu

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26988v1 | CC BY 4.0 Open Access | rec: 13 Jun 2018, publ: 13 Jun 2018

ABSTRACT

The study of how the observable features of organisms, i.e., their phenotypes, result from the complex interplay between genetics, development, and the environment, is central to much research in biology. The varied language used in the description of phenotypes, however, impedes the large scale and interdisciplinary analysis of phenotypes by computational methods. The Phenoscape project (www.phenoscape.org) has developed semantic annotation tools and a gene–phenotype knowledgebase, the Phenoscape KB, that uses machine reasoning to connect evolutionary phenotypes from the comparative literature to mutant phenotypes from model organisms. The semantically annotated data enables the linking of novel species phenotypes with candidate genes that may underlie them. Semantic annotation of evolutionary phenotypes further enables previously difficult or novel analyses of comparative anatomy and evolution. These include generating large, synthetic character matrices of presence/absence phenotypes based on inference, and searching for taxa and genes with similar variation profiles using semantic similarity. Phenoscape is further extending these tools to enable users to automatically generate synthetic supermatrices for diverse character types, and use the domain knowledge encoded in ontologies for evolutionary trait analysis. Curating the annotated phenotypes necessary for this research requires significant human curator effort, although semi-automated natural language processing tools promise to expedite the curation of free text. As semantic tools and methods are developed for the biodiversity sciences, new insights from the increasingly connected stores of interoperable phenotypic and genetic data are anticipated.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26988v1 | CC BY 4.0 Open Access | rec: 13 Jun 2018, publ: 13 Jun 2018

INTRODUCTION

There are over 20 million extant species on the planet, most of which can be described in relation to their unique and widely diverse phenotypes. Comparisons across species phenotypes, however, cannot yet readily be made using computer-assisted methods. This is because the rich legacy of comparative morphology has not yet been semantically enabled—that is, the corpus is in a free-text format that renders computation nearly impossible. This situation began to change almost two decades ago when model organism geneticists began representing the phenotypic changes resulting from experimental gene manipulations, with terms from anatomy or phenotype ontologies that they developed for each model organism (e.g., Sprague et al. 2001). More recently, the opportunity to enable interoperability from the phenotypes of biodiverse species to candidate genes from model species (Mabee et al. 2007a, 2007b) motivated the Phenoscape team to develop one of the first multispecies anatomy ontologies, the Teleost Anatomy Ontology (Dahdul et al. 2010b), based initially on the Zebrafish Anatomy Ontology (Ruzicka et al. 2015). Developing ontologies appropriate for biodiversity, including taxonomy ontologies (Midford et al. 2013) and scaling them up first to the level of teleost fishes (Dahdul et al. 2010), then to the level of vertebrates (Dahdul et al. 2012) and then to the level of metazoans (Mungall et al. 2012; Haendel et al. 2014), further enabled the automation of phenotypic comparisons across vertebrate species and discovery of candidate genes underlying evolutionarily novel phenotypes by the team (Edmunds et al. 2016). Over the past ten years a broad community of scientists invested in the development of shared community ontologies (e.g., Gkoutos et al. 2005; Haendel et al. 2008, 2014; Dahdul et al. 2014), annotation tools (Balhoff et al. 2010, 2014a; Yoder et al. 2010; Cui et al. 2016; The Gene Ontology Consortium 2017) and formats (Dahdul et al. 2010a; Vos et al. 2012) for phenotype annotation across biodiverse species (Dahdul et al. 2010a). These resources have made computational analyses possible and they have been leveraged to build a wealth of innovative applications (e.g., Deans et al. 2012; Mullins et al. 2012; Balhoff et al. 2013; Dececchi et al. 2015; Manda et al. 2015; Druzinsky et al. 2016; Jackson et al. 2018) across a variety of biodiversity-based research. The Phenoscape Knowledgebase (KB) (Figure 1) demonstrates these connections by integrating gene phenotype annotations from model organism databases with phenotype annotations from the biodiversity literature (Table 1). Compelling demonstrations of the utility of semantics for biodiversity studies are important because of the large and expensive investments in infrastructure and tool development required to curate the legacy literature and move the publication of phenotypic data into a natively semantic form.

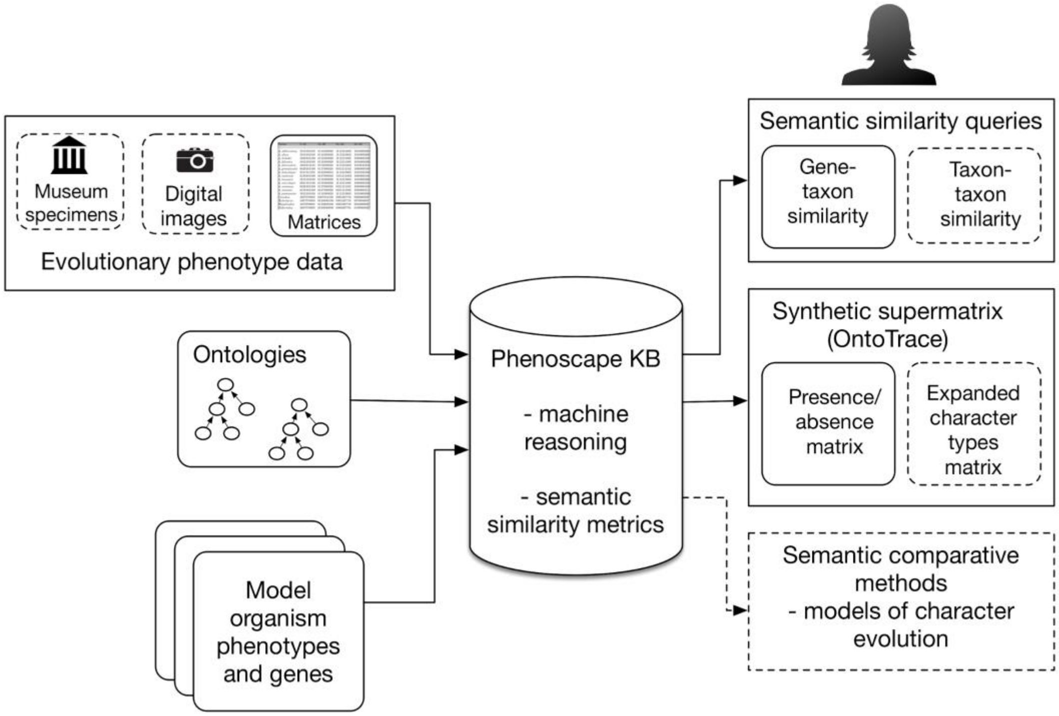
Museum specimens
Digital images
Matrices
Evolutionary phenotype data
Ontologies
Model organism phenotypes and genes
Phenoscape KB
- machine reasoning
- semantic similarity metrics
Semantic similarity queries
Gene-taxon similarity
Taxon-taxon similarity
Synthetic supermatrix (OntoTrace)
Presence/ absence matrix
Expanded character types matrix
Semantic comparative methods
- models of character evolution

Figure 1. Flow chart of currently existing data sources and tools (solid borders and lines) in the Phenoscape KB, and data and tools not yet integrated or developed (dotted borders and lines) but relevant to users in biodiversity research.

Table 1: Data for evolutionary and model organism phenotypes in the Phenoscape KB. (Data as of 2018-05-11)

Evolutionary Phenotypes

Annotated anatomical character states	22,321
Total number of annotated taxa (extant and fossil vertebrates)	5,310
Total number of taxon phenotypes	540,163
Terminal taxa (species) with at least one phenotype	4,260
Non-terminal taxa with at least one phenotype	1,050
Evolutionary phenotype profiles	682

Model Organism Phenotypes

	Zebrafish	Mouse	Xenopus	Human
Genes with at least one phenotype	5,883	7,758	12	3,717
Phenotype annotations	90,132	171,876	236	123,956
Genes with any expression data	12,509	10,599	15,062	0
Gene expression annotations	179,232	800,824	454,337	0

To date, only a small proportion of the biodiversity literature has been annotated semantically, and no publisher, to our knowledge, tags phenotypes with ontological terms that would support interoperability. The comparative study of organismal phenotypes, however, motivates research across diverse fields of biology, including evolution, paleontology, developmental biology, agriculture, and the veterinary and health sciences (Deans et al. 2015). The efficiency and potential of fundamental discoveries in the biodiversity arena would be dramatically expanded by the increased use of semantics. Further, few species, i.e., only model organisms, have curated phenotypic data that is linked to genetic and genomic data. The growth in sequencing technology, however, is changing this dynamic, resulting in the rapid expansion of genomic data for non-model species (e.g., Russell et al. 2017 and Chapter 10). However, without corresponding phenomic databases, the challenge of relating the growing volume of genetic knowledge in model and emerging model organisms to the diversity of phenotypes in nature cannot be met. In this chapter, through the description of driving research questions and by examples of the use of semantically annotated data in the Phenoscape KB, we provide a glimpse of the promise that semantic analysis tools hold in comparing phenotypes across species and globally associating genetic to phenotypic data.

1. Relating biodiverse phenotypes to candidate genes

Identifying the genetic and developmental changes that brought forth the incredible phenotypic diversification of life is a recalcitrant problem, but one where a basic semantic approach has shown promise and where more sophisticated approaches using semantic similarity may yet be even more valuable. Semantic similarity enables comparison and analysis of semantic annotations between entities (genes, taxa) using ontologies and computational reasoners to compute scores that reflect the level of similarity (e.g., Washington et al. 2009; Manda et al.

2015; see examples in Chapter 10). The Phenoscape team showed that ontology-driven information systems can generate thousands of testable hypotheses relating unique morphologies from non-model biodiverse species to candidate genes (Mabee et al. 2012). One of these, for example, connected the unique loss of a tongue ('basihyal element') in catfishes (Siluriformes) with several candidate genes from the zebrafish data. Edmunds et al. (2016) experimentally tested the candidates by examining their endogenous expression patterns in the channel catfish, *Ictalurus punctatus*, and found results consistent with the in silico hypothesis that the tongue evolved through disruption in developmental pathways at, or upstream of, *brpf1*.

The Phenoscape team recently extended this approach (Manda et al. 2015) by using semantic similarity to find matches between the full set of phenotypes described for a gene and the unique set of phenotypes that characterizes a clade of species, i.e., an 'evolutionary phenotype profile'. The effects from a gene knockdown range from several to hundreds of phenotypes, and the goal is to compare these in their entirety to the calculated set of phenotypes that are variable among the immediate descendants of a particular taxon. Using semantic similarity, the Phenoscape KB performs fuzzy matching between suites of phenotypes, and displays the taxonomic groups that vary in phenotypes that match most closely to the gene profile that results when the action of that gene is disrupted (e.g., knocked down). The user interface provides the statistical support for each match and allows the supporting evidence to be examined. There are some important caveats that must be considered when interpreting the results, such as the potential for some matches to result from differences in annotation coverage between genetic and evolutionary studies in the KB. Potentially spurious matches in that category are flagged by the KB. The KB also provides an interface for the reverse query: what genes have phenotypes that match most closely to the set of evolutionary phenotypes in a particular taxon under consideration? That is, a biologist who is curious about the genetic basis of taxonomic diversity might want to find genes that have phenotypes that resemble the phenotypic variation exhibited by a particular taxon.

2. Future applications of semantic similarity to phenotypes of biodiverse taxa

Questions of whether a particular combination of phenotypes in a taxon is unique, or what it might be similar to, are the types of broad questions that may be addressed in applying semantic similarity-based data mining to phenotypes across diverse taxa. Semantic similarity would retrieve taxa with similar phenotypic profiles; such similarity may have arisen because of common ancestry or independent origin (a 'homoplasy finder'). As described by Braun et al. (Chapter 10), predictive phenomics can, for example, be used to target desired phenotypes in species of interest - and together with recent gene editing capabilities, functional genomic analysis can be newly brought to bear on biodiverse species. The Phenoscape KB currently enables users to view taxa with variation similar to the phenotypic profile of a gene (and *vice versa*). In the future, they will also be able to query one custom set of phenotypes against another or a taxonomically selected subset, and obtain a ranked list of taxa with similar phenotypes. For example, miniature fishes in the genus *Paedocypris*, like many fishes that are evolutionarily reduced to an extremely small body size, exhibit the absence of bones including the interhyal, vomer, parietal, posttemporal, and supraneurals (Britz and Conway 2009). Are there other taxa that lack a highly similar set of bones? Enabling a comparison of these phenotypes across diverse taxa would allow a user to query for such matches; in this case, matches would include the ricefishes in the family Adrianichthyidae (Wiley and Johnson 2010), which similarly lack the interhyal, vomer, and supraneurals, and other bones such as the supracleithrum. Further, adrianichthyids may lack or possess extremely small or absent parietal bones and have structurally simple posttemporal bones, which biologists may recognize as reductive phenotypes on a continuum close to 'absent'. Methods that incorporate a framework of probabilistic reasoning for phenotype relatedness (e.g., Bauer et al. 2012) have the potential to improve precision of ontology-based queries.

3. Relating biodiverse phenotypes across studies: presence/absence

Addressing many of the questions in the biodiversity sciences involve knowing how a specific trait or set of traits has evolved across a group of species. Although the published literature is replete with research relating species and traits, and a few repositories hold phylogenetic trees, some of which are computed products from trait data, neither the traits nor the trees can be easily synthesized across studies. The OntoTrace tool was developed by the Phenoscape team (Balhoff et al. 2014b; Dececchi et al. 2015) to enable users to automatically pull together, from phenotype annotations made to published character matrices and monographic texts (Dececchi et al. 2015, 2016), a set of presence/absence data for specific traits for a set of taxa. For example, querying the Phenoscape KB for a supermatrix of traits of fins, limbs, girdles and their parts in sarcopterygian vertebrates (lobe-finned fishes and tetrapods), Dececchi et al. (2015) retrieved data for 1,052 taxa from 55 studies. The data, 1,759 synthetic presence/absence characters, were derived from 2,588 text-based character states (1,195 characters). The resultant character by taxon matrix was termed a 'synthetic morphological supermatrix'. Because of the ontological annotations, not only could these phenotypic data be automatically aggregated from multiple studies into a supermatrix, but the asserted data could be extended through inference to traits that were implied by, but not directly asserted in the original publications. For example, if an author observed a curved pectoral fin ray in a species, the machine would infer, based on the knowledge of anatomy encoded in the requisite ontology

(Uberon in this case), that a pectoral fin is present in that species (see Dececchi et al. 2015 and Jackson et al. 2018 for further examples). In this manner, the missing data in the variable character subset of the matrix (the subset containing only characters that include both present and absent states) was reduced from 98.5% to 78.2%. Further, 76% of the variable characters were made variable through the addition of inferred states. The authors pointed out that character conflicts and provenance reports from OntoTrace would support researchers review of large aggregated data sets and they showed how such machine reasoning enables quantification and new visualizations of the data, allowing the identification of undersampled character space.

4. Relating biodiverse phenotypes to phylogenetic trees

Using ontologies and machine reasoning to automatically generate large, synthetic character matrices of presence/absence phenotypes (as per above) set the stage for the research of Jackson et al. (2018), who took this a step further. They developed a bioinformatic pipeline to propagate data that was asserted to higher-level taxonomic nodes, to descendant species that were missing data. Similar to Dececchi et al. (2015), they showed that such logic inference significantly extended the asserted data (missing data were reduced from 98.0% to 85.9%), but additionally they showed the value of taxonomic data propagation, which extended the data further, reducing missing data to 34.8% (Jackson et al. 2018). Using the resultant matrix along with a synthetic phylogeny from the Open Tree of Life (Hinchliff et al. 2015), they mapped the full trait data set for 12,582 species to the tree and addressed the question of how often paired fins were lost in teleost fishes and whether they were ever regained (Jackson et al. 2018). Looking ahead, if all published traits and trees were made computable using these methods, any user could automatically generate a matrix for a specified set of traits and map it on various synthetic tree topologies, which in turn would allow addressing a host of questions regarding the pattern and tempo of phenotypic evolution and associations with genomic and environmental (Thessen et al. 2015) variables.

5. Relating biodiverse phenotypes across studies: future work

As described above, OntoTrace generates synthetic morphological supermatrices for presence/absence characters only (Dececchi et al. 2015). Expanding this functionality to automatically synthesize characters of other qualities, such as shape, size, structure, and color, is a current challenge that the Phenoscape team is addressing. For example, whereas characters in a presence/absence matrix are by definition limited to two states per character, the number of possible states for characters in other categories is *a priori* unconstrained. Thus, automatically synthesizing characters that, for example, describe ‘basihyal bone, shape’, can result in a large number of states per character because every originally published state that semantically is some type of ‘basihyal bone shape’ would have to be appended as a new state to the synthesized character. In the case of this example, there may be seven distinct shape terms used in its annotation (Box 1). The ontological relationships indicate that subsets of these states are more similar to each other than others. By adapting current semantic similarity metrics for the purpose of character and character state aggregation, and in effect, homology assignment, these distinct shape descriptors can be consolidated into new, synthetic states (see matrix in Box 1).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26988v1 | CC BY 4.0 Open Access | rec: 13 Jun 2018, publ: 13 Jun 2018

Box 1. Assembling a synthetic character and its states for 'basihyal bone, shape'.

Step 1: Assemble list of 'shape' (PATO:0000052) quality terms for all characters and states from multiple publications that include the entity 'basihyal bone' (UBERON:0011618):

'spiny' (PATO:0001365)

'folded' (PATO:0001910)

'upturned' (PATO:0002031)

'blade-like' (PATO:0002235)

'pointed' (PATO:0002258)

'curved ventral' (PATO:0001469)

'tapered' (PATO:0001500)

Step 2: Apply semantic similarity to above list of PATO terms for basihyal bone. Because of higher similarity among terms, three states (0, 1, 2) are generated from the seven phenotypes:

Character 1: Basihyal bone: shape

Synthetic State 0: 'sharp' (PATO:0001419) (includes 'blade-like', 'pointed', 'tapered')

Synthetic State 1: 'curved' (PATO:0000406) (includes 'upturned', 'curved ventral')

Synthetic State 2: 'surface feature shape' (PATO:0001925) (includes 'spiny', 'folded')

The Phenoscape team is now developing semantic similarity-based methods to cluster phenotypes across different character categories into characters and states, thus automating matrix construction, and enabling users to optimize the matrix for a variety of metrics. This

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26988v1 | CC BY 4.0 Open Access | rec: 13 Jun 2018, publ: 13 Jun 2018

would allow a user to constrain the number of characters in a synthesized matrix by excluding those with low information content (e.g., those for high level terms from the anatomy ontology such as ‘fin’’ vs. ‘pectoral fin’). Thus, employing semantic reasoning in matrix construction will allow a user to balance the properties of a synthetic matrix between, on the one hand, containing highly specific characters (and thus increased missing data), and on the other, including lower specificity characters (and thus decreasing missing data).

In addition to semantic tools for supermatrix construction, the Phenoscape team is developing enhanced semantics for addressing questions of trait evolution. Unlike the current tools available for analyzing molecular data, where each nucleotide site can be treated as independent of each other, evolutionary models for large morphological character matrices face significant challenges overcoming the strong conditional dependencies and correlations among morphological traits. Most existing methods ignore such dependencies and morphological characters are treated as independent. By leveraging domain knowledge relevant to assessing correlations of the traits underlying the characters, Phenoscape is developing tools that enable users to incorporate evidence of the relatedness of traits in a morphological matrix and into models of character evolution. These include measures of trait independence based on ontological relationships, distance (semantic similarity) of traits in the knowledge graph, and measures of genetic overlap (as derived from gene-phenotype annotations from model organism databases). Such dependencies can be directly built into the macroevolutionary model, or can be used to inform prior probabilities in Bayesian analyses when grouping traits into modules with shared evolutionary parameters or dynamics.

One of the challenges in conducting semantic similarity comparisons is the computational overhead of comparing EQ phenotypes over a large ontology space. Improvements in scalability of semantic similarity methods would enable fast, on-the-fly semantic similarity searches. Successfully applying these methods also currently depends on accurate *a posteriori* annotation of characters to capture the original author's intent. With only the published description and perhaps images to rely on, curators are unable to consistently apply standardized terms, a factor leading to lower consistency (Cui et al. 2015). For example, in a comparison of curator vs. machine generated annotations (Cui et al. 2015; in prep), three curators described the increased distance between the contralateral pelvic fins with three different quality descriptors: 'far from', 'separated from', and 'set apart from'. As methods and software tools develop, such that original authors are better empowered to apply the semantics themselves, the accuracy of character annotation, and thus, consolidation will increase. In the above example, the author would presumably be able to choose, based on the different definitions of the ontology terms, which term is most applicable to the phenotype observed.

6. Future challenges

A long-standing question, and one also being currently tackled by the Phenoscape team, is how the relationship of historical homology, i.e., similarity due to common ancestry, can most effectively be used in data retrieval. Recent work by Manda et al. (2016b), examined how semantic similarity is affected when external homology knowledge is included in an ontology. They measured phenotypic similarity between orthologous and non-orthologous gene pairs between humans and either mouse or zebrafish, and they compared the effect of including real vs. faux homology axioms. Semantic similarity was preferentially increased for orthologs when using real homology axioms, though only across the more divergent of the two species (human to zebrash, not human to mouse) (Manda et al., 2016). Overall, the effect of including homology axioms on cross-species semantic similarity was modest, though the authors suggested that the effect might be greater for more distant species comparisons. Current efforts include editing and clarifying the homology relationships in the Uberon ontology and investigating how reasoning on different models of homology affects information retrieval in the KB.

Another challenge for the broader application of semantics to biodiversity data is the significant, largely manual, effort necessary to annotate phenotypes from the published literature (Dahdul et al. 2015). Natural language processing tools are needed going forward to auto-annotate the legacy literature (Arighi et al. 2013; Cui et al. 2015; in prep). Further, in the future semantic phenotype data may increasingly come directly from publications, as semi-automated methods for marking up manuscripts at the time of publication become more accurate, mature, and thus prevalent. Evaluating, and hence continuously improving the accuracy of machine generated annotations depends on expert-curated “gold standard” data sets. To this end, Phenoscape has developed the first gold standard dataset for biodiversity phenotypes (in prep). Efforts to use ontologies in the process of new species descriptions are underway (Deans et al. 2012; Balhoff et al. 2013), and will contribute to achieving a vision of widely available linked species phenotype data.

As high-throughput phenotyping, typically involving image data collection, becomes more scalable, the application of semantic metadata would enable automated connections to the tools and computable datasets described herein. These digitization efforts can be new sources of phenotype information (Figure 1). Although broad domains of biology can be served if semantics are placed on digitized images and specimens, so far only a few projects are using semantics to label digitized specimens and their parts, despite promising prototypes (Maglia et al. 2007; Rámirez et al. 2007). If anatomical parts were tagged with ontology terms, then queries on basic trait distributions could be enabled (e.g., presence of pectoral fins in taxa a, b, c...). Although having a reduced information content compared to full Entity-Quality expressions, entity-only annotations have been shown to be informative for semantic similarity (Manda et al. 2016a). Thus, new sources of phenotypic data, such as those for specimens of extinct and extant taxa associated with institutional collections, can easily be made interoperable through shared semantics (Figure 1).

CONCLUSIONS

Over the past 10 years the development of shared cross-species community ontology resources such as Uberon and PATO has enabled interoperability of phenotype and genotype data. This in turn enables a wealth of potential applications and discoveries from semantic analysis of biodiverse taxa. Scientific attention continues to move toward gaining a deeper fundamental understanding of the developmental and evolutionary relationship between genotype and phenotype. The profound scale and scope of this problem will not only require interoperable big data, both genomic and phenomic, from a biodiverse set of taxa, but also new ways of using machines to enable this understanding. The applications of semantic analysis described herein only scratch the surface of what is possible. As scientific publication moves to incorporate semantic markup of phenotype data, and semi-automated tools are improved to annotate the phenotype legacy literature, knowledge of the rich phenotypic palette of life on our planet can be exposed to machine computation with great advantage to fundamental discovery across the life sciences.

ACKNOWLEDGEMENTS

We thank all participants in the Phenotype Ontology Research Coordination Network (RCN) (NSF 0956049) for their vision, contributions, and commitment to developing shared and interoperable resources. During the course of this work the Phenoscape project has been supported by NSF awards 1062404, 1062542, 0641025, 1661529, and the National Evolutionary Synthesis Center (NSF 0905606 and 0423641). This manuscript is based in part on work done by P.M.M. while serving at the U.S. National Science Foundation. The views expressed in this paper do not necessarily reflect those of the National Science Foundation or the United States Government.

REFERENCES

Arighi C.N., Carterette B., Cohen K.B., Krallinger M., Wilbur W.J., Fey P., Dodson R., Cooper L., Van Slyke C.E., Dahdul W., Mabee P., Li D., Harris B., Gillespie M., Jimenez S., Roberts P., Matthews L., Becker K., Drabkin H., Bello S., Licata L., Chatr-aryamontri A., Schaeffer M.L., Park J., Haendel M., Van Auken K., Li Y., Chan J., Muller H.-M., Cui H., Balhoff J.P., Chi-Yang Wu J., Lu Z., Wei C.-H., Tudor C.O., Raja K., Subramani S., Natarajan J., Cejuela J.M., Dubey P., Wu C. 2013. An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. Database. 2013:bas056.

Balhoff J.P., Dahdul W.M., Dececchi T.A., Lapp H., Mabee P.M., Vision T.J. 2014a. Annotation of phenotypic diversity: decoupling data curation and ontology curation using Phenex. J. Biomed. Semantics. 5:45.

Balhoff J.P., Dahdul W.M., Kothari C.R., Lapp H., Lundberg J.G., Mabee P., Midford P.E., Westerfield M., Vision T.J. 2010. Phenex: ontological annotation of phenotypic diversity. PLoS One. 5(5):e10500.

Balhoff J.P., Dececchi T.A., Mabee P.M., Lapp H. 2014b. Presence-absence reasoning for evolutionary phenotypes. Peer-reviewed conference paper, Bio-ontologies SIG at ISMB 2014.

Balhoff J.P., Mikó I., Yoder M.J., Mullins P.L., Deans A.R. 2013. A semantic model for species description applied to the ensign wasps (Hymenoptera: Evaniidae) of New Caledonia. Syst. Biol. 62:639–659.

Bauer S., Köhler S., Schulz M.H., Robinson P.N. 2012. Bayesian ontology querying for accurate and noise-tolerant semantic searches. Bioinformatics. 28:2502–2508.

Britz R., Conway K.W. 2009. Osteology of Paedocypris, a miniature and highly developmentally truncated fish (Teleostei: Ostariophysi: Cyprinidae). J. Morphol. 270:389–412.

Cui H., Dahdul W., Dececchi T.A., Ibrahim N., Mabee P., Balhoff J.P., Gopalakrishnan H. 2015. CharaParser+EQ: Performance Evaluation without Gold Standard. Proceedings of the 78th Annual Meeting of the Association for Information Science and Technology Annual (ASIS&T). Vol. 51.

Cui H., Xu D., Chong S.S., Ramirez M., Rodenhausen T., Macklin J.A., Ludäscher B., Morris R.A., Soto E.M., Koch N.M. 2016. Introducing Explorer of Taxon Concepts with a case study on spider measurement matrix building. BMC Bioinformatics. 17:471.

Dahdul W., Dececchi T.A., Ibrahim N., Lapp H., Mabee P. 2015. Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy. Database . 2015:bav040.

Dahdul W.M., Balhoff J.P., Blackburn D.C., Diehl A.D., Haendel M.A., Hall B.K., Lapp H., Lundberg J.G., Mungall C.J., Ringwald M., Segerdell E., Van Slyke C.E., Vickaryous M.K., Westerfield M., Mabee P.M. 2012. A unified anatomy ontology of the vertebrate skeletal system. PLoS One. 7(12):e51070.

Dahdul W.M., Balhoff J.P., Engeman J., Grande T., Hilton E.J., Kothari C., Lapp H., Lundberg J.G., Midford P.E., Vision T.J., Westerfield M., Mabee P.M. 2010a. Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. PLoS One. 5(5):e10708.

Dahdul W.M., Cui H., Mabee P.M., Mungall C.J., Osumi-Sutherland D., Walls R.L., Haendel M.A. 2014. Nose to tail, roots to shoots: spatial descriptors for phenotypic diversity in the Biological Spatial Ontology. J. Biomed. Semantics. 5:34.

Dahdul W.M., Lundberg J.G., Midford P.E., Balhoff J.P., Lapp H., Vision T.J., Haendel M.A., Westerfield M., Mabee P.M. 2010b. The Teleost Anatomy Ontology: anatomical representation for the genomics age. Syst. Biol. 59:369–383.

Deans A.R., Lewis S.E., Huala E., Anzaldo S.S., Ashburner M., Balhoff J.P., Blackburn D.C., Blake J.A., Burleigh J.G., Chanet B., Cooper L.D., Courtot M., Csösz S., Cui H., Dahdul W., Das S., Dececchi T.A., Dettai A., Diogo R., Druzinsky R.E., Dumontier M., Franz N.M., Friedrich F., Gkoutos G.V., Haendel M., Harmon L.J., Hayamizu T.F., He Y., Hines H.M., Ibrahim N., Jackson L.M., Jaiswal P., James-Zorn C., Köhler S., Lecointre G., Lapp H., Lawrence C.J., Le Novère N., Lundberg J.G., Macklin J., Mast A.R., Midford P.E., Mikó I., Mungall C.J., Oellrich A., Osumi-Sutherland D., Parkinson H., Ramírez M.J., Richter S., Robinson P.N., Ruttenberg A., Schulz K.S., Segerdell E., Seltmann K.C., Sharkey M.J., Smith A.D., Smith B., Specht C.D., Squires R.B., Thacker R.W., Thessen A., Fernandez-Triana J., Vihinen M., Vize P.D., Vogt L., Wall C.E., Walls R.L., Westerfeld M., Wharton R.A., Wirkner C.S., Woolley J.B., Yoder M.J., Zorn A.M., Mabee P. 2015. Finding our way through phenotypes. PLoS Biol. 13(1):e1002033.

Deans A.R., Yoder M.J., Balhoff J.P. 2012. Time to change how we describe biodiversity. Trends Ecol. Evol. 27:78–84.

Dececchi T.A., Balhoff J.P., Lapp H., Mabee P.M. 2015. Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. Syst. Biol. 64:936–952.

Dececchi T.A., Mabee P.M., Blackburn D.C. 2016. Data sources for trait databases: comparing the phenomic content of monographs and evolutionary matrices. PLoS One. 11(5):e0155680.

Druzinsky R.E., Balhoff J.P., Crompton A.W., Done J., German R.Z., Haendel M.A., Herrel A., Herring S.W., Lapp H., Mabee P.M., Muller H.-M., Mungall C.J., Sternberg P.W., Van Auken K., Vinyard C.J., Williams S.H., Wall C.E. 2016. Muscle Logic: new knowledge resource for anatomy enables comprehensive searches of the literature on the feeding muscles of mammals. PLoS One. 11(2):e0149102.

Edmunds R.C., Su B., Balhoff J.P., Eames B.F., Dahdul W.M., Lapp H., Lundberg J.G., Vision T.J., Dunham R.A., Mabee P.M., Others. 2016. Phenoscape: identifying candidate genes for evolutionary phenotypes. Mol. Biol. Evol. 33:13–24.

Gkoutos G.V., Green E.C.J., Mallon A.-M., Hancock J.M., Davidson D. 2005. Using ontologies to describe mouse phenotypes. Genome Biol. 6:R8.

Haendel M.A., Balhoff J.P., Bastian F.B., Blackburn D.C., Blake J.A., Bradford Y., Comte A., Dahdul W.M., Dececchi T.A., Druzinsky R.E., Hayamizu T.F., Ibrahim N., Lewis S.E., Mabee P.M., Niknejad A., Robinson-Rechavi M., Sereno P.C., Mungall C.J. 2014. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. J. Biomed. Semantics. 5:21.

Haendel M.A., Neuhaus F., Osumi-Sutherland D., Mabee P.M., Mejino J.L.V., Mungall C.J., Smith B. 2008. CARO--the Common Anatomy Reference Ontology. In: Burger A, Davidson D, Baldock R (eds) Anatomy Ontologies for Bioinformatics. Computational Biology, vol 6. Springer, London.

Hinchliff C.E., Smith S.A., Allman J.F., Burleigh J.G., Chaudhary R., Coghill L.M., Crandall K.A., Deng J., Drew B.T., Gazis R., Gude K., Hibbett D.S., Katz L.A., Laughinghouse H.D. 4th, McTavish E.J., Midford P.E., Owen C.L., Ree R.H., Rees J.A., Soltis D.E., Williams T., Cranston K.A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl. Acad. Sci. U. S. A.

Jackson L.M., Fernando P.C., Hanscom J.S., Balhoff J.P., Mabee P.M. 2018. Automated integration of trees and traits: a case study using paired fin loss across teleost fishes. Syst. Biol.:syx098.

Mabee P., Balhoff J.P., Dahdul W.M., Lapp H., Midford P.E., Vision T.J., Westerfield M. 2012. 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. J. Appl. Ichthyol. 28:300–305.

Mabee P.M., Arratia G., Coburn M., Haendel M., Hilton E.J., Lundberg J.G., Mayden R.L., Rios N., Westerfield M. 2007a. Connecting evolutionary morphology to genomics using ontologies: a case study from Cypriniformes including zebrafish. J. Exp. Zool. B Mol. Dev. Evol. 308:655–668.

Mabee P.M., Ashburner M., Cronk Q., Gkoutos G.V., Haendel M., Segerdell E., Mungall C., Westerfield M. 2007b. Phenotype ontologies: the bridge between genomics and evolution. Trends Ecol. Evol. 22:345–350.

Maglia A.M., Leopold J.L., Pugener L.A., Gauch S. 2007. An anatomical ontology for amphibians. Proceedings of the Pacific Symposium on Biocomputing 2007:367-378.

Manda P., Balhoff J.P., Lapp H., Mabee P., Vision T.J. 2015. Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution. Genesis. 53:561–571.

Manda P., Balhoff J.P., Vision T.J. 2016a. Measuring the importance of annotation granularity to the detection of semantic similarity between phenotype profiles. Proceedings of the International Conference on Biological Ontology 2016.

Manda P., Mungall C.J., Balhoff J.P., Lapp H., Vision T.J. 2016b. Investigating the importance of anatomical homology for cross-species phenotype comparisons using semantic similarity. Proceedings of the Pacific Symposium on Biocomputing 2016.:132–143.

Midford P.E., Dececchi T.A., Balhoff J.P., Dahdul W.M., Ibrahim N., Lapp H., Lundberg J.G., Mabee P.M., Sereno P.C., Westerfield M., Vision T.J., Blackburn D.C. 2013. The Vertebrate Taxonomy Ontology: a framework for reasoning across model organism and species phenotypes. J. Biomed. Semantics. 4:34.

Mullins P.L., Kawada R., Balhoff J.P., Deans A.R. 2012. A revision of Evaniscus (Hymenoptera, Evaniidae) using ontology-based semantic phenotype annotation. Zookeys:1–38.

Mungall C.J., Torniai C., Gkoutos G.V., Lewis S.E., Haendel M.A. 2012. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 13:R5.

Rámirez M.J., Coddington J.A., Maddison W.P., Midford P.E., Prendini L., Miller J., Griswold C.E., Hormiga G., Sierwald P., Scharff N., Benjamin S.P., Wheeler W.C. 2007. Linking of digital images to phylogenetic data matrices using a morphological ontology. Syst. Biol. 56:283–294.

Russell J.J., Theriot J.A., Sood P., Marshall W.F., Landweber L.F., Fritz-Laylin L., Polka J.K., Oliferenko S., Gerbich T., Gladfelter A., Umen J., Bezanilla M., Lancaster M.A., He S., Gibson M.C., Goldstein B., Tanaka E.M., Hu C.-K., Brunet A. 2017. Non-model model organisms. BMC Biol. 15:55.

Ruzicka L., Bradford Y.M., Frazer K., Howe D.G., Paddock H., Ramachandran S., Singer A., Toro S., Van Slyke C.E., Eagle A.E., Fashena D., Kalita P., Knight J., Mani P., Martin R., Moxon S.A.T., Pich C., Schaper K., Shao X., Westerfield M. 2015. ZFIN, The zebrafish model organism database: Updates and new directions. Genesis. 53:498–509.

Sprague J., Doerry E., Douglas S., Westerfield M. 2001. The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. Nucleic Acids Res. 29:87–90.

The Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 45:D331–D338.

Thessen A.E., Bunker D.E., Buttigieg P.L., Cooper L.D., Dahdul W.M., Domisch S., Franz

N.M., Jaiswal P., Lawrence-Dill C.J., Midford P.E., Mungall C.J., Ramírez M.J., Specht C.D., Vogt L., Vos R.A., Walls R.L., White J.W., Zhang G., Deans A.R., Huala E., Lewis S.E., Mabee P.M. 2015. Emerging semantics to link phenotype and environment. PeerJ. 3:e1470.

Vos R.A., Balhoff J.P., Caravas J.A., Holder M.T., Lapp H., Maddison W.P., Midford P.E., Priyam A., Sukumaran J., Xia X., Stoltzfus A. 2012. NeXML: rich, extensible, and verifiable representation of comparative data and metadata. Syst. Biol. 61:675–689.

Washington N.L., Haendel M.A., Mungall C.J., Ashburner M., Westerfield M., Lewis S.E. 2009. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol. 7(11):e1000247.

Wiley E.O., Johnson G.D. 2010. A Teleost Classification Based on Monophyletic Groups. In: Nelson J.S., Schultze H.P., Wilson M.V.H., editors. Origin and Phylogenetic Interrelationships of Teleosts. München, Germany: Verlag Dr. Friedrich Pfeil. p. 123–182.

Yoder M.J., Mikó I., Seltmann K.C., Bertone M.A., Deans A.R. 2010. A gross anatomy ontology for hymenoptera. PLoS One. 5(12):e15991.