Application of Semantic Technology in Biodiversity Science
Anne E. Thessen (Ed.)
ISBN 978-3-89838-733-0
© 2018 AKA Verlag Berlin

Chapter 4

Taxonomy and the Production of Semantic Phenotypes

Matthew J. YODER[a], Michael B. TWIDALE[b], Andrea K. THOMER[c], Lars VOGT[d], Nico M. FRANZ[e], Jinlong GUO[b], Andrew R. DEANS[f], and James BALHOFF[g]
[a]Illinois Natural History Survey, University of Illinois
[b]School of Information Sciences, University of Illinois
[c]School of Information, University of Michigan
[d]Institut für Evolutionsbiologie und Ökologie, Universität Bonn, An der Immenburg 1, Bonn D-53121, Germany
[e]Arizona State University, School of Life Sciences
[f]Department of Entomology, Penn State University, University Park, Pennsylvania, USA
[g]University of North Carolina at Chapel Hill, Renaissance Computing Institute

Taxonomists produce a myriad of phenotypic descriptions. Traditionally these are provided in terse (telegraphic) natural language. As seen in parallel within other fields of biology researchers are exploring ways to formalize parts of the taxonomic process so that aspects of it are more computational in nature. The currently used data formalizations, mechanisms for persisting data, applications, and computing approaches related to the production of semantic descriptions (phenotypes) are reviewed, they, and their adopters are limited in number. In order to move forward we step back and characterize taxonomists with respect to their typical workflow and tendencies. We then use these characteristics as a basis for exploring how we might

create software that taxonomists will find intuitive within their current workflows, providing interface examples as thought experiments.

1. Introduction

Taxonomists, those who describe and organize Earth's biodiversity, offer a unique perspective on life's phenotypes. With little exception, their work references phenotype information to circumscribe natural units of biodiversity. A taxonomist's hypotheses (concepts) typically rely on phenotypes, and their descriptions of phenotypes therefore serve as core evidence of a taxonomist's science. The primary product of taxonomy is not seen as a set of phenotypes, however, but rather as the conclusion: the taxon concept. A taxonomist's work is "validated" when their taxon concepts are applied, i.e. when others classify the world into their conclusions.

Some have argued[1] that taxonomists are under-selling their work by not recognizing the importance of their supporting data, particularly their phenotypic descriptions. One way to increase the utility of a taxonomist's anatomical observations is to generalize them into a "semantic phenotype"[1–3], i.e. a formalized, typically logically rooted, representation of an anatomical concept that is, at minimum, adapted to computational exploration. Phenotype data of this nature have recently been instrumental to a range of broader scientific explorations, both foundational/theoretical (e.g.[4–8] and methodological (e.g.[9–17]).

There are costs to producing semantic phenotypes, however, as researchers must be trained in the concepts; tools and supporting infrastructures must be built; benefits must be outlined; and incentives must be determined and implemented. A full cost-benefit study is beyond the scope of this work, in part because it must start with the baseline cost of taxonomic products as they currently exist, and this alone is an exceedingly difficult analysis in and of itself (consider assigning costs to the complexity documented by[18,19] and see potential model in[20]). Ultimately, regardless of whether semantic phenotypes are "cost-effective", their exploration in the context of the taxonomic process will help uncover the complexities, and therefore, ultimately the costs underlying their production.

For our current purposes we posit that taxonomists do care about the potentiality of semantic phenotypes. In part, this work is an updated roadmap to ideas proposed in Deans et al.[1]; it also seeks to serve as a summary introduction to taxonomists who want background on the field. Our goal is to provide a critical assessment of where we are at now with respect to taxonomists (specifically) adopting the principles and practices surrounding their production. With this background in place we then focus on the premise that new technologies, specifically software user interfaces, could catalyze the production of semantic phenotypes. The problem is therefore a general one: moving a community to adopt a new technology. Therefore it is best to start with a base level understanding and characterization of what that community does in the absence of that technology. From this basis we propose specific technologies that we hope will seed future exploration and discussion.

2. Approach

Our approach is to leverage insights derived from three core areas: firstly, research undertaken during two NSF Advances in Biological Informatics Projects; secondly, from our day-to-day efforts as taxonomists who wish to adopt a philosophy that embraces the production of semantic phenotypes; and thirdly, from our day-to-day interactions with collaborators working on related fields (e.g. morphology). Here we refine these insights into summaries that represent core issues with respect to taxonomists producing semantic phenotypes.

We begin by briefly describing the technologies taxonomists have utilized to add semantic layers to their phenotype data (specifically, taxonomic descriptions). Many of these technologies are cited elsewhere in this book. Our focus is to highlight issues specifically related to the field of taxonomy, and to clearly identify areas where taxonomists looking to enter the field would hit stumbling blocks.

We then step back and focus on how we might craft technologies that would support taxonomists in ultimately adopting a workflow that produces semantic phenotypes. The argument is as follows: The use and production of semantic phenotypes is dependant on technological advances in several aspects of computing. Therefore, if taxonomists are to produce semantic phenotypes they must adopt new technologies. With the goal of encouraging taxonomists to adopt new technologies we identify general characteristics of taxonomic work that lend themselves to technological solutions. In other words, any new tool for taxonomists must fit into, complement and/or enhance their existing work practices. Therefore, an in-depth understanding of taxonomists' existing work practices is needed before we can build new technological solutions, such as tools that produce semantic phenotypes. This philosophy and approach draws from research in the fields of Human Computer Interaction and Computer-Supported Cooperative Work as used by researchers in Information Science. We draw from from interviews of over 35 taxonomists and our own experiences as taxonomists. Each identified characteristic is cross-referenced to perceived issues specific to the production of semantic phenotypes.

We conclude by briefly reviewing the role of software interfaces in the production of semantic phenotypes. Interfaces have largely been an afterthought in the development of scientific computing, yet some of us feel they might be key to developing new systems. The use and adoption of semantic phenotypes can be broken down into a more base problem in software design, that of the adoption of a very large and highly interconnected systems. Navigation, display, editing, and updating these types of networks are difficult problems that should have generalizable solutions elsewhere. We do not seek to propose specific solutions along these lines but rather point out potential avenues of exploration. Nearly all of the topics touched on here are worthy of their own review papers, this work is intended provide an index to these yet unrealized reviews.

3. Discussion

3.1. Current Status

A system architected to produce semantic phenotypes for taxonomists must contain certain components. These generally include the *formalizations* themselves, a means to *persist* one or more formalizations, and the wrapping *applications*. To fully realize the importance of semantic phenotypes we also require *computing* or reasoning engines. With respect to taxonomy we feel that all of these are truly in their infancy. In the future we may see the current systems completely replaced by alternative solutions seeking similar goals.

Of the **formalizations** that have been implemented specifically to treat phenotype descriptions, the majority have defined *data structure*, rather than data meaning. For instance, early efforts like Delta[21] were developed to output taxonomic descriptions from matrix-like data, and the TDWG-based Structure of Descriptive Data (SDD) standard[22] was developed to "allow capture, transport, caching and archiving of descriptive data in all the forms shown above, using a platform- and application-independent, international standard" (from `https://github.com/tdwg/sdd`) (we note that SDD has only been adopted by only a few applications and is not currently exploited in any larger repositories). By far the most commonly used data structure is Nexus ([23], which is used to store character matrices). Lucid[24] also utilizes a simple table format with additional markup formats. Other models like NeXML[25] have not yet been adopted beyond simple experiments. We believe that lack of adoption is a reflection of 1) a the lack of applications that produce data of a given structure, and 2) the lack of repositories with specific capabilities for exploiting the underlying semantics.

Two approaches have been used to produce *semantically* based taxonomic descriptions – that is, they provide a formalization that reflects meaning. Cui[26] introduced CharaParser, which uses an XML markup to represent the results of Natural Language Processing algorithms. Balhoff et al.[2] introduced an approach that links matrix-based data in NeXML to phenotype descriptions in OWL (Web Ontology Language) which reference anatomy (e.g. the Hymenoptera Anatomy Ontology,[27]) and phenotype ontologies. The model was extended and "practiced in a series of follow-up papers[28–31]. The two approaches have somewhat different goals, the former being focused on mining phenotypes from published works, the latter focused on providing *de novo* formalizations. There is no lossless translation between the two formats available. A third approach that seeks to describe individual part instances (e.g. a single individual's head) using RDF is under development based on ideas put forth by Vogt[8,32], and Vogt et al.[33–36]. While not targeting taxonomic descriptions specifically, it has clear potential to be applicable to them. Related performance metrics, specifically those that test issues of repeatability and cross-community compatibility are key (see[37,38]).

There are very few **software applications** that have been used by taxonomists to produce semantic phenotypes for the purpose of taxonomic description. Huang et al.[39] created the Ontology Term Organizer (OTO), a tool that lets users ontologize anatomical terms, with specific application for taxonomic characters. This evolved into the "Exploring Taxon Concepts" (ETC) project, which is arguably the most integrated approach to producing semantic phenotypes specifically by and for taxonomists[10]. Balhoff et al.[2] used the matrix editing functionality in mx[40] to export NeXML, which could be edited in in Protegè (https://protege.stanford.edu/), and linked to phenotype descriptions. Both approaches have only been used by their creators and a limited number of collaborators. While not specifically used by taxonomists to produce descriptions, Phenex[41] has been the most extensively used software to generate semantic representations of data from published works[18]. Those matrices were primarily produced by taxonomists, but perhaps more specifically for evolutionary studies rather than taxon descriptions. Morph-D-base, in active development, will produce highly semantic instance ontologies[42].

Datasets of semantic phenotypes are **persisted** locally (that is, stored on users' machines) as XML documents, either in OWL format (approach of Balhoff et al.[2]), or the ETC format (approach of Cui et al.[10]). The latter persists in the extremely generalized RDF format. There is no standard relational database schema for either approach. Both formats are machine-readable, and not manually editable or easily examined without an application that can read them. It is unclear whether a more or less human readable file format for semantic phenotypes would encourage their adoption. There are no repositories specifically aimed at serving taxonomists, for example archives of data and their underlying semantics. Taxonomists who wish to share or publically archive their semantic data are currently only have general (rather than taxonomy-specific) solutions such as Github and Dryad, or must publish them as supplementary material on DOI serving electronic journals. None of these archives yet expose the underlying semantics as a queryable database. That said, Balhoff et al. recently developed Phenoscape KB (http://kb.phenoscape.org/), a SPARQL-based endpoint for data derived from Phenex. Vogt et al.[43] are re-developing Morph-D-Base as a knowledgebase for generating, storing, distributing and querying semantic phenotypes described as "instance-anatomies", i.e. highly semantic graphs. These latter two efforts have the most potential to be adopted as a more general purpose repository for taxonomists' semantic phenotypes.

One of the key (though sometimes only implied) benefits of producing semantic phenotypes is that they make data **computable** in a variety of ways. For example, they may be indexed and searched at a more finely atomic level than possible with unstructured natural language, or reasoned across using logical inference. However, none of the existing *taxonomic* efforts have demonstrated such applications beyond simple use of OWL reasoners (e.g. Elk) to classify data into partonomy based categories[2]. Franz[44], Fig. 7) has experimented with reasoning via application of the Euler/X reasoning engine over phenotype data classified via the ETC framework. The work of Balhoff[9] and Dececchi et al.[45] may present the most convincing demonstration to taxonomists of the potential use of computation with respect semantic phenotypes. They demonstrate that, in combination with a organismal classification, one could infer many gaps in a taxon by character presence/absence matrix. This work implies that with computational reasoning, taxonomists may be able to describe more taxa, with fewer observations. Ramírez and Michalik[15] also present a methodological- and visualization-based approach to employing semantic phenotypes which may be particularly compelling to taxonomists.

In summary, the range of semantic phenotype technology currently available to support taxonomists producing or refining new anatomical descriptions, or seeking to exploit past descriptions via new analytics, is very narrow. We do not mean to imply that the field has not deeply explored important issues, but rather, that much work is needed to make usable tools and thereby gain broader adoption. While we have demonstrated[28,31] that the underlying approaches can be learnt and advanced by graduate students, we note that most of the key advances have been facilitated by a very small number of highly technical individuals. Thus there is a substantial bottleneck in training semantic phenotype "producers."

3.2. Taxonomists adopting technologies: characterizing and supporting taxonomic work

One way to get past this bottleneck is to step back and acknowledge that catalyzing significant change in a scientific field will take time. If we accept this idea then we can afford to pause and carefully consider how best to enhance taxonomists' existing work platforms and practices. This means designing new technologies that first and foremost help taxonomists do what they already do. Once engaged, those technologies can be leveraged to slowly guide their users toward the production and application of semantic phenotypes. In order to implement this strategy we must first generalize characteristics of taxonomists and the field of taxonomy.

To that end, we interviewed over 30 taxonomists and held multiple workshops in conjunction with several NSF ABI related grants ("Collaborative Research: ABI Innovation: Rapid prototyping of semantic enhancements to biodiversity informatics platforms", "NSF Advances in Biological Informatics. The Hymenoptera Ontology: part_of a transformation in systematic and genome science."). From the discussions and review of those interactions several themes consistently arose. The results of those efforts are being expanded upon in forthcoming publications, but are broadly summarized here in the specific context of semantic phenotype production. While we by no means claim this to be a comprehensive or unbiasedly derived list, we do feel that these particular characteristics of taxonomists and their work lend themselves to consequences for the development of technologies. For each characteristic we briefly describe potential consequences, categorized into "pros" and "cons", for the production of semantic phenotypes.

Taxonomists are integrators. The product of a taxonomist's work typically summarizes everything that is known about a taxon, or unit of biodiversity. We have found that numerous taxonomists have independently developed their own complex systems for integrating their data, and there is a high degree of convergence towards the need for large, integrative software tools. *Pro:* Taxonomists understand the difficulties of pulling together disparate types of data, and may view semantic phenotypes as just another kinds of data they need to integrate and work with. *Con:* Existing approaches to semantic phenotype production require a cobbling together of software and techniques. Consequently, their integration is going to be seen by taxonomists as requiring more work than their existing workflows.

Taxonomists are illuminators of the never before seen. By the very nature of their work, taxonomists must frequently describe things that have never before been recognized. This has critical consequences for workflows that reference semantic standards, in that those standards will almost certainly not express all that the taxonomist needs them to. *Pro:* Taxonomists are the perfect type of researchers to extend and expand underlying standards (e.g. anatomy ontologies). *Con:* Software and tools that build semantic phenotypes must allow the user to formalize their data using temporary standards which become fully realized after the fact. This may be quite challenging to implement, given current systems' challenges in handling semantic uncertainty.

Taxonomists are "within-ers", not betweeners. By this we mean that taxonomists are primarily interested in defining taxa such that they are "locally" recognizable; there is an assumption that one will start with an existing set of potential unknowns, not all unknowns. This is commonly illustrated in their work via statements like "species A has a smaller head than species B", or "A has a spine more curved than the spine of B". In this example, the description would be sufficient if a researcher only has A's and B's to look at, but insufficient if C's, D's and Z's are introduced. In other words, a lot of taxonomic description is about relative values rather than absolute values, and relative to other near species rather than relative to all species. *Pro:* Combining formalized semantics with natural language processing could help universalize this class of statements by identifying them to the taxonomists prior to their publication or by linking to broader ontologies that could appropriately contextualize relative terms and descriptors. *Con:* Taxonomists may push back against the need to make their semantic phenotypes to be more globally interpretable, because it may require changes to their methods or the language they're more comfortable using.

Taxonomists work iteratively. Taxonomists' workflows are decidedly non-linear. They continually return to past observations for refinement. *Pro:* In an integrated system this tendency could lead them to continually refine the supporting semantics (e.g. reference ontologies). *Con:* Referenced formalizations such as anatomy ontologies cannot be built as a first step, but need to be iteratively updatable in real time. This requires a complex software design pattern to properly be addressed.

Taxonomists implicitly reference the past. Through a lifetime of experience that includes not only exposure to published work, but also tacit knowledge gained through mentoship and collaboration, taxonomists work with anatomical concepts that are understood but that may have never been fully or explicitly defined (semantically or otherwise). *Pro:* Taxonomists are a source of previously unpublished concepts that could potentially be drawn out during their workflow. *Con:* It is unclear whether semantic phenotypes can be defined to fully reflect the intent or meaning as understood by the taxonomist.

Taxonomists are set builders. Taxonomists build diverse kinds of sets (e.g. their character matrices), in the mathematical sense. *Pro:* Computers excel at manipulating sets. Interfaces which mimic the natural way taxonomists build and interact with sets, for example aggregating and sorting through physical specimens, are largely unexplored. *Con:* There is not nearly enough logical exploitation of this principle. There is great potential to reason or compute over the datasets that are basic elements of a taxonomist's work. Semantic phenotypes should extend the types of sets a taxonomists can make and refine to explore their data.

Taxonomists are visually driven. Taxonomists transcribe concepts as text or annotated images for publications. Neither of these formats fully represent what the taxonomist understands and sees about these concepts. *Pro:* The space for developing novel visually based interfaces is for all intents and purposes completely unexplored. *Con:* Semantics for describing complex phenotypes may require extensions far beyond what has been done to date, such as models for 3D representations.

Taxonomists are short term localists and long term globalists. Taxonomic work at the level of species definition is about attempting precise delineations between similar species, by noting differences that allow a distinction to be made between very similar species, and not by describing in relation to all species. However this detailed work about often small and subtle differences to help avoid misclassification is then published as a contribution to a global collaborative effort of describing all the world's species, that has been successfully sustained over decades, even centuries. That globalised effort benefits from different approaches to integration and standardization than if the effort was solely focused as the level of say genera. *Pro:* The way the work is done allows for both local and global work. *Con:* the needs of both perspectives can lead to contradictory requirements at different points of doing the work.

3.3. Interfaces

To illustrate how we might practically apply the generalized principles discussed above, we conclude by describing interface concepts for a taxonomist's workbench that aids in the production and annotation of semantic phenotypes. These interfaces are premised on two of the principles identified above: that taxonomists are visual by nature, and that taxonomists often "localize" or describe their results in relative terms.

While there are a vast number of visually varying interfaces employed in software in general, those used by taxonomists tend to be simple: spreadsheets, documents, or database entry forms. As noted above, taxonomists are set builders; they build groupings of specimens, literature, anatomical descriptions, and images. Though some of this set-building takes place in a spreadsheet, it's just as likely to take place in on a physical table or lab. We propose a set building interface that functions more like a table than a spreadsheet: individuals in these sets could be manipulated in *virtual light tables* (Fig. 1). The concept here is that a taxonomist's physical workspace is often a table littered (literally) with specimens, pieces of paper, containers and tools. This physical space is configured for the task at hand: labelling specimens, exploring them for new diagnostic phenotypes, comparing them to descriptions in the past literature, and so on. Providing an analogous space within software may let taxonomists organize and examine their data in a more free-form, iterative manner and may support the iterative work of grouping, lumping and splitting, and determining of best distinguishing features within the software. While the obvious use of a virtual light table is to examine images, including zooming, rotating, annotating them, and moving them side by side, we can also envision laying out specimens, labels, notes or other data in a symbolic fashion. These layouts could be visually enhanced by the semantics that relate the core data. For example laying out specimens as circles on a light table, with colorization based on the number of measurements taken on them.

Taxonomists are constantly *labelling* these sets (e.g. paper attached to insect pins) both physically and virtually. A richer, more graphically immersive interface that facilitates annotation, i.e. labelling, is another potential means to bridge the relationship between physical and virtual workspaces. *Radial flyout menus* (Fig. 3) provide a unified framework for letting users quickly select from a range of types of inputs. We envision allowing the user to tag, take notes, add images, or qualify the quality of existing data/observations. When combined with a light table concept the first stages of taxonomic discovery (e.g. "these hairs are interesting", "red legs seems to group these specimens", "these specimens need a another look") become

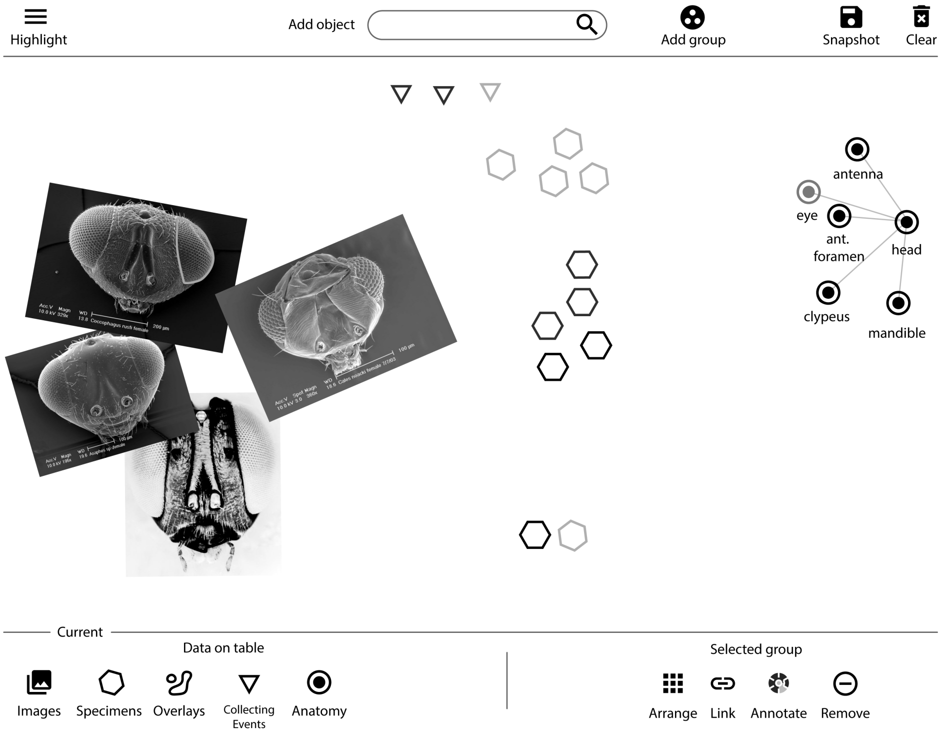


Figure 1. A virtual light table. The table space include images, and symbolic representations of specimens, collecting events, and an anatomy ontology arranged in a freeform open space. Basic functionality includes adding objects, highlighting objects, and saving the workspace (top), selecting, manipulating, and annotating the current contents of the table (bottom). See also Fig. 2

more digitally integrated. Taxonomists link the physical and digital world with paper labels, a flyout annotator could further allow the user to indicate that they want to queue a physical version of their annotation for print.

We know of only a few cases where taxonomists have sought to make 3D anatomical models a part of their workflow for describing taxa. While training taxonomists to describe taxa in 3D modellers (e.g. Blender, `https://www.blender.org/`) is likely some distance off, there is great potential for exploiting 3D spaces within workbench interfaces. We see their use falling into three categories: 1) spatially binding anatomical concepts to approximate their real-life position (e.g. Fig. 2-I) using symbolic representations of these spatially bound concepts to report metadata (e.g. as heat-mapped values); and 3) ex-

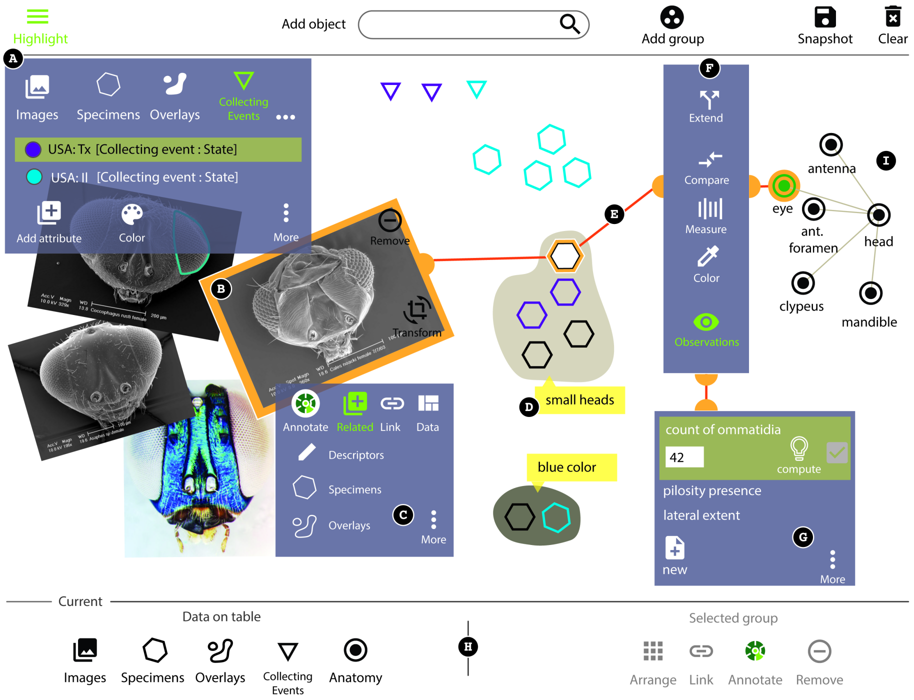
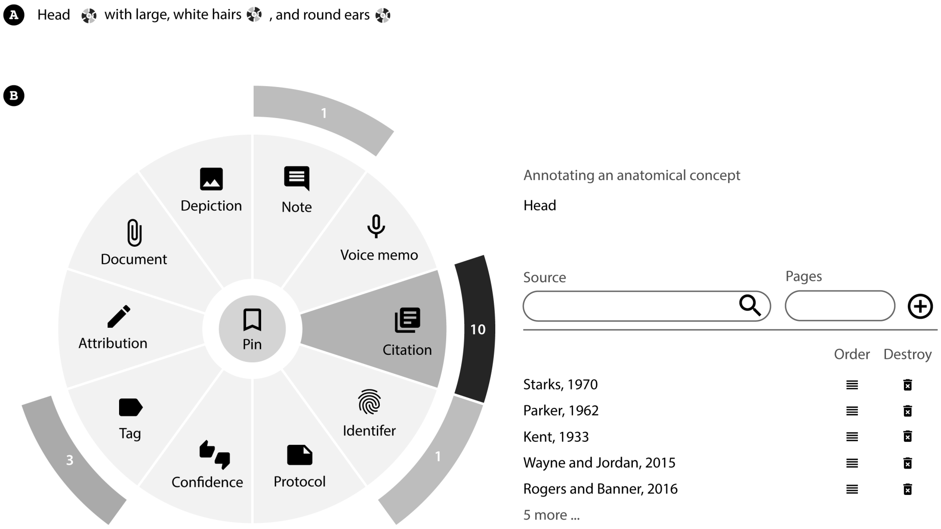


Figure 2. A virtual light table with actionable interfaces. Menus are presented as simultaneously open for display purposes, in practice they would open independently. The interface designed with touch-screen use in mind. A) Individual attributes can be selected and assigned a color, this highlights symbolic representations of the data, for example here we see specimens and collecting events from particular geographic areas correspondingly highlighted. B) Images can be manipulated in freeform. Clicking them brings up a menu C) which lets the user quickly add related data to the table, provide annotations, or new links (semantics). D) Arbitrary groups of data can be defined in freeform by the user, and given simple textual annotations. This is particularly important during the discovery phase of a taxonomist's research in which novel phenotypes are being understood and circumscribed for the first time. E) As an object is selected or function triggered linkages between symbolic data dynamically appear, allowing the user to quickly see and assign new observations at many different levels. F) Phenotypes can be quickly defined based on set classes (e.g. size, shape, color, and relative nature to other phenotypes). G) Traditional (e.g. qualitative or quantitative) observations can be gathered as well, these are precursors to semantic representations. H) Objects (data) and their groups on the table can be quickly selected and annotated. I) Anatomy classes can be displayed in correspondence to their physical position on the specimen.



Head, with large, white hairs, and round ears

Annotating an anatomical concept

Head

Source

Pages

	Order	Destroy
Starks, 1970	≡	🗑
Parker, 1962	≡	🗑
Kent, 1933	≡	🗑
Wayne and Jordan, 2015	≡	🗑
Rogers and Banner, 2016	≡	🗑
5 more ...		

Figure 3. A radial menu. A) The initial menu is compact, and can be inserted within complex natural language statements without overburdening the interface. B) On click a radial menu is opened. Clicking a slice of the menu opens the corresponding form. When the user is done the menu and form collapses in place. See implementation within the light table (Fig. 2-C.)

ploiting these same models to permit the user to navigate into specific phenotype description templates (this being particularly important with the advent of virtual headset technologies). Collectively these concepts could, again, provide a more intuitive parallel between the physical and digital world, closing the space between the abstractions in the mind of the taxonomist and their digital manifestations.

A major promise of novel interfaces is to provide new ways to express information that is currently almost exclusively shared in telegraphic natural language or annotated images. We envision navigating from a symbolic 3D representation of a taxon's anatomy into templates that map to specific phenotype types. To realize the full use of phenotype templates they need to be classified into categories that taxonomists frequently think of, for example, color, size, shape (e.g. Fig. 2-F). There are various classifications that could be used as the foundation for deriving phenotype template categories[7,8,32,34,46], though little has been done to provide a formal classification from which to base application development. Lessons from 3D modelling

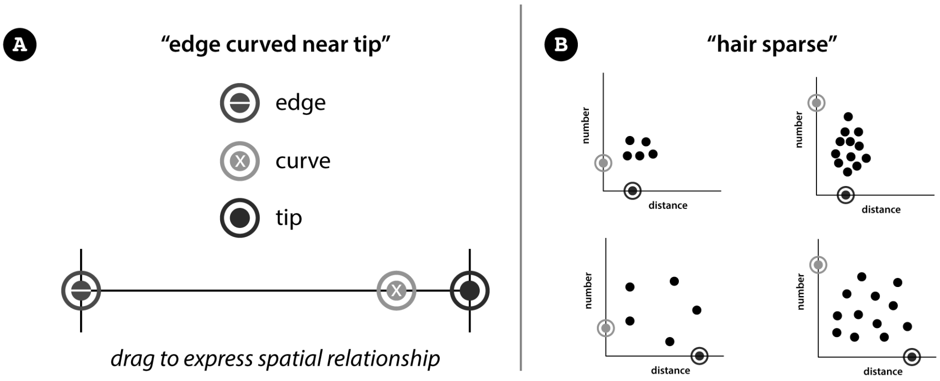


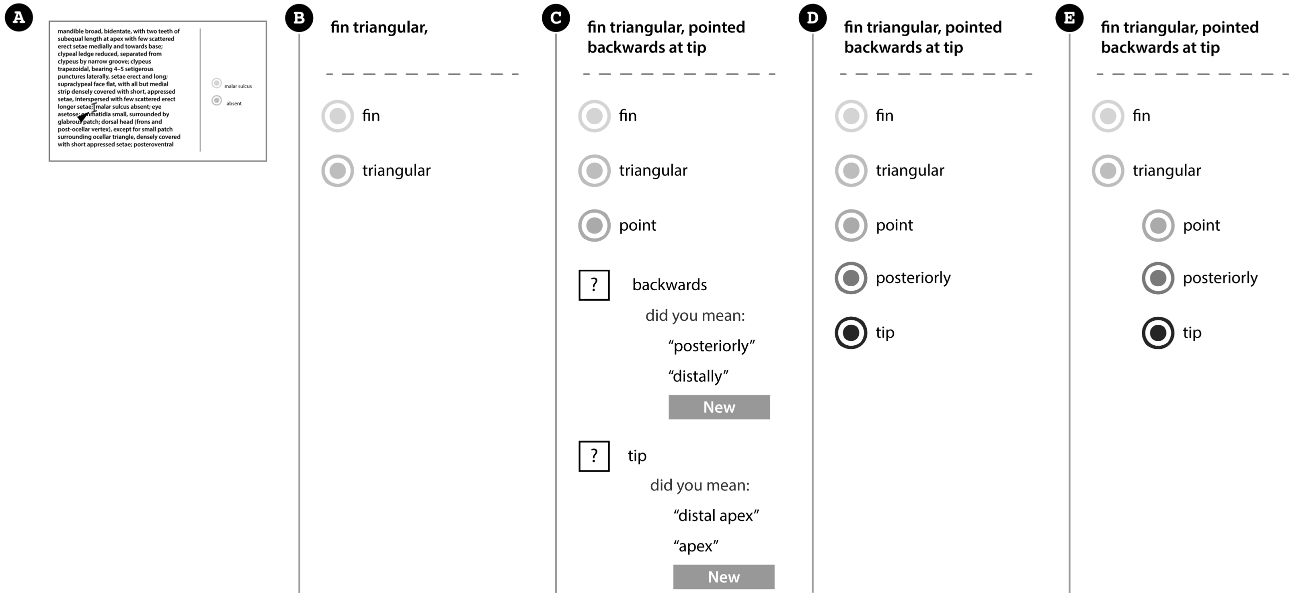
Figure 4. Symbolic interfaces. A) Statements like "edge curved near tip" are not globally comparable amongst taxa. Allowing the taxonomist to express a visual relationship via a simple interface (drag anchors along a line) quantifies the expression in a manner that is globally comparable. B) Statements like "hair sparse" are useful only in the context of extra metadata (images, figures). Simple interfaces that let the user choose an approximation of what they mean by "sparse", or tune their own approximations, result in a globally comparable quantification.

software (e.g. how that software lays out options for color, texture, size) should be explored. We anticipate that *symbolically based interfaces* (Fig. 4) hold potential for clarifying semantic phenotypes. These types of interfaces are specifically useful in cases where we are making "within" type comparisons (e.g. "edge curved near tip", "protrusion far from edge", "setae dense", "setae sparse"). Here the idea is to give the taxonomists the confidence to express what is a relative NL statement into a visual expression that is numerically quantifiable (e.g. has richer semantics). These expressions are not necessarily perfect, but they are formalized, and as such ultimately comparable in "between" studies.

A generalizable set of interface improvements are possible if we can exploit the semantics of existing data while the user is providing new data (sometimes referred to as reasoning on the fly). In this case assertions made by the taxonomists like "the specimen has an orange tongue" are parsed for semantic links to underlying semantics. The results can be fed back to the user via multiple mechanisms, possibly including visualizations using symbolic representations. Feedback can be of the autocorrect type (you said 'hair', did you mean 'setae') in which labels are tested against the concepts they are bound to

(Fig. 5), or in which detected concepts are being described in a way that is logically inconsistent (e.g. user error, “the head is attached to the leg” should not be a legal expression according to some ontology). A third type of auto-feedback is more complex and reflects the within/between dichotomy. In this case we can imagine a taxonomist describing a new taxon, while they populate a set of phenotypes those data are analyzed in real time against existing statements for related taxa. These analyses should be the basis for returning to the taxonomists and prompting them to 1) make new statements; 2) qualify existing statements; or 3) fix inconsistent statements regarding their phenotypes. For a trivial example, imagine a group of species presently diagnosed by head color. A taxonomist seeks to describe a new member of what he or she believes to belong in this clade. Upon completion, the software detects that the taxonomist has not provided a phenotype that references head, and it suggests that this be added so that all taxa in the clade, both previously and newly described, can be cross compared.

We conclude by imagining a more abstract interface that adds simple, but powerful semantics to the underlying data (Fig. 6-B). This interface’s goal is to allow the taxonomist to give their “within” (local) phenotype a “between” (global) context. This interface maps specific concepts like “roundness”, “blueness”, “straightness”, “nearness”, “hairiness” 1:1 with a globally accessible endpoint (knowledgebase). A taxonomist is prompted to slot his or her phenotype concept between existing concepts, in essence making assertions that “my phenotype is hairy, it is more hairy than this, but less hairy than that”. Within the interface they can step back and forth between nearby phenotypes, or make more radical jumps to something “much hairier”. Results from this type of character are not necessarily locally accurate, for example your concept of blue may differ from mine, but they are globally relative and scoped. The distribution of data in a particular endpoint (all blue things) can be broken into groups and given user-desired labels (Fig. 6-b). For example: 1) I’m using the label “light blue” for records 0...1000, 2) “blue” for records 1001...1020, and 3) “dark blue” for records 1021..20000. The user could choose to exclude records from series if they don’t fit the concept.



mandible broad, bidentate, with two teeth of subequal length at apex with few scattered erect setae medially and towards base; clypeal ledge reduced, separated from clypeus by narrow groove; clypeus trapezoidal, bearing 4–5 setigerous punctures laterally, setae erect and long; supraclypeal face flat, with all but medial strip densely covered with short, appressed setae, interspersed with few scattered erect longer setae; malar sulcus absent; eye asetose; ommatidia small, surrounded by glabrous patch; dorsal head (frons and post-ocellar vertex), except for small patch surrounding ocellar triangle, densely covered with short appressed setae; posteroventral

Figure 5. A simple, non-intrusive real-time concept matcher. A) The text box/screen, user provides standard telegraphic natural language in the editor on the left of the split screen, on the right real-time feedback matches text to the current statement based on the cursor location (see arrow); B) As the user types, concepts from existing standards are suggested; C) Some concepts can not be matched, and the user is prompted to select semantically similar ones, note that the user's phrase is not automatically changed, this allows them to express themselves in the manner they see fit, while adding a general level of semantics to their statement; D) The user has selected a related concept and also elected to create a new concept; E) The interface could be extended in many ways, for example it could allow the user to express the "nestedness" of the related concepts by allowing them to be indented via a dragging action.

Because these labels (breaks in the distribution) are specifically bound to a position relative to other data they are vastly more informative than had the data been provided without external reference. Furthermore, as a collective community expands the number of assertions within a particular endpoint each particular assertion becomes more powerful (there are now X more things to compare it against) and more precise, it's blue somewhere between these 1,000 examples on one side and those 1,000 examples on another versus "it's light blue". This approach is inherently a consensus building mechanism similar to CAPTCHA-based systems (when five users say the picture contains a dog, we can be confident that it does indeed contain a dog). If the system permits endpoints (e.g. "blueness") to be cloned, or split into new endpoints, then as end-users find problems with the distribution or definitions they can easily make assertions that they feel more confident with, i.e. the system can evolve. The beauty behind the system is that it allows a taxonomists to assert a broader context for their data by using a simple, intuitive interface with minimal decisions points: 1) bump my phenotype to the left; 2) bump it to the right; 3) leave it here, I'm done!

4. Summary

The production of semantic phenotypes originates either from the processing of previously published data - at this point exclusively natural language (though image post-processing is conceivable) - or from the taxonomists as they produce never before recorded observations. While in the former case the resultant utility of a given semantic phenotypes is greatly limited by the abilities of the parsing algorithm to interpret NL or by the annotator deriving a semantic phenotype from their understanding of NL statement, in the latter case what can be expressed is bound only by the interface. In other words the roadblocks preventing taxonomists from producing de novo semantics phenotypes are the lack of novel, imaginative interfaces. These interfaces must reflect the general principles that govern what a taxonomist does if they are to provide a system that resonates with the taxonomist.

The interface and functionality ideas outlined above are just ideas - points in a larger design space. However it is a region of the design space that we believe is worth exploring. The taxonomic work

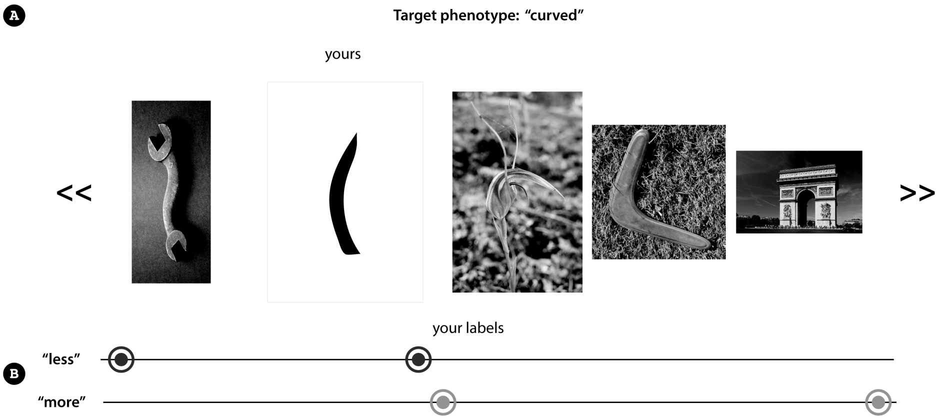


Figure 6. An interface to define "curved" using a global context. A) User selects "curved" and their target is randomly placed within the knowledgebase of curved things (as represented by images, textual descriptions, or other interpretable data). By sliding their target between other objects (concepts) that are asserted to "be curved" the user expresses the nature of their curve. B) If the user wishes they can provide labels for what they mean by the degree of "curved". The endpoints of these ranges are globally referenced within the knowledge-base, in theory automatically accessioning the user's data as a URI referenced object if desired.

process is extremely visual, involving numerous different comparisons by a trained expert eye in order to make appropriate, useful, actionable and replicable distinctions. This visual adjacency comparative work is finally translated into a textual description which has conventionally been free form natural language albeit using standardized terminology and structural conventions. We believe that software for taxonomists should support not only the product of structured text, but also the process of getting to that text, acknowledging its visual, comparative and iterative aspects. This somewhat structured natural language is relatively easily shared in worldwide databases aggregating the work of taxonomists across space and time. Semantic phenotypes offer great potential as a way to use even greater structure to support inferencing. Tools that make it easier for taxonomists to work towards both textual species definitions and develop semantic phenotypes without substantially increasing the work that the taxonomist must do are highly desirable. Additionally a tool that supports comparative iterative work enables a recording of all the steps along the way. Such a history makes it easier for a taxonomist to review her work process, recover from dead ends and revert to earlier possibilities, and benefit from reuse of work on very similar species. It also at least offers the possibility of helping others to learn by making more visible their own work practices and the work practices of more expert practitioners through visualizations of the twists and turns of the taxonomic work process.

In conclusion, we see no lack of interest from our fellow taxonomists and researchers, they want to build anatomy ontologies, formalize descriptions, and take advantage of the quantitative potential of the data, the vision presented here and throughout this book. However, evolving the few existing methods for producing semantic phenotypes into methods that can scale to meet this interest and demand remains a major challenge.

5. Acknowledgements

The work was made possible by a grant from the U.S. National Science Foundation (grant # DBI-1356381). Conversations with participants of the Phenotype Research Coordination Network (NSF 0956049) resulted in the seeds of many of the ideas presented here. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Special thanks to Jim Woolley, John Heraty, István Mikó, and Roger Burks for their ideas and feedback regarding interfaces. Images used here are from the Wikimedia Commons:

- Plant- By Glenda Wood [CC BY 3.0 au], via Wikimedia Commons
- Curved wrench - By Tomas Castelazo (Own work) [CC BY-SA 3.0] or GFDL (`http://www.gnu.org/copyleft/fdl.html`), via Wikimedia Commons
- Boomerang - Pic taken by Adrian Barnett, using Olympus C-2000Z), from `http://en.wikipedia.org/wiki/Image:Boomerang.jpg` Public Domain
- Arch - By Norio NAKAYAMA from Saitama, Japan (Arc de Triomphe Paris France) [CC BY-SA 2.0], via Wikimedia Commons

6. References

[1] Deans AR, Yoder MJ, Balhoff JP (2012) Time to change how we describe biodiversity. Trends in Ecology & Evolution 27(2):78-84.

[2] Balhoff JP, Mikó I, Yoder MJ, Mullins PL, Deans AR (2013) A semantic model for species description applied to the ensign wasps (Hymenoptera: Evaniidae) of New Caledonia. Systematic Biology 62(5):639-659.

[3] Deans AR, Mikó I, Wipfler B, Friedrich F (2012) Evolutionary phenomics and the emerging enlightenment of arthropod systematics. Invertebrate Systematics 26:323-330. doi:10.1071/IS12063.

[4] Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, et al. (2015) Finding our way through phenotypes. PLoS Biology 13(1):e1002033.

[5] Mikó I, Friedrich F, Yoder MJ, Hines HM, Deitz LL, Bertone MA, Seltmann KC, Wallace MS, Deans AR (2012) On dorsal prothoracic appendages in treehoppers (Hemiptera: Membracidae) and the nature of morphological evidence. PLoS ONE 7(1):e30137. doi:10.1371/journal.pone.0030137.

[6] Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA (2012) Uberon, an integrative multi-species anatomy ontology. Genome Biology 13(1):R5.

[7] Wirkner CS, Göpel T, Runge J , Keiler J, Klussmann-Fricke B-J, Huckstorf K, Scholz S, Mikó I, Yoder M, Richter S (2017) The first organ based ontology for arthropods (Ontology of Arthropod Circulatory Systems OArCS) and a semantic model for the formalization of morphological descriptions. Systematic Biology. doi: https://doi.org/10.1093/sysbio/syw108

[8] Vogt L (2016) Assessing similarity: on homology, characters and the need for a semantic approach to non-evolutionary comparative homology. Cladistics EarlyView. doi:10.1111/cla.12179

[9] Balhoff JP, Dececchi TA, Mabee PM, Lapp H (2014) Presence-absence reasoning for evolutionary phenotypes. arXiv preprint arXiv:1410.3862.

[10] Cui H, Xu D, Chong SS, Ramírez M, Rodenhausen T, Macklin JA, Ludäscher B, Morris RA, Soto EM, Mongiardino Koch E (2016) Introducing Explorer of Taxon Concepts with a case

study on spider measurement matrix building. BMC bioinformatics 17(1):471.

[11] Edmunds, RC, Su B, Balhoff JP, Eames BF, Dahdul WM, Lapp H, Lundberg JG, et al. (2015) Phenoscape: identifying candidate genes for evolutionary phenotypes. Molecular biology and evolution 33(1):13-24.

[12] Mabee PM, Balhoff JP, Dahdul WM, Lapp H, Midford PE, Vision TJ, Westerfield M (2012) 500,000 fish phenotypes: The new informatics landscape for evolutionary and developmental biology of the vertebrate skeleton. Journal of Applied Ichthyology 28(3):300-305.

[13] Manda P, Balhoff JP, Lapp H, Mabee PM, Vision TJ (2015) Using the Phenoscape Knowledgebase to relate genetic perturbations to phenotypic evolution. Genesis 53(8):561-571.

[14] Ramírez, MJ, Coddington JA, Maddison WP, Midford PE, Prendini L, Miller J, Griswold CE, et al. (2007) Linking of digital images to phylogenetic data matrices using a morphological ontology. Systematic Biology 56(2):283-294.

[15] Ramírez, MJ, Michalik P (2014) Calculating structural complexity in phylogenies using ancestral ontologies. Cladistics 30(6):635-649.

[16] Thessen AE, Bunker DE, Buttigieg PL, Cooper LD, Dahdul WM, Domisch S, Franz NM, et al. (2015) Emerging semantics to link phenotype and environment. PeerJ 3:e1470.

[17] Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE (2009) Linking human diseases to animal models using ontology-based phenotype annotation. PLoS biology 7(11):e1000247.

[18] Dahdul WM, Balhoff JP, Engeman J, Grande T, Hilton EJ, Kothari C, Lapp H, et al. (2010) Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. PLoS One 5(5):e10708.

[19] Franz NM (2014) Anatomy of a cladistic analysis. Cladistics 30(3):294-321.

[20] ten Hoopen P, Amid C, Buttigieg PL, Pafilis E, Bravakos P, Cerdeño-Tárraga AM, Gibson R, et al. (2016) Value, but high costs in post-deposition data curation. Database 2016.

[21] Dallwitz M (1980) A general system for coding taxonomic descriptions. Taxon 29:41-46. http://delta-

intkey.com/www/dallwitz-1980.pdf

[22] Hagedorn G, Thiele K, Morris R, Heidorn PB (2005) Structured Descriptive Data (SDD) w3c-xml-schema, Version 1.0. Biodiversity Information Standards (TDWG) `http://www.tdwg.org/standards/116`

[23] Maddison D, Swofford D, Maddison W (1997) NEXUS: An extensible file format for systematic information. Systematic Biology 46(4):590-621.

[24] Norton GA, Patterson DJ, Schneider M (2012) LucID: A multimedia educational tool for identification and diagnostics. International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International) 4(1).

[25] Vos RA, Balhoff JP, Caravas JA, Holder MT, Lapp H, Maddison WP, Midford PE, et al. (2012) NeXML: rich, extensible, and verifiable representation of comparative data and metadata. Systematic biology 61(4):675-689.

[26] Hong C (2012) CharaParser for fine-grained semantic annotation of organism morphological descriptions. Journal of the Association for Information Science and Technology 63(4):738-754.

[27] Yoder MJ, Mikó I, Seltmann KC, Bertone MA, Deans AR (2010) A gross anatomy ontology for Hymenoptera. PloS ONE 5(12):e15991.

[28] Mikó I, Trietsch C, Sandall E, Yoder MJ, Hines H, Deans AR (2016) Malagasy Conostigmus and the secret of scutes. PeerJ 4:e2682 doi:https://doi.org/10.7717/peerj.2682

[29] Mikó I, Masner L, Johannes E, Yoder MJ, Deans AR (2013) Male terminalia of Ceraphronoidea: diversity in an otherwise monotonous taxon. Insect Systematics and Evolution 44:261-347. doi:10.1163/1876312X-04402002

[30] Mikó I, Copeland RS, Balhoff JP, Yoder MJ, Deans AR (2014) Folding wings like a cockroach: a review of transverse wing folding ensign wasps (Hymenoptera: Evaniidae: Afrevania and Trissevania). PLoS ONE 9(5):e94056. doi:10.1371/journal.pone.0094056

[31] Trietsch C, Deans AR, Mikó I (2015) Redescription of *Conostigmus albovarius* Dodd, 1915 (Hymenoptera, Megaspilidae), a metallic ceraphronoid, with the first description of males. Journal of Hymenoptera Research 46:137. doi:10.3897/JHR.46.5534

[32] Vogt L (2010) Spatio-structural granularity of biological material entities. BMC Bioinformatics 11:289.

[33] Vogt L, Bartolomaeus T, Giribet G (2010) The linguistic problem of morphology: structure versus homology and the standardization of morphological data. Cladistics 26(3):301-325.

[34] Vogt L, Grobe P, Quast B, Bartolomaeus T (2012) Fiat or bona fide boundary - a matter of granular perspective. PLoS ONE 7(12):e48603.

[35] Vogt L (2017) The logical basis for coding ontologically dependent characters. Cladistics EarlyView. doi:10.1111/cla.12209

[36] Vogt L (2017) Towards a semantic approach to numerical tree inference in phylogenetics. Cladistics EarlyView. doi:10.1111/cla.12195

[37] Bertone MA, Mikó I, Yoder MJ, Seltmann KC, Balhoff JP, Deans AR (2013) Matching arthropod anatomy ontologies to the Hymenoptera Anatomy Ontology: results from a manual alignment. Database 2013:bas057. doi:10.1093/database/bas057

[38] Cui H, Dahdul W, Dececchi AT, Ibrahim N, Mabee PM, Balhoff JP, Gopalakrishnan H (2015) CharaParser+ EQ: Performance evaluation without gold standard. Proceedings of the Association for Information Science and Technology 52(1):1-10.

[39] Huang F, Macklin JA, Cui H, Cole HA, Endara L (2015) OTO: Ontology Term Organizer. BMC Bioinformatics. 16:47 doi:10.1186/s12859-015-0488-1

[40] Yoder MJ, Dole K, Deans AR “Mx” `http://purl.oclc.org/NET/mx-database`. Accessed 1 Sept. 2017

[41] Balhoff JP, Dahdul WM, Kothari CR, Lapp H, Lundberg JG, Mabee PM, Midford PE, Westerfield M, Vision TJ (2010) Phenex: ontological annotation of phenotypic diversity. PLoS One 5(5):e10500.

[42] Meid S, Baum R, Bhatty P, Grobe P, Köhler C, Quast B, Vogt L (2017) Developing a Module for Generating Formalized Semantic Morphological Descriptions for Morph-D-Base. Proceedings of TDWG 1:e15141. https://doi.org/10.3897/tdwgproceedings.1.15141

[43] Vogt L (2017) Assessing similarity: on homology, characters and the need for a semantic approach to non-evolutionary comparative homology. Cladistics 33:513-539. doi:10.1111/cla.12179

[44] Franz NM, Pier NM, Reeder DM, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher B. Two influential primate classifications logically aligned. Systematic biology 65(4):561-582.

[45] Dececchi TA, Balhoff JP, Lapp H, Mabee PM (2015) Toward synthesizing our knowledge of morphology: Using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. Systematic biology 64(6):936-952.

[46] Vogt L (2008) Learning from Linnaeus: towards developing the foundation for a general structure concept for morphology. Zootaxa 1950:123-152

[47] Vogt L (2009) The future role of bio-ontologies for developing a general data standard in biology: chance and challenge for zoo-morphology. Zoomorphology 128(3):201-217.

[48] Vogt L (2011) Signs and terminology: science caught between language and perception. Bionomina 4:1-41.