Special Issue: Ecological and evolutionary informatics

Time to change how we describe biodiversity

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

[1] Department of Entomology, North Carolina State University, Raleigh, NC 27695, USA
[2] National Evolutionary Synthesis Center, 2024 W. Main Street, Suite A200, Durham, NC 27705, USA
[3] Department of Biology, University of North Carolina, Chapel Hill, NC 27599, USA

Taxonomists are arguably the most active annotators of the natural world, collecting and publishing millions of phenotype data annually through descriptions of new taxa. By formalizing these data, preferably as they are collected, taxonomists stand to contribute a data set with research potential that rivals or even surpasses genomics. Over a decade of electronic innovation and debate has initiated a revolution in the way that the biodiversity is described. Here, we opine that a new generation of semantically based digital scaffolding, presently in various stages of completeness, and a commitment by taxonomists and their colleagues to undertake this transformation, are required to complete the taxonomic revolution and critically broaden the relevance of its products.

Golden age of taxonomy

A decade has passed since a surge of informatics resources, catalyzed by the adoption of the World Wide Web, revolutionized taxonomy [1]: classifications now provide an extensible, digital scaffold upon which exponentially increasing data are integrated [2], infrastructure needed to disseminate biodiversity information has been established [3–5], and byproducts of the taxonomic process, mainly images, DNA and electronic keys, stand as new methods for specimen diagnosis [6,7]. The core missions of taxonomy (to name and classify units of biodiversity according to their evolutionary history and to support these hypotheses with data, most frequently phenotypes) remain highly relevant and more urgent than ever. The knowledge generated through this process is fundamental to all biology and must be made available to address questions outside of taxonomy.

Phenotypes in current taxonomy

Taxonomists examine specimens, sort them into taxonomic concepts worthy of names, organize their associated data (e.g. where and when they were collected and by whom) and then describe what they look like: leaves pinnate, femur longer than tibia, abdomen red, and so on. These phenome annotations (see Glossary) are compiled into 'descriptions', here including diagnoses, which mostly serve to formalize taxon concepts for future taxonomists.

Glossary

Alpha-taxonomy: the science of describing species (as opposed to higher-level taxa, which include multiple species).
Description: in the context of descriptive taxonomy, text describing some portion of the phenome of a taxon or specimen. A description typically comprises a relatively long list (often dozens) of character states that enable one to mentally reconstruct the phenome of a taxon. They are often written with the assumption that readers have sufficient background knowledge to interpret the domain-specific terminology used.
Determination: an application of a taxonomic name to a specimen by a particular person (or machine) at a particular time. For example, Jane Smith identifies an ant specimen as *Solenopsis invicta* in 2002 by writing this information on a piece of paper and attaching it to a physical specimen.
Diagnosis: a concise list of characters (usually less than ten) that differentiates one taxon from other, similar-looking and (usually) closely related taxa.
Domain expert: one who provides data from a particular area or topic. Within the realm of ontology development, this term is typically applied to biologists, rather than to individuals who create models or software for handling these data. For example, a shark taxonomist is a domain expert, who might contribute content to a fish anatomy ontology.
Evaluation metrics: a computable means of scoring the completeness or some other property of a body of work. In the present context, evaluation metrics might be used to determine whether a description is missing character annotations. For example, if all species descriptions in Genus A included statements about head color, a metric could be developed to remind future taxonomists to annotate that region of the phenome in all subsequent descriptions of new species in Genus A.
Entity: within the EQ model, a phenotype (typically an anatomical structure), about which some qualitative aspect is described.
Entity-Quality (EQ): a conceptual model for phenotypic description, pioneered in the context of model organism databases. This compositional approach allows independent development of organism-specific anatomy ontologies (providing entities) and generalized phenotypic trait ontologies (providing qualities).
Gaming mechanism: a software-based approach to formalizing data that uses a game with simple rules to enable players (typically non-experts) to produce scientifically meaningful output, through progress in the game. These games frequently exploit the highly refined ability of humans to pattern match (e.g. the puzzle website foldit, which uses gaming to understand protein folding; http://fold.it/portal/).
Natural language processing: a text-mining method, whose algorithms take into account rules of human grammar and syntax.
OWL (Web ontology language): a knowledge representation language built on formal logical semantics and designed for compatibility with the World Wide Web.
Phenobank: a database of semantic phenotypes. As envisioned here, phenobanks could be either specific domains within the Semantic Web or more narrowly defined specific-use databases (e.g. Phenoscape, http://kb.phenoscape.org).
Phenome: in the context of descriptive taxonomy, the set of all phenotypes expressed by a specimen or taxon. In contrast to the discretely characterizable genome (i.e. characteristics can be represented using only a few symbols), there are innumerable ways to describe a phenome.
Quality: within the EQ model, a characteristic property or trait value, such as 'shortened', 'round', or 'dark red', of a particular entity.
Rich search: computer searches that take into account the underlying structure of the data to both query and display results. For example, if one searches for images of 'eyes', the rich search might recognize, based on an anatomy ontology, that heads have eyes and, therefore, would return images annotated either as 'eye' or 'head'.
Text-mining: a general class of methods to extract discrete data from human-written text by use of computer algorithms.

Corresponding author: Deans, A.R. (andy_deans@ncsu.edu)

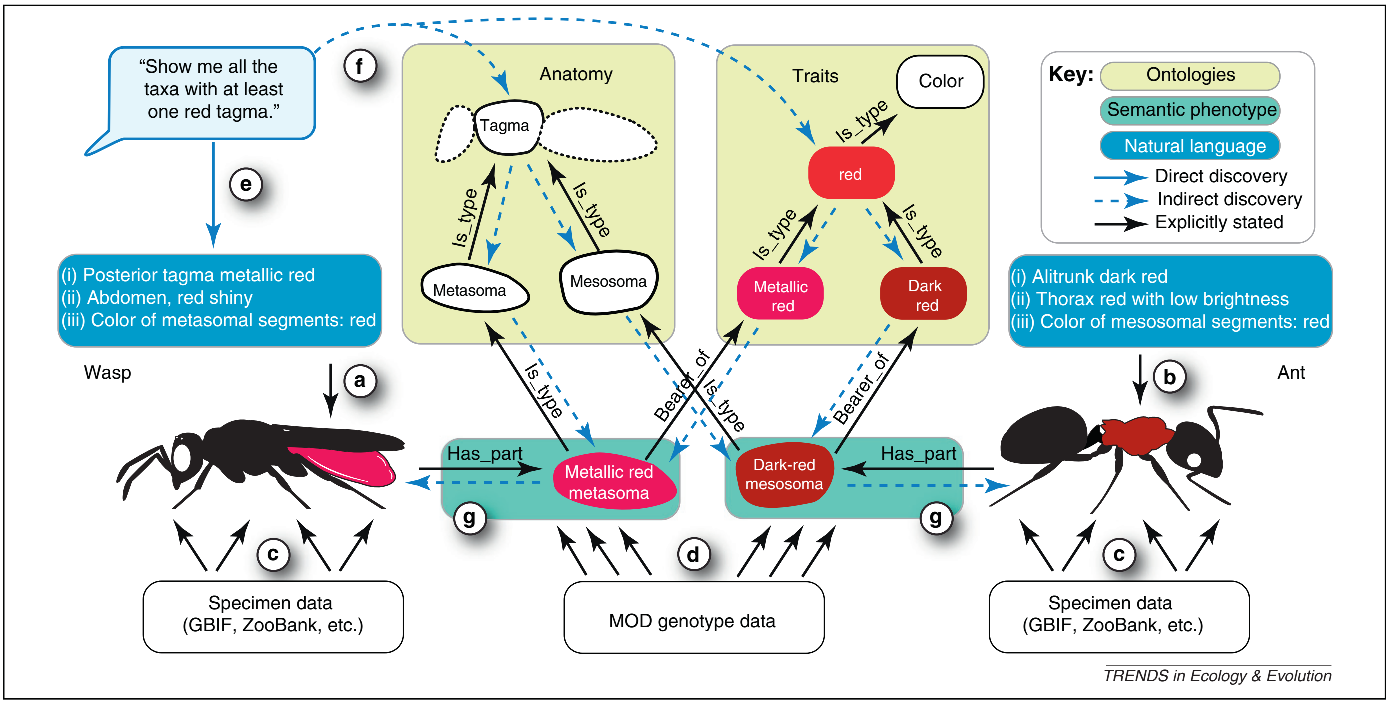
0169-5347/$ – see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2011.11.007

Descriptions are published in analog resources [at least for animals; see current International Commission on Zoological Nomenclature (ICZN) guidelines (http://iczn.org/code)] and are infrequently, if ever, used again. In fact, annotations are usually reproduced *de novo* when a taxon is revised, a redundant process that results in dozens of descriptions for some taxa, almost all of which are wasted data. Given this inefficiency, it is no wonder that many taxonomists consider descriptions to be a nuisance (see [8]) and, in some cases, have discussed taxon delimitation based exclusively on molecular data [7,9] (although not without concern [10]). The limited utility of descriptions (i.e. they are written for, and consumed almost exclusively by, taxonomists, who specialize on limited sets of organisms) may also explain, in part, the recurrent stagnation in funding and training (the taxonomic impediment [11–14]) and the subsequent low morale that plagues this fundamental science.

Three main issues prevent these phenome annotations from being broadly accessible: (i) they are published on paper {although this is rapidly changing [Cressey, D. (2011) Botanists shred paperwork in taxonomy reforms. *Nature News* 20 July 2011 (http://www.nature.com/news/2011/110720/full/news.2011.428.html); Rinaldo, C. and Norton, C. (2009) BHL, The Biodiversity Heritage Library: an expanding international collaboration. *Nature Precedings* 17 August 2009 (http://precedings.nature.com/documents/3620/version/1)] [15,16]); (ii) they are composed in natural language and, lacking standardization, are difficult to data mine effectively; and (iii) they typically do not reference explicit, logical definitions of concepts (i.e. homonymy and synonymy are rampant [17]). The wasp in Figure 1, for example, has a distinct reddish structure that taxonomists annotate in numerous ways: 'abdomen metallic red', 'metasoma shiny and red', 'posterior tagma with reflectance and red hue', and so on. Whereas a human reader with sufficient background in insect terminology can interpret these statements, text-mining tools would struggle to extract these annotations as equivalent [18].

Phenotypes in future taxonomy

What if all those descriptive data, minimally dozens of phenome annotations for every species on Earth, were available in a database (a 'Phenobank' or the Semantic Web), to be queried and repurposed for scientific questions? One might recognize this premise as a logical extension of the 'electronic page for each species of organism on Earth' of the Encyclopedia of Life [3]. We argue that the time has come to transform the way that biodiversity is described, to make phenotypes computable and linkable to the wider world of digital data. As a community, taxonomists publish more than 16 000 new, partially annotated phenomes (descriptions of new species) annually [19], alongside countless refinements to existing phenomes (redescriptions). This growing body of data could, if captured in a way that was semantic, extensible and broadly accessible, act as a foundation for huge discoveries across the life sciences. Questions of phenotype correlations, with each other, through time, or with the environment, for example, could be answered on a large scale if fine-grained, semantic

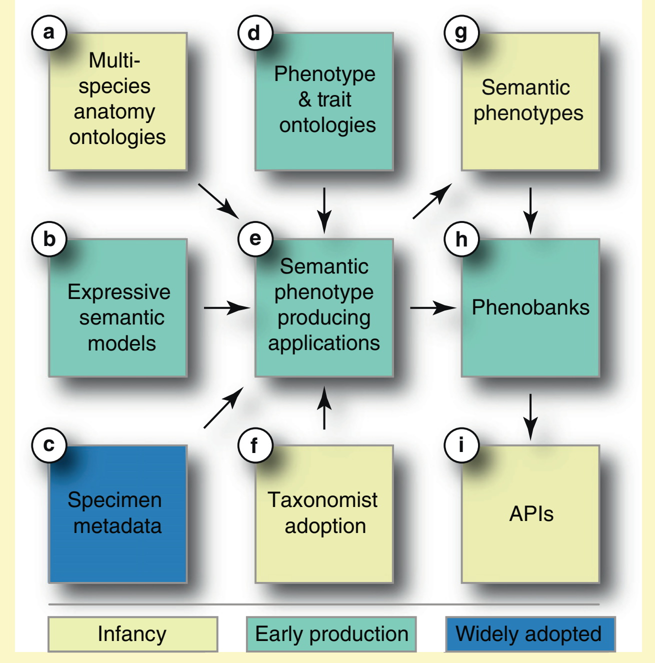


TRENDS in Ecology & Evolution

Figure 1. Benefits of semantic phenotypes. The wasp and the ant are described in natural language **(a,b)** and connected to data in existing databases **(c)**; this is the extent of current taxonomic practice. A biologist (upper left) queries against the corpus of biodiversity knowledge and fails to discover the ant when manually reading natural language **(e)**, but finds the wasp and the ant when using the logic inherent in ontologies to query across semantic phenotypes **(f)**. Model organism curators already collect genotype data and apply them to phenotypes using semantic methods **(d)** [45]. In our proposed method, taxonomists or other annotators produce phenome annotations **(g)** built following logical rules and referencing anatomy and trait ontologies (tan boxes). The example here is simplified to illustrate the major connections. The graph representing an actual semantic phenotype within an explicit logical model would be more complex: briefly, an individual specimen might be asserted to be an instance of an OWL class expression such as 'has_part *some* (metasoma *and* bearer_of *some* metallic_red)'. Abbreviations: GBIF, Global Biodiversity Information Facility; MOD, model organism database; OWL, Web ontology language.

Box 1. Supporting the production of semantic phenotypes by taxonomists

Multispecies anatomy ontologies [(Figure Ia), e.g. the HAO [17], Teleost Anatomy Ontology [46] and Plant Ontology [47]] and phenotype ontologies [(Figure Ib), e.g. PATO or Biospatial Ontology) contain the basic elements from which semantic phenotypes (Figure Ic) are built. These ontologies, which persist in OWL or Open Biomedical Ontology [48] (OBO) format, are combined [49] in meta-models [(Figure Id), e.g. EQ, see Figure 1, main text) through applications [(Figure Ie), e.g. Phenex [50], Phenote (http://phenote.org/) or Protégé (http://protege.stanford.edu/)] that ultimately export phenotypes to phenobanks [(Figure If), e.g. the Zebrafish Information Network [51]], or the Semantic Web. These applications must also manage and present a broad array of specimen metadata [(Figure Ig), e.g. collector, locality and habitat data) to the taxonomist. A major challenge is to develop applications that allow users to create semantic phenotypes (i.e. data with complex formats, see Figure 1, main text) with minimal effort. This step is key if we hope to establish broad adoption of this method by taxonomists (Figure Ih) necessary for the approach to succeed. Data from phenobanks are shared with the world through application programming interfaces [APIs (Figure Ii)] that return web-standard data (e.g. XML, RDF or JSON). The facilitation of taxonomists in this process requires investment in infrastructure; some components (specimen metadata) are well worked out, whereas others (supporting ontologies, applications and APIs) are in their infancy. As of yet, there are no point-and-click solutions to semantic phenotype production, but there are many opportunities to contribute to the development of this infrastructure at all stages.


a
Multi-species anatomy ontologies
d
Phenotype & trait ontologies
g
Semantic phenotypes
b
Expressive semantic models
e
Semantic phenotype producing applications
h
Phenobanks
c
Specimen metadata
f
Taxonomist adoption
i
APIs
Infancy
Early production
Widely adopted

TRENDS in Ecology & Evolution

Figure I. Infrastructure for semantic phenotype production.

phenotype (Box 1) data were available that were connected to a phylogenetic classification and specimen information for all species. Does organism 'fuzziness' exhibit latitudinal or altitudinal trends? Are species darker than they used to be? If flowers of a certain phenotype disappear, can one predict which pollinators would be at risk? Further networking these data with genotype–phenotype information of the model organism would open doors to dimensions of biodiversity previously unattainable. If this purple berry that grows on the south side of mountains yields a cancer-fighting agent, would one know where and when to look for other purple fruits? Is it possible to learn more about blindness in humans by studying a broader array of cave-dwelling organisms? What genes have been associated with wing development and patterning across insects and, therefore, would be good candidates for the genes controlling wing variation in one's study organism?

A best guess is that only 10–20% of Life has been discovered [3], and a recent calculation indicates that describing the remaining biodiversity (perhaps 8 million or more species, not to mention countless higher-level taxa) will cost approximately US$263 billion [19]. With this level of investment, are we going to encourage the continued collection and publication of hundreds of millions of data that are only marginally useful?

Lessons from biomedicine

Biomedicine suffers from similar symptoms, as the rate of data collection and publication, mostly in natural language, has rapidly outpaced the ability of researchers to synthesize emerging knowledge. To surmount this 'literature overload' [20], informaticians have invested primarily in a two-pronged *a posteriori* approach for retrieving information [18,21,22]: natural language processing [23,24] (which sacrifices precision for speed), combined with human-driven annotation of the literature [25–27] (which sacrifices speed for precision). Each model organism community hires curators, who peruse relevant journals, interpret the context and information presented therein, and then translate the data into a more structured format that can be queried across databases. A *Drosophila* lab, for example, might have five to seven curators, who regularly read a couple of dozen journals for information relevant to FlyBase [27,28]. This

Box 2. Ontologies and semantic phenotypes

An ontology is a formal representation of concepts within a domain and the logical relationships between those concepts. For anatomy, the concepts are typically body parts, each of which is formally represented by a structured definition or annotated image(s) and related to other body parts via properties, such as 'part_of' or 'is_a' (see [17; Washington, N. and Lewis, S. (2008) Ontologies: scientific data sharing made easy. *Nature Education* 1] and http://glossary.hymao.org for further explanation). In the HAO, for example, the concept 'protibia' bears the textual definition 'the tibia that is located on the fore leg' and is related to other anatomical parts as 'part_of fore leg' and 'is_a tibia'.

Semantic phenotypes are structured annotations that represent observations of the phenome. They are constructed using concepts from ontologies, in a model that facilitates computational analysis. Within a taxonomic framework, for example, the natural language description 'protibia color: dark red', as applied to wasp specimen A, could have the semantic phenotype annotation 'wasp specimen A has_part some (protibia and bearer_of some dark red)'. Protibia is drawn from the HAO (concept HAO:0000350) and dark red is from the PATO (concept PATO:0001261). Software can reason across the network of properties among these concepts, for example finding wasp specimen A as a result for a query on 'organisms with a red part of the fore leg' (because the protibia is part of the fore leg and, in wasp specimen A, the protibia is a bearer of a type of red).

process, although labor intensive and inefficient [29], has led to break-through discoveries in medicine and evolutionary biology [30,31], even when annotations are only at a very coarse level.

Taxonomists can do better; and, given the low level of funding currently allocated to biodiversity research, relative to genomics and medicine, and the fact that species descriptions are distributed across thousands of journals, they must do better. Text-mining already augments the production of semantic phenotype statements derived from the historical literature (e.g. [32]), a relatively new and growing area of research. However, what about future descriptions? Current annotations, which are usually telegraphic in style and typically present a body part and then the characteristics of that part, already mimic formalized annotations. Why not capture semantic phenotype statements at the very moment the phenotype is observed? There are myriad ways to create descriptive statements within this framework, and the tools and standards developed in other domains (Box 1) could be adapted to help meet the semantic phenotype challenge. As a starting point for discussion, we outline a knowledge representation method for taxon descriptions that is modeled after the Entity–Quality (EQ) [33] formalism originally developed for capturing gene expression and mutant data for model organism databases (MODs) and recently adapted for candidate gene prediction in evolutionary developmental biology [34].

Semantic phenotype descriptions for biodiversity

Ontologies (Box 2) are becoming well established within the biological sciences as tools for representing knowledge in a computable way, by providing shared definitions for concepts and the relationships between those concepts (for a primer on ontologies, see Washington, N. and Lewis, S. (2008) Ontologies: scientific data sharing made easy. *Nature Education* 1). Semantic phenotype annotations are constructed by selecting relevant concepts from source ontologies, singly or in combination, to denote logically the meaning of natural language descriptive statements. In the EQ approach, an entity (E) from an organism-specific ontology is associated with a quality (Q) from a generic trait ontology, a composition that has been shown to facilitate interoperability across the data of diverse research communities [31]. For the ant specimen in Figure 1, one could describe the color of the functional thorax in several ways using natural language (Figure 1b) but would annotate this free text with an explicit anatomical concept from the Hymenoptera Anatomy Ontology (HAO) [17] (mesosoma, which has the unique identifier HAO:0000576) and a quality class from the Phenotype and Trait Ontology (PATO) [33] (dark red, with unique identifier PATO:0001261). The natural language description remains available to the domain expert (i.e. taxonomy proceeds unabated), and the machine-readable annotation enters a system that facilitates computation across phenotypes.

Semantic annotations, as envisioned here, should ultimately be attached to individual specimens rather than taxonomic concepts. This added level of granularity will facilitate the development of logically based taxon concepts, in which membership of a taxon is defined via a set of necessary and sufficient trait values; an automated reasoner could be used to evaluate specimen membership, as well as to determine disjointness or subsumption of taxon concepts. It would have been inconceivable in the past to have taxonomists annotate thousands of individual specimens for a given study, but we anticipate that modern applications will ameliorate this problem through proxy mechanisms, such as rich search and filtering capabilities.

The broad benefits of semantic phenotypes

For taxonomists

By applying this (or a similar) model to descriptions, taxonomists would inherit a wealth of refined bioinformatics. Evaluation metrics based on knowledgebases (phenobanks in Box 1), for example, could be applied to elucidate the ‘completeness’ [29] of descriptions. Using the logic encoded within ontologies, taxonomists could be warned that proposed characters conflict with earlier statements (similar to the quasi-ontological application rules that have been employed for collaborative scoring of large phylogenetic matrices [35]) or that alternate terminology was used in a conflicting manner in previous publications. This logic could also be co-opted to drive smarter, more flexible diagnostic tools, where determinations are checked against prior publications for their sufficiency. Semantic phenotypes will also lead to a new level of rigor (see discussion in [36]) and integration (see discussion in [37]) among taxonomists, thanks in part to their very nature (i.e. they are rigorously defined and easily searched or filtered). In our experience, the exercise of encoding traditionally formulated taxonomic characters as semantic phenotypes has resulted in a more thoughtful, introspective refinement of characters [38].

It is premature to equate the adoption of semantic phenotype frameworks with predicted increases in taxonomic productivity, at least as measured by the number of taxa described. We do predict increases in efficiency, for example by ending the practice of redescription (but not the amending or appending of prior descriptions) and by informing a taxonomist’s choice of character systems on which to focus for diagnosis (e.g. which anatomical complexes seem to be the most informative for species delimitation in soil-dwelling, worm-like organisms?) We also predict that changing to this new workflow will rapidly increase the utility of taxonomy, which may therefore result in increased reinvestment in descriptive science. Our premise (and experience) is that meaningful taxonomy is constrained mainly by the limited supply of person-hours. The most efficient solution to speeding taxonomic production, ultimately, is to train and hire more taxonomists. Applications that facilitate big, new science by sharing the spoils of taxonomy through semantic bioinformatics just might catalyze a reinvestment in the taxonomic infrastructure needed to describe finally all of Life.

For all biologists

As alluded to above, formalized taxon descriptions will make available millions of explicit data for researchers interested in testing hypotheses of adaptation, phenotypic plasticity, development and medicine [39], as well as for developing new diagnostic tools. If applied to specimens, semantic annotations will connect phenotype to environment, phenology and even natural history data (through collecting event labels) and to evolutionary history through a classification that increasingly reflects phylogenetic relationships.

The discrete nature of semantic phenotypes predisposes them to broad use. That is, individual observations can be made outside of the collections produced by taxonomists and added to larger systems; semantic phenotypes make science more extensible. For example, discretely defined observations are more easily integrated into (i.e. displayed in the context of) the phylogeny of life. Semantic phenotypes become powerful when sampled across huge numbers of organisms. A recent study [40] reports on the large-scale citizen science project that looked at trends in snail color and patterning, wherein only a few simple annotations were captured for more than half a million specimens: shell color and shell banding. Imagine if there was a semantic version of these data but for a much larger swath of the biodiversity of Earth. Gaming mechanisms have already been proposed for annotating [41] and building the Semantic Web. An imaginative adaptation of these approaches could be adopted for the large-scale production and harvesting of phenotype data.

Finally, semantic phenotype data could also facilitate a better understanding of the scientific practice itself (as has been recognized in biomedicine [42]). Which character complexes do taxonomists tend to neglect and, therefore, deserve a more focused character exploration? Based on trends in phenotype descriptions over time, are there missed opportunities for collaborations between domains? Do certain high profile misinterpretations of anatomy persist in descriptions, even after they have been exposed?

Challenges

Logical limitations

Although we would like computable semantic descriptions to be capable of completely replacing free-form, natural language statements, in practice the two must coexist (at least for now), augmenting one another. Shortcomings in ontology content, as well as limitations of the logic frameworks themselves, often prevent the creation of semantic phenotypic descriptions that are expressive enough to stand in wholly for traditional statements, let alone form the basis for necessary and sufficient requirements for taxon membership. For example, the PATO ontology currently includes a simple class hierarchy representing specific colors (‘yellow’, ‘red’, ‘dark red’). As more precise phenotypic annotations are created, we anticipate that users of PATO will drive development of the ontology, such that axioms equating these color classes to specific value ranges of hue, saturation and brightness (for example) are included. A greater challenge is the ability of the logic frameworks to express all the conceptual relationships one might encounter within taxonomic characters. As an example, we have found that the inability to include variables within concept definitions using the Web ontology language (OWL) hinders satisfactory representation of characters that compare the length of one structure to another (‘antenna twice as long as eye height’). These logical limitations are certainly not showstoppers. On the one hand, ontology development is ideally a community process, and the content of ontologies can be continually improved to meet the needs of their users. The technological landscape of automated reasoning and rule systems is evolving rapidly, removing computational limitations along the way. On the other hand, semantic descriptions, even those that do not fully replace natural language descriptions, still provide nearly all of the computational benefits related to aggregating, querying and linking the data described above, including the case presented in Figure 1. With the current state of semantic annotation, by annotating the kind of phenotype denoted by a descriptive statement, rather than providing an ontological expression that can fully replace the descriptive statement, it is already possible to support a query such as, ‘Show me characters referencing anatomy that is part of the head’.

A note on homology

Taxonomists, who study their organisms through the lens of homology, may struggle to reconcile the relationship of homology to the concept hierarchy represented by an ontology. Existing anatomy ontologies are primarily developed using structural concept definitions, a strategy that works well in the single-species anatomy ontology case. The increased use of multi-species and higher-level anatomy ontologies, however, now forces taxonomists to address evolutionary issues that were previously irrelevant to ontology development. The hierarchical arrangements of anatomical classes themselves do not imply homology. Rather, homology hypotheses could be incorporated explicitly into the ontology via other (non-hierarchical) mechanisms, for example as accessory homology statements (i.e. in another data structure) that tie anatomical concepts to taxa in the context of a phylogeny. The integration of homology and anatomical classes is also expressed in natural language discussion, for example ‘class X in species A is homologous to class Y in species B as class Z in species C’. Based on our experience, the formalization of classes within multi-species anatomy ontologies forces one to review what was previously assumed to hold true for each anatomical concept, a process that ultimately forges stronger homology hypotheses. Homology does have important consequences as to the practical boundaries of a given anatomy ontology. Should there be many ontologies that each cover a small clade, for example, or a few large anatomy ontologies that cover major domains of life? This is a pragmatic issue, without obvious resolution. The semantics of homology statements and other strategies for addressing homology are under very active discussion by the phenotype ontology user and developer community (e.g. Phenotype RCN; http://phenotypercn.org/).

The taxonomist’s workflow

Semantic phenotypes are tied to specimens for both logical and data-modeling purposes. We acknowledge that this requirement introduces a major change to the typical alpha-taxonomy workflow. Very few taxonomists currently tie phenotype data to individual specimens. Rather, they imply that the material examined, which typically corresponds to populations or species, falls within the continuous range of variation provided by their descriptions. The technology needed to enable this shift in methodology exists already (i.e. databases and the Web), and could be programmed into workbench applications that facilitate specimen annotation. A larger issue is the cultural shift: how do we convince taxonomists that they need make a fundamental change in how they describe taxa? Nevertheless, carefully thought out and realized workbenches that accommodate the taxonomist's needs should significantly increase the probability of success.

Future directions

Our proposal to transform the descriptive process requires that significant new infrastructure be developed. Critics may worry that yet more resources will be diverted from taxonomic prospecting; the alternative, however, is to fund a practice (i.e. production of analog descriptions) that is trending towards irrelevance. Furthermore, it does not necessarily follow that these resources come from the same pool. Large meta-analyses based on interconnected databases of taxonomic names depend on the assumption that the supporting data are of equal quantity and quality, and this is clearly not the case right now. Meta-analyses based on the supporting data themselves would be far more powerful, eliminating the troublesome middleman (taxon name) that usually represents an unquantified hypothesis (i.e. the species circumscription). We have outlined a potential core infrastructure necessary for the production of semantic phenotypes (Box 1). Much of this infrastructure is undergoing a rapid evolution in parallel subdomains of the life sciences and, as such, the taxonomic world is well positioned to use it. Given the overall infancy of these methods and tools (Box 1, Figure I), there is little question that developing the virtual infrastructure and human capital necessary to meet this goal is a grand challenge. The task is not insurmountable; however, neither do we require a 'complete' infrastructure to make tangible progress towards semantic phenotype descriptions. The transformation will undoubtedly proceed in a stepwise fashion, beginning, for example, with references to concept uniform resource identifiers (URIs) in one's revisions [43] to provide explicit anchors for searching and indexing.

Concluding remarks

Taxonomists operate in the frontiers of biodiversity, making big discoveries, describing new species and bringing previously unknown phenotypes to the broader world. Almost every source of data that taxonomists use or generate has been digitized [37] or is the target of ongoing digitization efforts. Standards, databases and applications that facilitate collaborative taxonomy and the rapid publication of said data have been developed [44]. We can hardly imagine a more exciting field to be in, or a better time to be a taxonomist. Yet the most time-consuming, expansive and potentially transformative data that taxonomists collect, phenome annotations, remain largely unchanged since the days of Linnaeus and almost completely inaccessible by other domains of science. Most of the resources and know-how for making this data source broadly useful already exist, and it is time to catalyze a transformation in the way that biodiversity is described.

Acknowledgments

The ideas presented herein were developed during meetings funded by the US National Science Foundation (NSF; grants DBI-0850223, DEB-0956049) and the Biodiversity Synthesis Center of the Encyclopedia of Life, and in rigorous conversation with participants of the Phenotype RCN (http://www.phenotypercn.org). We thank the NSF-funded National Evolutionary Synthesis Center (EF-0905606) for providing the support for synergistic interactions. JPB is supported by the Phenoscape project (NSF DBI-0641025 and DBI-1062542), whose members have significantly contributed to refining these ideas. Matt Bertone assisted with the artwork in Figure 1. We are indebted to Katja Seltmann, István Mikó, and the anonymous reviewers, who provided invaluable critical feedback and insights that greatly improved the manuscript.

References

1 Bisby, F.A. (2000) The quiet revolution: biodiversity informatics and the internet. *Science* 289, 2309–2312

2 Patterson, D.J. *et al.* (2010) Names are key to the big new biology. *Trends Ecol. Evol.* 25, 686–691

3 Wilson, E. (2003) The encyclopedia of life. *Trends Ecol. Evol.* 18, 77–80

4 Page, R.D.M. (2010) Wikipedia as an encyclopaedia of life. *Organ. Divers. Evol.* 10, 343–349

5 Mindell, D.P. *et al.* (2011) Aggregating, tagging and integrating biodiversity research. *PLoS ONE* 6, e19491

6 MacLeod, N. *et al.* (2010) Time to automate identification. *Nature* 467, 154–155

7 Hebert, P.D.N. *et al.* (2003) Biological identifications through DNA barcodes. *Philos. Trans. R. Soc. Lond. Ser. B: Biol. Sci.* 270, 313–321

8 Evenhuis, N.L. (2007) Helping solve the 'other' taxonomic impediment: completing the eight steps to total enlightenment and taxonomic Nirvana. *Zootaxa* 1407, 3–12

9 Cook, L.G. *et al.* (2010) Need morphology always be required for new species descriptions? *Inv. Syst.* 24, 322–326

10 Brower, A.V.Z. (2010) Alleviating the taxonomic impediment of DNA barcoding and setting a bad precedent: names for ten species of '*Astraptes fulgerator*' (Lepidoptera: Hesperiidae: Eudaminae with DNA-based diagnoses. *Syst. Biodivers.* 8, 485–491

11 Agnarsson, I. and Kuntner, M. (2007) Taxonomy in a changing world: seeking solutions for a science in crisis. *Syst. Biol.* 56, 531–539

12 Coleman, C.O. *et al.* (2010) DELTA for Beginners: an introduction into the taxonomy software package DELTA. *ZooKeys* 45, 1–75

13 Godfray, H.C.J. (2007) Linnaeus in the information age. *Nature* 446, 259–260

14 Carvalho, M.R. *et al.* (2007) Taxonomic impediment or impediment to taxonomy?. A commentary on systematics and the cybertaxonomic-automation paradigm. *Evol. Biol.* 34, 140–143

15 Penev, L. *et al.* (2010) Semantic tagging of and semantic enhancements to systematics papers: ZooKeys working examples. *ZooKeys* 50, 1–16

16 Knapp, S. *et al.* (2007) Spreading the word. *Nature* 446, 261–262

17 Yoder, M.J. *et al.* (2010) A gross anatomy ontology for Hymenoptera. *PLoS ONE* 5, e15991

18 Rebholz-Schuhmann, D. *et al.* (2005) Facts from text – is text mining ready to deliver? *PLoS Biol.* 3, e65

19 Carbayo, F. and Marques, A.C. (2011) The costs of describing the entire animal kingdom. *Trends Ecol. Evol.* 26, 154–155

20 Hunter, L. and Cohen, K.B. (2006) Biomedical language processing: what's beyond PubMed? *Mol. Cell* 21, 589–594

21 Antezana, E. *et al.* (2009) Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief. Bioinform.* 10, 392–407

22 Alex, B. *et al.* (2008) Assisted curation: does text mining really help? *Pac. Symp. Biocomp.* 2008, 556–567

23 Frijters, R. *et al.* (2010) Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comp. Biol.* 6, 11

24 Zweigenbaum, P. *et al.* (2007) Frontiers of biomedical text mining: current progress. *Brief. Bioinform.* 8, 358–375

25 Hoehndorf, R. *et al.* (2011) PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res.* 39, 1–12

26 Jensen, L.J. and Bork, P. (2010) Ontologies in quantitative biology: A basis for comparison, integration, and discovery. *PLoS Biol.* 8, e1000374

27 Karamanis, N. *et al.* (2007) Integrating natural language processing with FlyBase curation. *Pac. Symp. Biocomput.* 2007, 245–256

28 Giles, J. (2007) Key biology databases go wiki. *Nature* 445, 691

29 Baumgartner, W.A. *et al.* (2007) Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics* 23, i41–i48

30 Mabee, P.M. *et al.* (2007) Connecting evolutionary morphology to genomics using ontologies: a case study from Cypriniformes including zebrafish. *J. Exp. Zool. B: Mol. Dev. Evol.* 308, 1552–5015

31 Washington, N.L. *et al.* (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol.* 7, e1000247

32 Cui, H. *et al.* (2009) Semantic annotation of biosystematics literature without training examples. *J. Am. Soc. Inf. Sci. Tech.* 61, 522–542

33 Mungall, C.J. *et al.* (2010) Integrating phenotype ontologies across multiple species. *Genome Biol.* 11, R2

34 Mabee, P.M. *et al.* (2007) Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol. Evol.* 22, 345–350

35 O'Leary, M.A. and Kaufman, S. (2011) MorphoBank: phylophenomics in the 'cloud'. *Cladistics* 27, 1–9

36 Vogt, L. *et al.* (2009) The linguistic problem of morphology: structure versus homology and the standardization of morphological data. *Cladistics* 26, 301–325

37 Padial, J.M. *et al.* (2010) The integrative future of taxonomy. *Front. Zool.* 7, 1–16

38 Mikó, I. and Deans, A.R. (2009) *Masner*, a new genus of Ceraphronidae (Hymenoptera: Ceraphronoidea) described using controlled vocabularies. *ZooKeys* 20, 127–153

39 Groth, P. and Weiss, B. (2006) Phenotype data: a neglected resource in biomedical research? *Curr. Bioinform.* 1, 347–358

40 Silvertown, J. *et al.* (2011) Citizen science reveals unexpected continental-scale evolutionary change in a model organism. *PLoS ONE* 6, e18927

41 Steggink, J. and Snoek, C.G.M. (2011) Adding semantics to image-region annotations with the Name-It-Game. *Multimedia Syst.* 17, 367–378

42 Rzhetsky, A. *et al.* (2008) Seeking a new biology through text mining. *Cell* 134, 9–13

43 Talamas, E. *et al.* (2011) Revision of the *Paridris nephta* species group (Hymenoptera, Platygastroidea, Platygastridae). *ZooKeys* 133, 49

44 Smith, V.S. *et al.* (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. *BMC Bioinform.* 10 (Suppl. 14), S6

45 Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29

46 Dahdul, W.M. *et al.* (2010) The Teleost Anatomy Ontology: anatomical representation for the genomics age. *Syst. Biol.* 59, 369–383

47 Jaiswal, P. *et al.* (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genomics* 6, 388–397

48 Smith, B. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255

49 Dahdul, W.M. *et al.* (2010) Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. *PLoS ONE* 5, e10708

50 Balhoff, J.P. *et al.* (2010) Phenex: ontological annotation of phenotypic diversity. *PLoS ONE* 5, e10500

51 Sprague, J. *et al.* (2008) The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.* 36, D768–D772