Phenex: Ontological Annotation of Phenotypic Diversity

James P. Balhoff[1,6*], Wasila M. Dahdul[1,2], Cartik R. Kothari[1,6], Hilmar Lapp[1], John G. Lundberg[3], Paula Mabee[2], Peter E. Midford[4], Monte Westerfield[5], Todd J. Vision[1,6]

1 National Evolutionary Synthesis Center, Durham, North Carolina, United States of America, **2** Department of Biology, University of South Dakota, Vermillion, South Dakota, United States of America, **3** Academy of Natural Sciences, Philadelphia, Pennsylvania, United States of America, **4** Department of Ecology and Evolutionary Biology.,University of Kansas, Lawrence, Kansas, United States of America, **5** Institute of Neuroscience, University of Oregon, Eugene, Oregon, United States of America, **6** Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

*corresponding author:

address:

National Evolutionary Synthesis Center

2024 W. Main St., Ste. A200

Durham, NC 27705

email: balhoff@nescent.org

Abstract

Background

Phenotypic differences among species have long been systematically itemized and described by biologists in the process of investigating phylogenetic relationships and trait evolution. Traditionally, these descriptions have been expressed in natural language within the context of individual journal publications or monographs. Thus, this rich store of phenotype data has been largely unavailable for statistical and computational comparisons across studies or integration with other biological knowledge.

Methodology/Principal Findings

Here we describe Phenex, a platform-independent desktop application designed to facilitate efficient and consistent annotation of phenotypic similarities and differences using Entity-Quality syntax, drawing on terms from community ontologies for anatomical entities, phenotypic qualities, and taxonomic names. Phenex can be be configured to load ontologies for different taxonomic groups. The graphical user interface was developed for, and tested by, evolutionary biologists accustomed to working with lists of taxa, characters, character states, and character-by-taxon matrices.

Conclusions/Significance

Annotation of phenotypic data using ontologies and globally unique taxonomic identifiers will allow biologists to better leverage decades of work in systematics and comparative morphology and contribute to an ever more useful web of linked biological data.

Introduction

The manifestation of evolution at the organismal level is the phenotype, or the set of observable traits inhering in an individual organism as a result of the interaction of heredity, environmental influences, and developmental processes. Biologists from different subdisciplines have approached the study of phenotypes in different ways. The unfolding of the phenotype from a fertilized egg is at the core of developmental biology, the inference of gene function through the phenotypic effect of allelic differences is a major focus of genetics, and using phenotypes to inform and interpret phylogenies in living and fossil organisms is at the core of systematics and comparative biology.

Despite the centrality of the phenotype to so much of biology, traditions for communicating information about phenotypes are idiosyncratic to different disciplines. Phenotypes seem to elude standardized descriptions due to the variety of traits that compose them and the difficulty of capturing the complex forms and subtle differences among organisms that we can readily observe. Consequently, phenotypes are largely inaccessible to the growing network of linked machine-readable biological data [1] and impenetrable to computational analysis. The problem is not simply a lack of syntactical conventions for phenotype data, but the lack of shared semantic conventions, as well.

An ontology is a type of structured vocabulary in which the terms and the logical relationships that hold between them are well-defined. Ontologies have become a foundational technology for capturing and computing with biological knowledge. They are particularly well-suited to providing machine-readable context for qualitative biological data. The Gene Ontology (GO) is the best-known example [2]. By describing gene products from many different organisms with GO terms, a very broad scientific community has been able to communicate knowledge about gene function in a way that is simultaneously human-readable and machine-readable, and innumerable applications have been developed that exploit this fact [3]. The widespread adoption of GO has demonstrated the value of ontologies for describing polymorphous and qualitative data.

The Open Biomedical Ontologies (OBO) Consortium serves as an umbrella organization for a wide variety of ontology efforts in the life sciences. OBO has promoted a set of principles for ontology development: they must be available for use without restriction, the community of users must have a means to improve them, they should be orthogonal in content, they must be instantiated in a well-specified syntax, and they must all share a common space of identifiers [4]. By following these principles, the ontologies that have been and are being developed in many different realms of biology (http://www.obofoundry.org/) can help overcome barriers to both human and machine communication across traditional boundaries (of discipline, language, etc.) and serve as a foundational technology for a web of linked biological data.

The genetic model organism community has pioneered the application of OBO ontologies to phenotype data, specifically for describing the extraordinary diversity of shape, size, position, composition, etc. in the observable physical characteristics in mutant genotypes relative to the wild type [5]. The 'Entity-Quality' (EQ) formalism is a rigorous syntactic convention well-suited to phenotype data [6,7], in which an entity term can be drawn from an organism-specific ontology (e.g. "fin", "vertebra", or "skull" from the Zebrafish Anatomy Ontology), and a quality term from the generic Phenotype and Trait Ontology (PATO) that describes the quality or value of some attribute of the entity (such as its color, size, shape, or number). EQ syntax has a number of desirable computational features [8]: it leads to compact representations because one need not specify qualities whose values are null, while at the same time there are no arbitrary limits on the number of attributes per entity; it supports complex queries over the structure of the ontologies; separating E and Q terms permits development of a small generic quality ontology independently of another relatively simple domain-specific entity ontology; and this separation also helps to insulate the system from schema instability. The model organism community has developed specialized software to represent phenotypes in EQ syntax. Phenote, for example, is used for phenotype annotation by curators at the ZFIN, WormBase, and Flybase databases [5].

Evolutionary biologists have compared and described phenotypic differences among species—extinct and extant—in the systematics and paleontological literature for many

decades [9]. One of the commonest and most formalized approaches within evolutionary biology is in the field of phylogenetic systematics where the variable organismal features (or *characters*) and their variants in different taxa (or *character states*), are itemized and given numeric codes in a *character-by-taxon matrix*. Importantly, the character and character state descriptions themselves are expressed in natural language. In the character-by-taxon matrix, the rows index *operational taxonomic units*, and the columns index *characters*. Individual cells contain a numeric code for a particular *character state*. For example, a group of species may vary in the character *opercle shape*, with some species exhibiting the character state *triangular* and represented by a '0' in the matrix, whereas other species may exhibit another character state *round* and be represented by a '1' in the matrix. Character-by-taxon matrices are analyzed for the purpose of recovering phylogenetic relationships or for examining patterns of character change on a given phylogeny.

Character-by-taxon matrices can be created or edited in software programs such as Mesquite [10] and MacClade [11] or web-based systems such as Morphobank (http://www.morphobank.org/) and MX (http://hymenoptera.tamu.edu/wiki/index.php/Main_Page). While the most common representation of these data is in NEXUS format [12], other standards have been developed for specific ends [13,14]. It is not yet general practice, and is not currently possible with many of these tools and standards, to link ontology terms to the text strings used to describe characters, character states, and taxonomic names. While free text strings are adequate for human interpretation (though not without ambiguity), they are semantically opaque to computers, and so preclude computational comparison of phenotype data across multiple evolutionary studies and the integration of these data with other biological knowledge on the web.

Here we describe a new platform for phenotypic data curation software named Phenex. The software was developed as part of the Phenoscape project (http://www.phenoscape.org), which is curating a large body of phenotypic descriptions from the legacy systematics literature using the EQ formalism [9,15,16]. Phenoscape has initially focused on capturing phenotype data from the ichthyological systematics

literature to enable linkages with the rich store of phenotype data from genetic studies in zebrafish [7]. A companion paper describes the standards that have been developed within the Phenoscape project for the curation of phenotypic data from the literature using ontologies, and the results obtained [15].

Methods

Software Design

Phenex was developed using agile development methodology—regular iterations of software improvement based upon continuous feedback from users. The initial requirements for Phenex grew out of early experiences using Phenote [5]. The major design difference in Phenex is the ability to easily manage data in the form of large character-by-taxon matrices. Development proceeded through multiple iterations, first extending Phenote with features to simplify management of multiple species, and then creating an application (using the same application framework) in which the interface was explicitly oriented around the character-by-taxon matrix. The resulting implementation resulted in the compartmentalization of data entry, with separate panels for characters and states, phenotypes, matrix, and specimens list. The requirements continued to be refined through input from individual curators in the course of their work, as well as during periodic data curation jamborees. The jamborees brought together domain experts with varying levels of familiarity with ontologies and EQ syntax. By pairing experts with Phenoscape personnel, we obtained user feedback on the overall curation workflow, the application of EQ syntax to character data, and the mechanics of working with the curation tool. User feedback led to changes in interface design, a variety of new features, and numerous bug fixes.

Implementation

Phenex is a desktop application that will run on systems with Java 5 or higher. It makes heavy reuse of the application framework developed for the OBO-Edit ontology editor [17], which provides the ontology object model, ontology reading capabilities, and interface framework. Like OBO-Edit, Phenex is open source, released under the MIT

license (http://www.opensource.org/licenses/mit-license.php). The Phenex homepage (https://www.phenoscape.org/wiki/Phenex) includes links to download the latest release as well as user documentation. Source code is available from the OBO project Sourceforge repository (https://sourceforge.net/projects/obo/). This paper describes version 1.0.

Input and Output

Phenex can import lists of taxa, characters and character states, as well as character-by-taxon matrices, from NEXUS, the most widely-used file format in systematics [12]. Phenex also exports the original character-by-taxon matrix as well as the free-text character and state descriptions and corresponding EQ statements in a tab-delimited format compatible with Microsoft Excel. The native Phenex file format is NeXML (http://www.nexml.org), an XML-based phylogenetic data exchange standard, based upon NEXUS, that provides a means to embed ontology-based annotations such as EQ statements within standard character-by-taxon matrix data. These embedded annotations adhere to the RDFa syntax for the inclusion of metadata within XML documents (http://www.w3.org/TR/rdfa-syntax/). Metadata relations from existing standards are used where possible to facilitate repurposing of the annotations for other applications (Table 1). Terms from the Darwin Core metadata standard (http://rs.tdwg.org/dwc/) are used to link taxa to unique identifiers for taxa and specimens (Figure 1A), and terms from the Dublin Core Metadata Initiative (http://dublincore.org) are used for general document information such as identification of the data curator. EQ phenotypes are embedded within character states and are serialized using the PhenoXML schema (http://www.fruitfly.org/~cjm/obd/formats.html), a standard for representing Entity-Quality phenotype descriptions developed by the OBO community (Figure 1B).

Results

Systematic characters and the EQ formalism

A core function of Phenex is to allow users to construct EQ statements, i.e. 'phenotypes' corresponding to each combination of character and character states present in a character-by-taxon matrix. While character/character state definitions and EQ statements are both used to represent phenotype descriptions, there is not a one-to-one mapping between the two. A character (such as *opercle shape*), typically includes both an entity term (*opercle*) as well as the particular variable attribute of that entity (*shape)* (Figure 2A, top). The character state (in this case, *triangular*), describes the value that the attribute takes in some specimen or taxon. In contrast, the attribute is implicit in the equivalent EQ statement (Figure 2A, bottom). This is possible because PATO represents qualities in a subtype hierarchy in which value qualities are specific kinds of attribute qualities [6,7]. For example, something that can be *triangular* in PATO is by necessity a type of *shape* (Figure 2B). It would be superfluous to qualify what attribute of the entity is variable in an EQ statement.

The EQ representation of a description such as "the opercle is approximately triangular in shape" [18] is formally represented as a type of the quality 'triangular' which *inheres_in* the entity 'opercle'. The phenotype is linked to the quality term *triangular* (PATO:0001875) from the Phenotype and Trait Ontology through the *is_a* relation, and the *opercle* term (TAO:0000250) from the Teleost Anatomy Ontology through the *inheres_in* relation. These relations are derived from the OBO Relations Ontology ([19]; http://obofoundry.org/cgi-bin/detail.cgi?id=relationship) and from proposed extensions to this ontology (http://www.bioontology.org/wiki/index.php/RO:Main_Page).

In applying the EQ formalism to systematic characters, it is helpful to distinguish several categories of characters that are commonly found in the systematics literature (Dahdul et al., in review). A character may fall into more than one of these categories. In describing these below, we use the following abbreviations: E, entity; Q, quality; C, count; RE, related entity. We describe the support that was built into Phenex for the representation of these characters in subsequent sections.

Monadic (non-relational) characters and states are those that involve single entities or anatomical structures. These characters are annotated with quality terms from PATO that are children of *quality of single physical entity* (PATO:0001237), such as *shape*, *size*, and *structure* and their children. For example, the caudal fin is described as having a deeply forked margin in some gonorynchiform fishes [20]. This is annotated as: E: *caudal fin*; Q: *bifurcated*.

Relational characters and states are those that involve two entities or anatomical structures. Such characters are annotated with quality terms from PATO that are children of *quality of related physical entities* (PATO:0001238) and these quality terms describe a phenotype that exists between two entities. For example, the two bones of the caudal fin (hypural 2 and hypural 3) are described as fused in some characiform taxa [21]. This is annotated as: E: *hypural 2*, Q: *fused_with*, RE: *hypural 3*.

Composite character states involve multiple phenotypes for a single character state. These character states can be monadic or relational. Systematists often describe multiple features in a character state to capture anatomical complexity, to document non-independent properties of a single anatomical entity, or to represent what is assumed to be a 'character complex'. For this reason, Phenex allows for multiple phenotype descriptions per character state. For example, the shape of the caudal fin of catfishes is described with the following three states [22]: forked with pointed lobes (0); forked with rounded lobes (1); scarcely emarginate to rounded (2). Each of these states, however, requires three distinct EQ phenotypes, e.g. state 0 is annotated with: E: *caudal fin*, Q: *bifurcated*, E: *caudal fin upper lobe*, Q: *sharp* and E: *caudal fin lower lobe*, Q: *sharp*.

Quantitative characters provide a literal value for a variable phenotypic feature (e.g., size, area, count). For example, characters involving counts of entities are annotated using the *count* quality and the literal values are recorded in the count field. Variation in meristics such as vertebral number are commonly described across species, e.g. [23]: state 0: 40-42; state 1: 43; state 2: 44-45. State 0 is annotated as: E: *vertebra*, Q: *count*, C: 40-42.

Use of the Phenex software for EQ annotation of phenotypes

Loading ontologies

Phenex can load any OBO-formatted ontology available on the web or in a local file, and it uses these ontologies for nearly every data type in Phenex. Using the ontology configuration panel, users can specify URLs from which Phenex should load ontology terms. The most recent version of each ontology is loaded each time Phenex is launched. Users can specify a term filter for each type of entry field, which determines the collection of terms provided as autocomplete suggestions for that field. These filters are commonly used to specify that an entry field uses terms which are drawn from a particular namespace or ontology subset (known as a "slim" in OBO parlance), or have specific relationships to other terms, so that only relevant suggestions are provided to the user. For example, the entity field only draws on terms from the Teleost Anatomy Ontology, Spatial Ontology, or Gene Ontology Biological Process namespace in the Phenoscape configuration of Phenex.

Interface

Phenex provides a straightforward and familiar graphical user interface (GUI) to evolutionary biologists who work routinely with lists of taxa, characters, character states, and character-by-taxon matrices. In this way, it differs from the closely related Phenote software, which is commonly used for EQ annotation of mutant phenotypes in the genetics community (Washington et al., in review). Both Phenote and Phenex inherit a flexible, modular interface design from the OBO-Edit application framework. The primary GUI components in Phenex include newly developed panels for editing taxa, specimens, characters, character states, character-by-taxon matrices, EQ statements, and literature citations, as shown in Figures 3-5 and described below. In addition, Phenex reuses several GUI components provided by OBO-Edit and Phenote that assist users to quickly find and evaluate ontology terms. These include graphical displays of term relationships ('Complete Ontology Tree View'), a sophisticated term 'Search Panel', and a textual 'Term Info' panel that displays synonyms, definitions, and relationships (Figure

3). All panels displaying ontology term information are updated with the currently selected term in the primary editing interface.

Taxa Panel: While phenotypic descriptions are always associated with particular taxa, the taxonomic names used in the publication may be obsolete or otherwise invalid (Dahdul et al., in review). Phenex allows the curator to relate the taxonomic names used in the original publication to currently valid names, as represented within a taxonomic ontology (Midford et al. in prep.). These correspond to 'Publication Taxon' and 'Valid Taxon', respectively, in Figure 4. Also, because many publications report phenotypes for higher taxa (genera, families, etc.) rather than species, Phenex allows entry of a 'Matrix Taxon' that corresponds to the set of species with specimens that were actually examined. Comments and illustrative figure references may be associated with a particular taxon entry.

Specimens Panel: It is standard and often required practice for publications in systematics to include a list of the voucher specimens on which phenotypic observations were based. Voucher specimens are deposited and catalogued (or registered) in the permanent collections of natural history museums. Future investigators may reexamine these physical specimens to validate or extend the original observations, or to retrieve additional information that pertains to each voucher, such as the collection locality. Recognizing the importance of this information to the reusability of the data, Phenex facilitates selection of a museum or institution code from a look-up table and manual entry of the catalog number for each voucher specimen reported in the publication (Figure 4). Increasingly museums have publicly accessible electronic collection databases allowing users to look up collection and other specimen metadata for cataloged specimens.

Characters and States-for-Character Panels: In each publication, the characters and character states are typically reported in a numbered list, where the numbers index the rows of the character-by-taxon matrix. The 'Characters' panel autonumbers free-text characters as they are manually entered, and the Comment column provides, e.g. for entry of the English translation of the character text when the original publication is in another

language (Figure 5). The Figure field allows for entry of references to any illustrative figures for that specific character. The 'States for Character' panel (Figure 5) requires manual entry of the symbol used for a character state. Typically the states are '0' or '1', but sometimes authors use 'a', 'b' or other variations. Similar to the 'Characters' panel, entry of free-text description of the character state, comments, and figures is enabled.

Phenotypes Panel: Phenex enables curators to create Entity-Quality statements for each character state in the Phenotypes panel (Figure 5). The ontologies provide terms for Entity, Quality, and Related Entity fields, thus supporting annotation of monadic and relational characters. The 'Add Phenotype' button (+ on Phenotypes panel, Figure 5) facilitates the entry of multiple phenotypes for a single character state, thus supporting annotation of composite characters. The values for quantitative characters may be entered in a structured manner using the free-text fields for counts and measurements. Units (e.g. millimeters, milligrams, etc.) are recorded within the ontology-enabled unit field.

To fully describe some characters, it can be helpful to use post-composed entities, new terms created from the semantic intersections of existing terms in one or more ontologies [24]. For example, many skeletal structures vary in the presence, shape, and size of their "margins", "processes", or "regions". Terms for these do not need to be defined exhaustively in the ontology, and thus post-composition is the preferred method for capturing these phenotypes. Variation in a skull bone, the epiotic, for example is described as [25]: 'Epiotic process, pointed (0) or bifurcated distally (1)'. 'Epiotic process' is represented as the post-composition 'process(part_of(epiotic))'. Post-compositions can be constructed on the fly in Phenex using a dedicated pop-up editor. The curator can add human-readable comments, to express, e.g., difficulties in interpretation of the published character description.

Application of Phenex to Phenoscape domain data

The aim of Phenoscape is to develop the means for comparison of phenotypic variation among species with phenotypic variants resulting from genetic manipulations in model organisms. This is achieved in part by expressing both forms of phenotypic data using

EQ syntax with terms and relations drawn from mutually-understood OBO ontologies. The initial focus of data curation in Phenoscape has been the Ostariophysi, a large clade of teleost fishes that includes zebrafish (*Danio rerio*), due to the richness of the systemic literature for these organisms and the abundance of EQ phenotype data already in the ZFIN database.

The Phenoscape curation workflow, of which Phenex was a critical piece, resulted in 12,397 high quality, consistent EQ annotations or 'phenotypes' for a large collection (47) of evolutionary character matrices (Dahdul et al., in review). For Phenoscape, Phenex was configured to load the Teleost Anatomy Ontology (TAO), Phenotype and Trait Ontology (PATO), and Teleost Taxonomy Ontology (TTO, Midford et al., in prep.) as well the Gene Ontology (GO), the Spatial Ontology (BSPO), the Relations Ontology (RO), the Evidence Code Ontology (ECO), the Unit Ontology (UO), and a list of museum identifiers. All the ontologies are available from the OBO Foundry, while the list of museum identifiers is publicly available (http://phenoscape.svn.sourceforge.net/viewvc/phenoscape/trunk/vocab/fish_collection_abbreviation.obo). Phenex is easily configured to generate EQ statements for a different set of organisms by simply loading different anatomy and taxonomy ontologies upon startup.

Phenex enabled effective division of labor between novices and experts within a collaborative curation workflow. Research assistants entered free-text data from systematic publications into Phenex for characters and character states, taxa, and voucher specimens; they also entered character-by-taxon matrices into Mesquite and exported these data in NEXUS format for import into Phenex. Character-by-taxon matrix data could also be entered directly using the 'Matrix' panel in Phenex. Using Phenex, ichthyological experts linked each taxonomic name used in the publication to the current valid taxon name (as represented in the TTO), and used the ontology search tools provided within Phenex to compose EQ statements corresponding to each character–character-state combination. The character-by-taxon matrix then automatically provided the mapping of EQ statements to individual taxa without curator intervention.

NeXML output from Phenex has been regularly uploaded to the Phenoscape Knowledgebase (http://kb.phenoscape.org). The logical structure of the ontologies, coupled with the deductive reasoning and the query interface provided by the underlying Ontology-Based Database (OBD) [24], enable simple and powerful queries across phenotype data from both systematic studies of the Ostariophysi and genetic experiments in zebrafish. Using the Phenoscape Knowledgebase (kb.phenoscape.org) to amalgamate EQ annotations created within Phenex demonstrates relationships among data from multiple studies that would be extremely difficult to discover without exploiting the structure of the ontologies (Figure 6).

Discussion

There is a wealth of information about phenotypic diversity among taxa to be found as text descriptions in the scientific literature, but it can only realize a fraction of its value in that medium. The challenge facing evolutionary bioinformatics is how to efficiently put this information into a form that allows comparisons of data across studies, and allows linkages to relevant data from other sources (such as genetically characterized phenotypes, geographic localities, phylogenetic relationships, etc.). User-friendly tools for curation of the systematic biology literature, such as Phenex, will be critical to the success of this effort. Annotating phenotypes and taxa from character-by-taxon matrices using ontologically defined terms and relations will open the door to a wide array of powerful ontology-driven applications for studying evolutionarily variable phenotypes.

While Phenex was designed in the context of a curation workflow for legacy data, the software may also be used to enter new character-by-taxon matrix data. While some extra effort would be necessary on the part of the researcher to construct EQ statements for the characters and character states, it would likely be more efficient and accurate than curation after the fact. Because Phenex can be configured to load terms from any OBO ontologies, data curation can be done in any taxonomic group as long as there exist appropriate anatomy and taxonomy ontologies. Employing Phenex for the creation of new datasets and within novel taxonomic groups will likely drive development of the Phenex user interface in ways that further increase the breadth of its applicability.

While ontologically annotated character-by-taxon matrices are tremendously useful for applications that compare data published by different authors at different times in different publications, there are challenges inherent in using these data to construct a phylogenetic *supermatrix* composed of EQ statements from multiple studies [26]. In particular, the dissociation of character states into EQs makes this difficult in two ways. First, the mapping between a character state and an EQ is not necessarily one-to-one, i.e. there may be multiple EQs for a single state, and this in and of itself precludes a simple substitution of EQs for character states in a matrix. Second, the *attribute* that forms part of a traditional character description is instead implicit in the hierarchical structure of the quality ontology (Figure 2) and so it is not straightforward to infer that two EQ descriptions represent alternative states for a character. On the one hand, this can be advantageous, because EQ descriptions from unrelated studies are readily combined into a unified knowledgebase (as we have done with the Phenoscape Knowledgebase), in contrast to the difficulty and uncertainty associated with combining characters from different character-by-taxon matrices. From this knowledgebase, related taxon–phenotype annotations can be easily discovered by searching higher level anatomical or quality terms. On the other hand, the comparative context provided by a character-by-taxon matrix, in which phenotypes stand as alternative values for an evolutionary character, is not as readily apparent. It remains an open question how best to to relate EQ phenotypes to alternative character states, and thus to aggregate EQs into supermatrices for subsequent phylogenetic analysis.

A valuable future extension to Phenex would be the ability to load ontologies in OWL format (http://www.w3.org/TR/owl2-overview/) in addition to OBO. While OWL is more widely used in the computer science and semantic web community, OBO is still far more prevalent among biological ontologies, and so this is not presently seen as a major deficiency. Irrespective of ontology format, the use of globally unique identifiers and rigorous semantics reduces ambiguity and promotes data integration across studies and domains. Related to the use of OWL ontologies for source concepts, Phenex could be enhanced with the ability to directly export datasets in RDF-triple format. The NeXML schema and the Comparative Data Analysis Ontology (CDAO) have been developed in

tandem so that it is straightforward to map NeXML data directly to CDAO concepts expressed in RDF [27]. Upon export, Phenex could augment these data with the ontological annotations it enables. In contrast to our current method of importing NeXML files into our OBD-based Knowledgebase, standard RDF triples would be more readily integrated with other data resources on the Semantic Web [28].

Directly exporting data as RDF triples would also make explicit some of the semantics that are currently only implied within our NeXML file format. Our Phenoscape Knowledgebase data loading software uses the mapping provided by the character-by-taxon matrix to create *exhibits* links between taxa and EQ phenotypes in our knowledgebase. Additionally, the loading software provides interpretation of embedded PhenoXML-structured data to create EQ phenotypes using the standard semantics, i.e. 'phenotype X' *is_a* 'quality Y' and *inheres_in* 'entity Z'. Ensuring a compatible data model among different users would increase interoperability across data sets.

It is our hope that current and future versions of Phenex will help put the vast stores of information about phenotype variation among taxa into a machine readable form, thereby enabling new directions in comparative biology and new contributions to the emerging web of linked biological data.

Acknowledgments

We thank the following for their contributions to the development of this software: Chris Mungall, Mark Gibson, Suzanna Lewis, Nicole Washington, John Day-Richter, Amina Abdulla, and Rutger Vos. We also thank the many workshop participants and systematists who provided examples and input that contributed to the Phenex user interface (see http://kb.phenoscape.org/contributors/).

References

1. Goble C, Stevens R (2008) State of the nation in data integration for bioinformatics. Journal of Biomedical Informatics 41: 687-693.
2. Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25: 25-29.
3. Azuaje F, Al-Shahrour F, Dopazo J Ontology-driven approaches to analyzing data in functional genomics. Methods Mol Biol 316: 67-68.
4. Ashburner M, Mungall CJ, Lewis SE (2003) Ontologies for biologists: a community model for the annotation of genomic data. Cold Spring Harb Symp Quant Biol 68: 227-235.
5. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, et al. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biology: in press.
6. Gkoutos GV, Green EC, Mallon AM, Hancock JM, Davidson D (2004) Using ontologies to describe mouse phenotypes. Genome Biology 6: R8.
7. Sprague J, Bayraktaroglu L, Bradford Y, Conlin T, Dunn N, et al. (2008) The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. Nucleic Acids Research 36: D768-D772.
8. Nadkarni PM, Marenco L, Chen R, Skoufos E, Shepherd G, et al. (1999) Organization of heterogeneous scientific data using the EAV/CR representation. Journal of the American Medical Informatics Association 6: 478-493.
9. Mabee P, Ashburner M, Cronk Q, Gkoutos G, Haendel M, et al. (2007) Phenotype ontologies: the bridge between genomics and evolution. Trends in Ecology & Evolution 22: 345-350.
10. Maddison WP, Maddison DR (2009) Mesquite: a modular system for evolutionary analysis. Version 2.6 http://mesquiteproject.org.
11. Maddison WP, Maddison DR (1992) MacClade: Analysis of phylogeny and character evolution, version 3.0. Sunderland, Massachusetts: Sinauer.
12. Maddison DR, Swofford DL, Maddison WP (1997) NEXUS: an extensible file format for systematic information. Syst Biol 46: 590-621.
13. Dallwitz M (1980) A general system for coding taxonomic descriptions. Taxon 29: 41-46.
14. Hagedorn G, et al. (2005) The Structured Descriptive Data (SDD) w3c-xml-schema, version 1.0. Available from: http://www.tdwg.org/standards/116/.
15. Dahdul WM, Lundberg JG, Midford PE, Balhoff JP, Lapp H, et al. The Teleost Anatomy Ontology: Anatomical representation for the genomics age. Systematic Biology: in press.
16. Mabee PM, Arratia G, Coburn M, Haendel M, Hilton EJ, et al. (2007) Connecting evolutionary morphology to genomics using ontologies: A case study from Cypriniformes including zebrafish. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 308B: 655-668.

17. Day-Richter J, Harris M, Haendel M, The Gene Ontology OBO-Edit Working Group, Lewis S (2007) OBO-Edit - an ontology editor for biologists. Bioinformatics Applications Note 23: 2198-2200.
18. Fink SV, Fink WL (1981) Interrelationships of the Ostariophysan fishes (Teleostei). Zoological Journal of the Linnean Society 72: 297-353.
19. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, et al. (2005) Relations in biomedical ontologies. Genome Biology 6: R46.
20. Grande T, Poyato-Ariza FJ (1999) Phylogenetic relationships of fossil and Recent gonorynchiform fishes (Teleostei: Ostariophysi). . Zoological Journal of the Linnean Society 125: 197-238.
21. Toledo-Piza M (2007) Phylogenetic relationships among Acestrorhynchus species (Ostariophysi: Characiformes: Acestrorhynchidae). Zoological Journal of the Linnean Society 151: 691-757.
22. Lundberg JG (1992) The phylogeny of ictalurid catfishes: A synthesis of recent work. In: Mayden RL, editor. Systematics, Historical Ecology, and North American Freshwater Fishes. Stanford: Stanford University Press. pp. 392-420.
23. Smith GR (1992) Phylogeny and biogeography of the Catostomidae, freshwater fishes of North America and Asia. In: Mayden RL, editor. Systematics, HIstorical Ecology, and North American Freshwater Fishes. Stanford, CA: Stanford University Press. pp. 778-826.
24. Mungall CJ, Gkoutos GG, Smith CL, Haendel MA, Lewis SE, et al. Integrating phenotype ontologies across multiple species. Genome Biology: in press.
25. Royero R (1999) Studies on the systematics and phylogeny of the catfish family Auchenipteridae (Teleostei: Siluriformes). Bristol: University of Bristol.
26. de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. Trends in Ecology and Evolution 22: 34-41.
27. Prosdocimi F, Chisham B, Pontelli E, Thompson JD, Stoltzfus A (2009) Initial implementation of a comparative data analysis ontology. Evolutionary Bioinformatics 5: 47-66.
28. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, et al. (2007) Advancing translational research with the Semantic Web. BMC Bioinformatics 8: S2.

Figure Legends

Figure 1. NeXML fragments demonstrating embedded Phenex annotations. A. A taxon. B. A character and character state.

Figure 2. Correspondence between Entity-Quality statements and evolutionary characters. A. Comparison of the structure of phenotypic descriptions using character-character state vs. Entity-Quality (= 'Phenotype') syntaxes. B. The defined relationship between an attribute quality type (shape) and a value quality type (triangular) within the Phenotype and Trait Ontology (PATO).

Figure 3. Phenex screenshot of window configured with panels for browsing and searching of ontology terms and relationships. Note that users can configure the position and size of each panel on the fly. See text for details of each panel.

Figure 4. Phenex screenshot of window configured with panels for editing of taxon lists, voucher specimens, and publication information.

Figure 5. Phenex screenshot of window configured with panels for editing of character and character states data, phenotypes (i.e. EQ statements), and character-by-taxon matrix.

Figure 6. An example of lexigraphically dissimilar phenotype descriptions from two publications that are semantically similar in that they pertain to the same anatomical structure. The 'dorsal arrector' and the 'posterior pectoral-spine serrae' are both parts of the pectoral fin, which is immediately apparent to both humans and computers from the structure of the anatomy ontology. Some of the data relationships shown, such as *PHENOSCAPE:exhibits* and those from CDAO (Comparative Data Analysis Ontology, [27]), are not explicit in Phenex. Instead, these are generated by the

interpretation of NeXML documents within the Phenoscape Knowledgebase data loading software.

Tables

Table 1. Metadata identifiers and XML elements used by Phenex to embed annotations with NeXML documents.

Identifier	Namespace	Source	Usage
creator	http://purl.org/dc/terms/	Dublin Core	Relates NeXML document to curators of content.
references	http://purl.org/dc/terms/	Dublin Core	Relates NeXML document to publication source.
description	http://purl.org/dc/terms/	Dublin Core	Relates NeXML document to contents of Phenex "Publication Notes" field.
taxonID	http://rs.tdwg.org/dwc/terms/	Darwin Core	Relates each NeXML taxon to the OBO identifier used for the Phenex "valid name".
individualID	http://rs.tdwg.org/dwc/terms/	Darwin Core	Relates each NeXML taxon to each specimen entry.
collectionID	http://rs.tdwg.org/dwc/terms/	Darwin Core	Relates each specimen entry to a museum collection OBO identifier.
catalogNumber	http://rs.tdwg.org/dwc/terms/	Darwin Core	Relates each specimen entry to an accession code for a museum collection.
comment	http://www.w3.org/2000/01/rdf-schema#	RDF-Schema	Relates NeXML taxon, character, and state elements to curator comments.
hasMatrixName	http://vocab.phenoscape.org/	Phenoscape	Relates each NeXML taxon to the Phenex "matrix name".

inFigure	http://vocab.phenoscape.org/	Phenoscape	Relates NeXML taxon, character, and state elements to figure references.
describesPhenotype	http://vocab.phenoscape.org/	Phenoscape	Relates NeXML state elements to a block of PhenoXML data representing EQ phenotypes.
phenotype	http://www.bioontologies.org/obd/schema/pheno	PhenoXML	Represents a collection of EQ statements.
phenotype_character	http://www.bioontologies.org/obd/schema/pheno	PhenoXML	Represents a single EQ statement.
bearer	http://www.bioontologies.org/obd/schema/pheno	PhenoXML	Represents the entity component of an EQ statement.
quality	http://www.bioontologies.org/obd/schema/pheno	PhenoXML	Represents the quality component of an EQ statement.
related_entity	http://www.bioontologies.org/obd/schema/pheno	PhenoXML	Represents the related entity component of a relational EQ statement.
typeref	http://www.bioontologies.org/obd/schema/pheno	PhenoXML	Represents a reference to a particular OBO ontology term or post-composition.

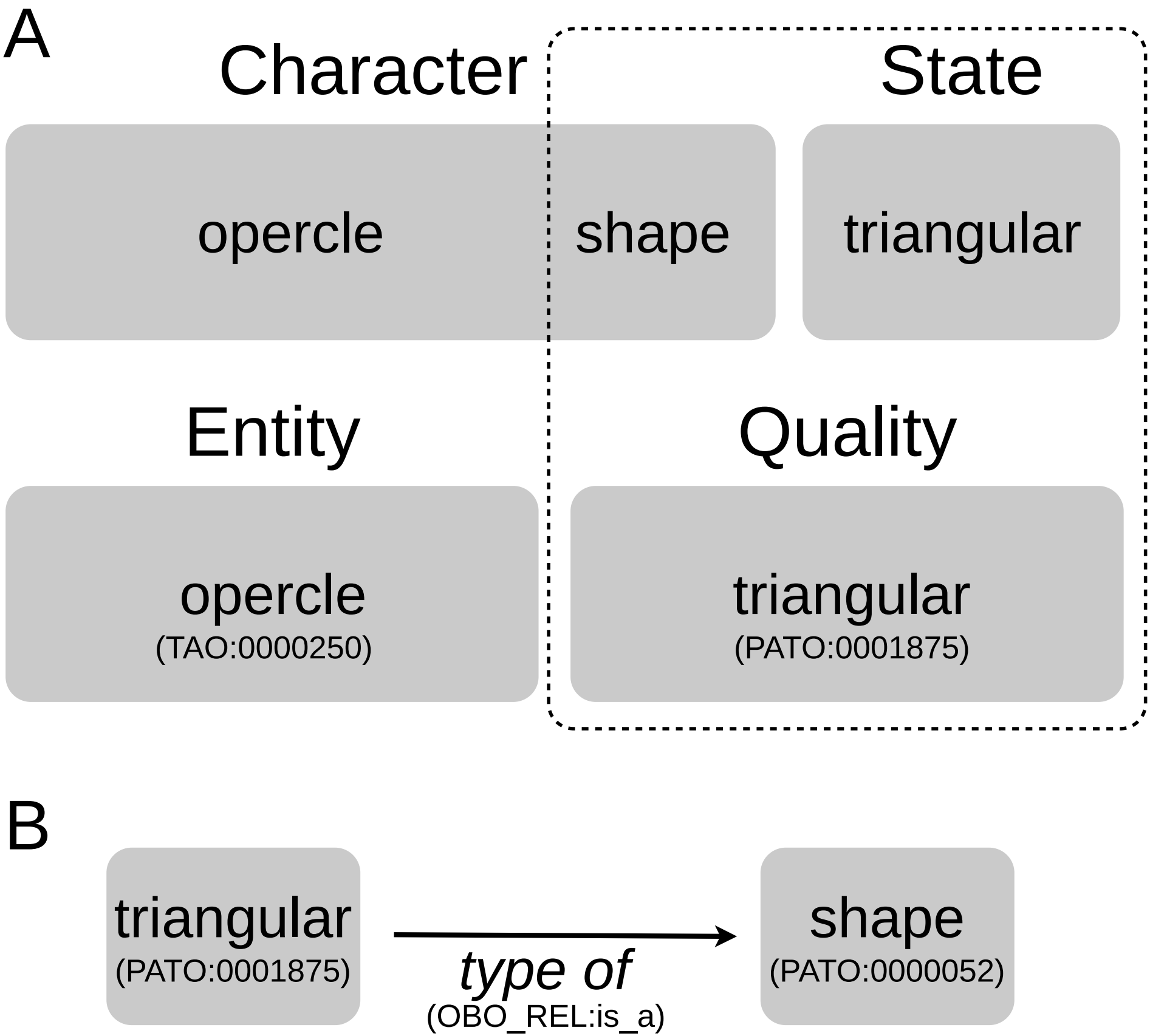
Identifiers in the http://vocab.phenoscape.org/ namespace are intended to be replaced with community standards as they become available.

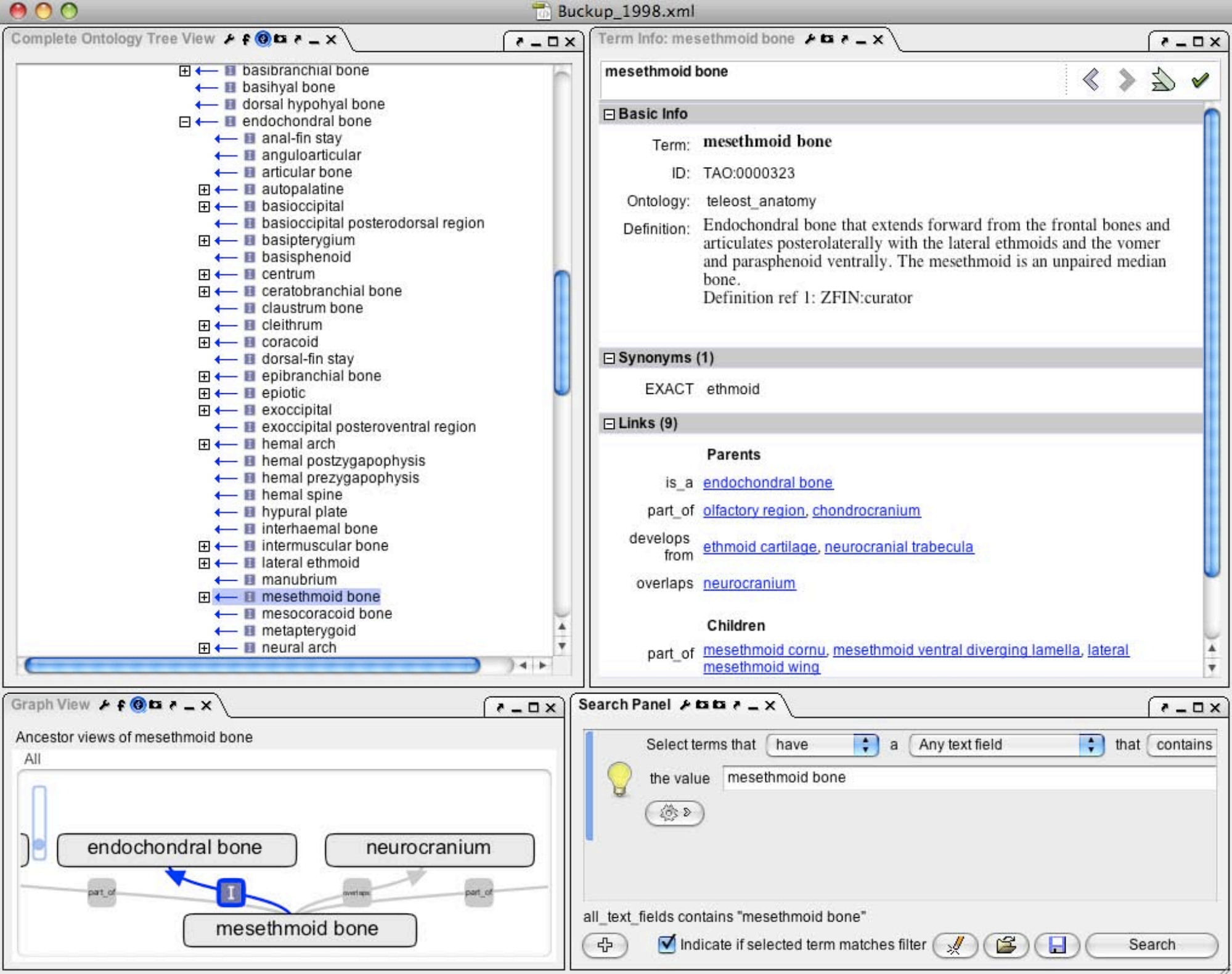
A

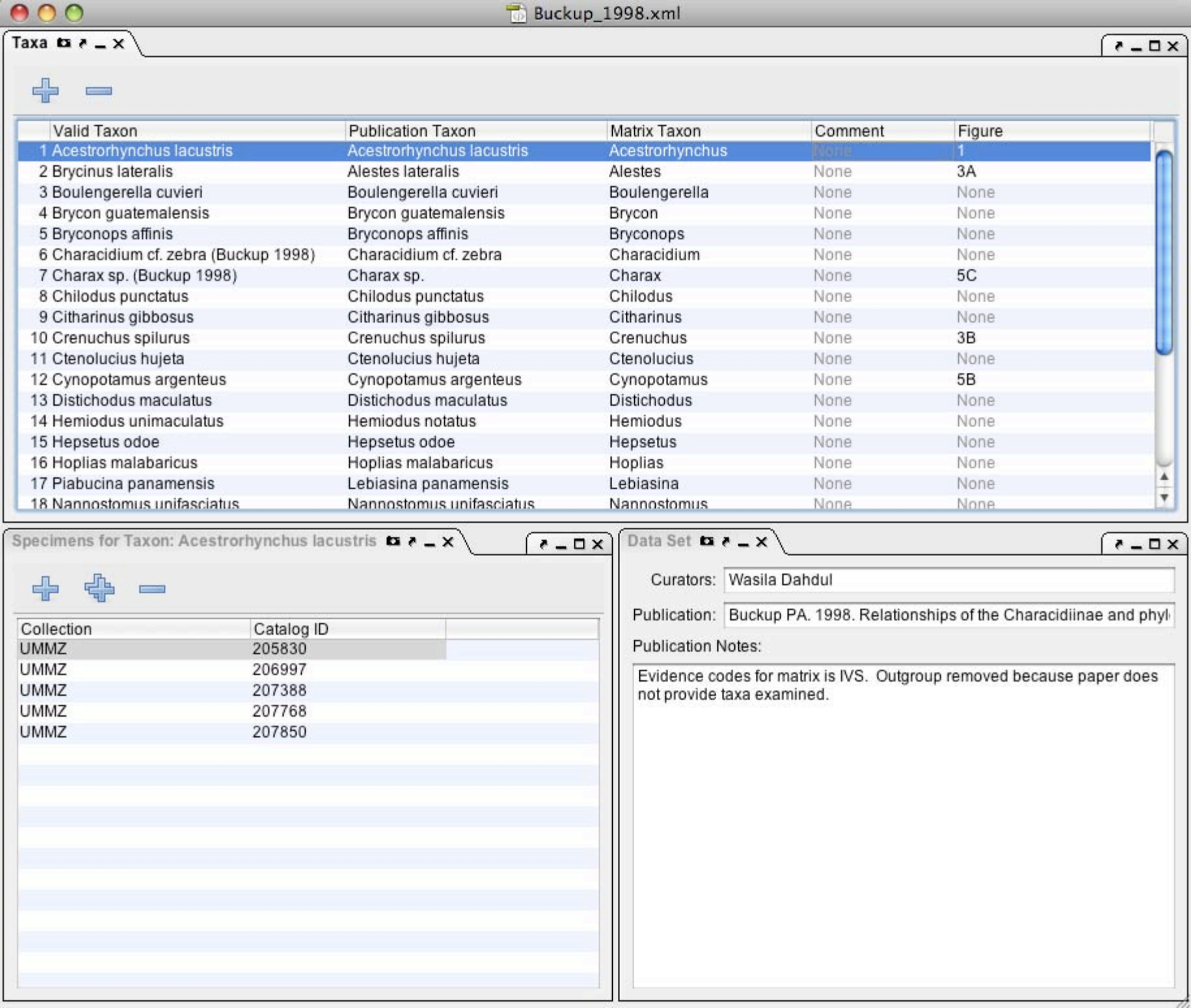
`<otu xmlns:dwc=”http://rs.tdwg.org/dwc/terms/” about="#00001" id="00001" label="Acestrorhynchus lacustris">`	*A NeXML “otu” element representing a study taxon, Acestrorhynchus lacustris.*
`<meta rel="dwc:taxonID" href="http://purl.obolibrary.org/obo/TTO_1030219" xsi:type="nex:ResourceMeta"/>`	*A metadata annotation linking this taxon to an identifier (TTO:1030219) from the Teleost Taxonomy Ontology.*
`<meta rel=”dwc:individualID” xsi:type="nex:ResourceMeta">`	*A metadata annotation linking this taxon to a specimen, UMMZ 205830.*
`<meta rel=”dwc:collectionID” href=”http://purl.obolibrary.org/obo/COLLECTION_0000403” xsi:type="nex:ResourceMeta"/>`	*A nested metadata annotation referencing the museum collection identifier (COLLECTION:0000403) for the specimen.*
`<meta property=”dwc:catalogNumber” xsi:type="nex:LiteralMeta">205830</meta>`	*A nested metadata annotation referencing the catalog number (205830) for the specimen.*
`</meta>` `</otu>`	*Closing XML tags.*

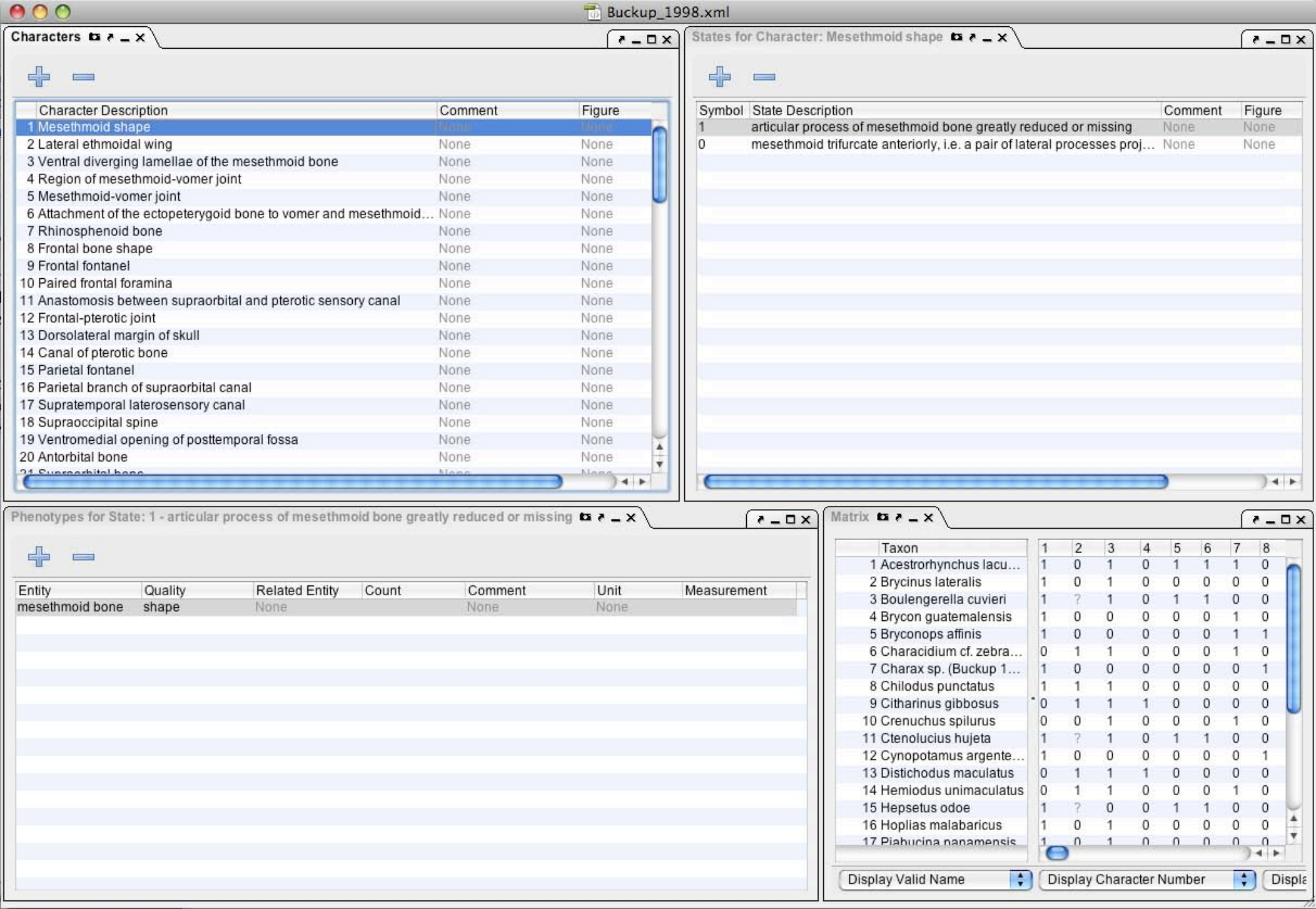
B

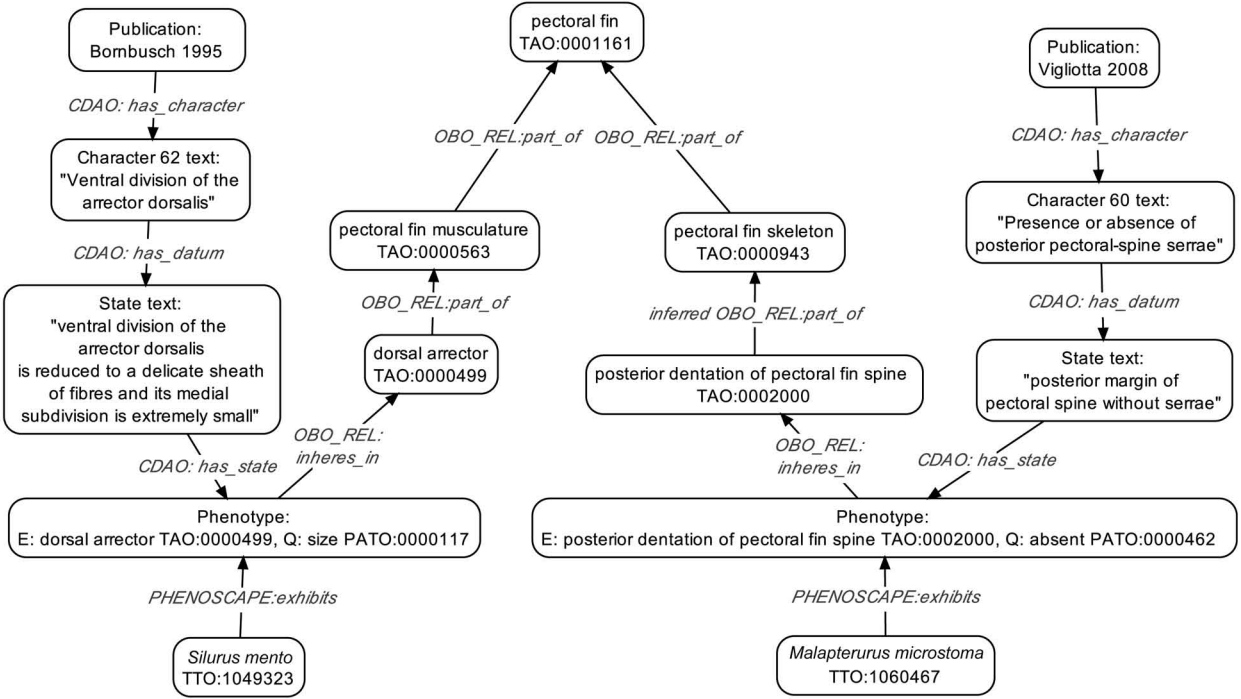
`<format xmlns:ps="http://vocab.phenoscape.org/" xmlns:phen="http://www.bioontologies.org/obd/schema/pheno">`	*A NeXML “format” element containing character and state definitions.*
`<states id="states01">` `<state id="state0102" about="#state0102" label="absent" symbol="1">`	*A “states” element with one state, “absent”.*
`<meta property="ps:describesPhenotype" xsi:type="nex:LiteralMeta">`	*A metadata annotation linking this state to an EQ phenotype description.*
`<phen:phenotype>` `<phen:phenotype_character>` `<phen:bearer>` `<phen:typeref about="TAO:0000127"/>` `</phen:bearer>` `<phen:quality>` `<phen:typeref about="PATO:0000462"/>` `</phen:quality>` `</phen:phenotype_character>` `</phen:phenotype>`	*An embedded block of PhenoXML data representing an EQ phenotype with entity “antorbital” (TAO:0000127) and quality “absent” (PATO:0000462).*
`</meta>` `</state>` `</states>`	*Closing XML tags.*
`<char id="char01" label="Presence or absence of antorbital" states="states01"/>` `</format>`	*The character definition related to the above state.*


A
Character
State
opercle
shape
triangular
Entity
Quality
opercle
(TAO:0000250)
triangular
(PATO:0001875)
B
triangular
(PATO:0001875)
type of
(OBO_REL:is_a)
shape
(PATO:0000052)


Buckup_1998.xml
Complete Ontology Tree View
basibranchial bone
basihyal bone
dorsal hypohyal bone
endochondral bone
anal-fin stay
anguloarticular
articular bone
autopalatine
basioccipital
basioccipital posterodorsal region
basipterygium
basisphenoid
centrum
ceratobranchial bone
claustrum bone
cleithrum
coracoid
dorsal-fin stay
epibranchial bone
epiotic
exoccipital
exoccipital posteroventral region
hemal arch
hemal postzygapophysis
hemal prezygapophysis
hemal spine
hypural plate
interhaemal bone
intermuscular bone
lateral ethmoid
manubrium
mesethmoid bone
mesocoracoid bone
metapterygoid
neural arch
Term Info: mesethmoid bone
mesethmoid bone
Basic Info
Term: mesethmoid bone
ID: TAO:0000323
Ontology: teleost_anatomy
Definition: Endochondral bone that extends forward from the frontal bones and articulates posterolaterally with the lateral ethmoids and the vomer and parasphenoid ventrally. The mesethmoid is an unpaired median bone.
Definition ref 1: ZFIN:curator
Synonyms (1)
EXACT ethmoid
Links (9)
Parents
is_a endochondral bone
part_of olfactory region, chondrocranium
develops from ethmoid cartilage, neurocranial trabecula
overlaps neurocranium
Children
part_of mesethmoid cornu, mesethmoid ventral diverging lamella, lateral mesethmoid wing
Graph View
Ancestor views of mesethmoid bone
All
endochondral bone
neurocranium
part_of
overlaps
part_of
mesethmoid bone
Search Panel
Select terms that have a Any text field that contains
the value mesethmoid bone
all_text_fields contains "mesethmoid bone"
Indicate if selected term matches filter
Search


Buckup_1998.xml
Taxa
Valid Taxon
Publication Taxon
Matrix Taxon
Comment
Figure
1 Acestrorhynchus lacustris
Acestrorhynchus lacustris
Acestrorhynchus
None
1
2 Brycinus lateralis
Alestes lateralis
Alestes
None
3A
3 Boulengerella cuvieri
Boulengerella cuvieri
Boulengerella
None
None
4 Brycon guatemalensis
Brycon guatemalensis
Brycon
None
None
5 Bryconops affinis
Bryconops affinis
Bryconops
None
None
6 Characidium cf. zebra (Buckup 1998)
Characidium cf. zebra
Characidium
None
None
7 Charax sp. (Buckup 1998)
Charax sp.
Charax
None
5C
8 Chilodus punctatus
Chilodus punctatus
Chilodus
None
None
9 Citharinus gibbosus
Citharinus gibbosus
Citharinus
None
None
10 Crenuchus spilurus
Crenuchus spilurus
Crenuchus
None
3B
11 Ctenolucius hujeta
Ctenolucius hujeta
Ctenolucius
None
None
12 Cynopotamus argenteus
Cynopotamus argenteus
Cynopotamus
None
5B
13 Distichodus maculatus
Distichodus maculatus
Distichodus
None
None
14 Hemiodus unimaculatus
Hemiodus notatus
Hemiodus
None
None
15 Hepsetus odoe
Hepsetus odoe
Hepsetus
None
None
16 Hoplias malabaricus
Hoplias malabaricus
Hoplias
None
None
17 Piabucina panamensis
Lebiasina panamensis
Lebiasina
None
None
18 Nannostomus unifasciatus
Nannostomus unifasciatus
Nannostomus
None
None
Specimens for Taxon: Acestrorhynchus lacustris
Collection
Catalog ID
UMMZ
205830
UMMZ
206997
UMMZ
207388
UMMZ
207768
UMMZ
207850
Data Set
Curators:
Wasila Dahdul
Publication:
Buckup PA. 1998. Relationships of the Characidiinae and phyl
Publication Notes:
Evidence codes for matrix is IVS. Outgroup removed because paper does not provide taxa examined.


Buckup_1998.xml
Characters
Character Description
Comment
Figure
1 Mesethmoid shape
2 Lateral ethmoidal wing
3 Ventral diverging lamellae of the mesethmoid bone
4 Region of mesethmoid-vomer joint
5 Mesethmoid-vomer joint
6 Attachment of the ectopeterygoid bone to vomer and mesethmoid…
7 Rhinosphenoid bone
8 Frontal bone shape
9 Frontal fontanel
10 Paired frontal foramina
11 Anastomosis between supraorbital and pterotic sensory canal
12 Frontal-pterotic joint
13 Dorsolateral margin of skull
14 Canal of pterotic bone
15 Parietal fontanel
16 Parietal branch of supraorbital canal
17 Supratemporal laterosensory canal
18 Supraoccipital spine
19 Ventromedial opening of posttemporal fossa
20 Antorbital bone
States for Character: Mesethmoid shape
Symbol
State Description
Comment
Figure
1
articular process of mesethmoid bone greatly reduced or missing
0
mesethmoid trifurcate anteriorly, i.e. a pair of lateral processes proj…
Phenotypes for State: 1 - articular process of mesethmoid bone greatly reduced or missing
Entity
Quality
Related Entity
Count
Comment
Unit
Measurement
mesethmoid bone
shape
Matrix
Taxon
1 Acestrorhynchus lacu…
2 Brycinus lateralis
3 Boulengerella cuvieri
4 Brycon guatemalensis
5 Bryconops affinis
6 Characidium cf. zebra…
7 Charax sp. (Buckup 1…
8 Chilodus punctatus
9 Citharinus gibbosus
10 Crenuchus spilurus
11 Ctenolucius hujeta
12 Cynopotamus argente…
13 Distichodus maculatus
14 Hemiodus unimaculatus
15 Hepsetus odoe
16 Hoplias malabaricus
Display Valid Name
Display Character Number


Publication:
Bombusch 1995
CDAO: has_character
Character 62 text:
"Ventral division of the
arrector dorsalis"
CDAO: has_datum
State text:
"ventral division of the
arrector dorsalis
is reduced to a delicate sheath
of fibres and its medial
subdivision is extremely small"
CDAO: has_state
Phenotype:
E: dorsal arrector TAO:0000499, Q: size PATO:0000117
PHENOSCAPE:exhibits
Silurus mento
TTO:1049323
OBO_REL:
inheres_in
dorsal arrector
TAO:0000499
OBO_REL:part_of
pectoral fin musculature
TAO:0000563
OBO_REL:part_of
pectoral fin
TAO:0001161
OBO_REL:part_of
pectoral fin skeleton
TAO:0000943
inferred OBO_REL:part_of
posterior dentation of pectoral fin spine
TAO:0002000
OBO_REL:
inheres_in
Phenotype:
E: posterior dentation of pectoral fin spine TAO:0002000, Q: absent PATO:0000462
PHENOSCAPE:exhibits
Malapterurus microstoma
TTO:1060467
CDAO: has_state
State text:
"posterior margin of
pectoral spine without serrae"
CDAO: has_datum
Character 60 text:
"Presence or absence of
posterior pectoral-spine serrae"
CDAO: has_character
Publication:
Vigliotta 2008