Phenoscape: Ontologies for Large Multi-species Phenotype Datasets

Peter Midford[1], Paula Mabee[2], Todd Vision[3,4], Hilmar Lapp[3],Jim Balhoff[3,4], Wasila Dahdul[2,3], Cartik Kothari[3,4], John Lundberg[5], Monte Westerfield[6]

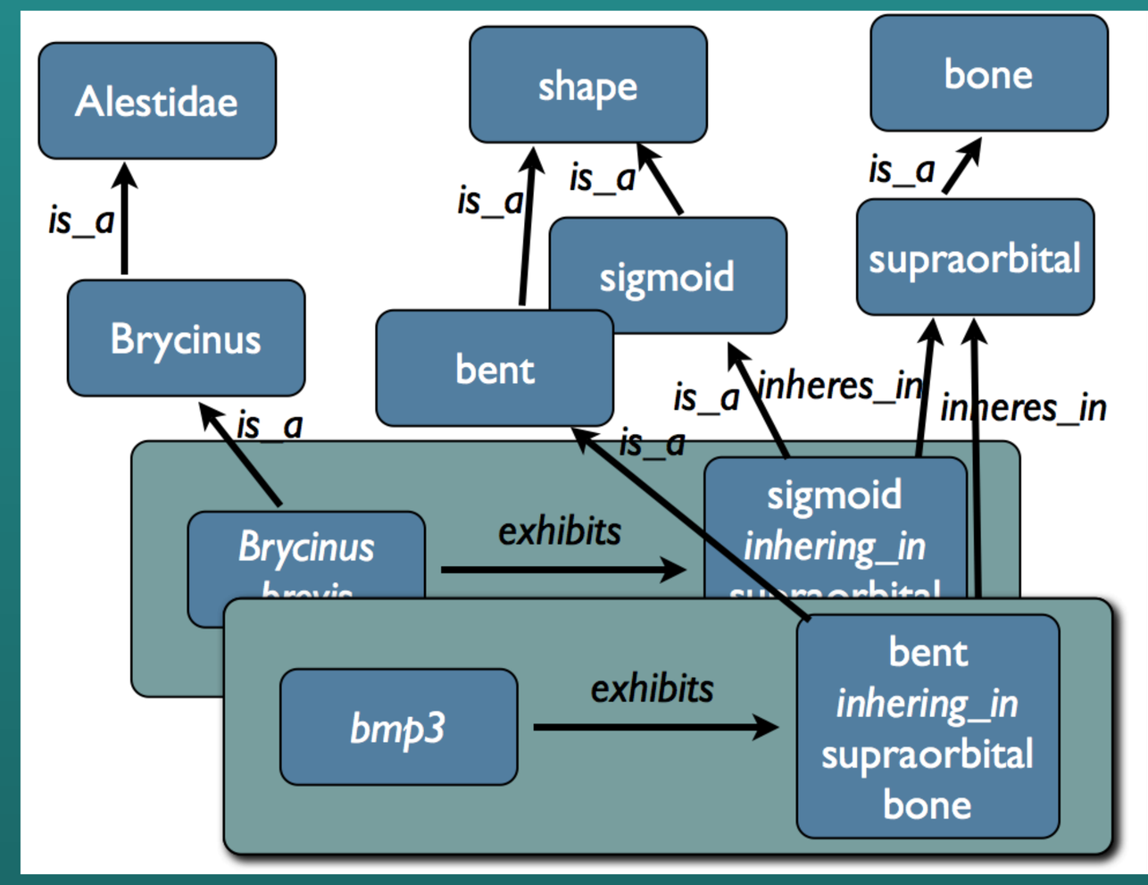
[1]University of Kansas; [2]The University of South Dakota; [3]National Evolutionary Synthesis Center (NESCent); [4]University of North Carolina at Chapel Hill; [5]Academy of Natural Sciences, Dept. Ichthyology; [6]University of Oregon, Zebrafish Information Network (ZFIN)

Introduction

The Phenoscape project has developed ontologies and tools to integrate morphological and genomic data. This integration allows us to address comparative questions in evolutionary biology. We have curated 81 publications, describing nearly 6,000 phenotypic characters in 4,000 species of Ostariophysian fish. Our database of ontology-based annotations is now available via a web-based interface at http://kb.phenoscape.org.

Until recently, biological ontologies have either focused on single model organisms, or, like the Gene Ontology, attempted to span the tree of life. Phenoscape is a project to develop ontologies and a database to describe the phenotypes of the Osteriophysi, a group of >9,000 species of teleost fish. Among the Osteriophysi is the Zebrafish, an established model organism, frequently used in developmental biology. Because Phenoscape shares anatomical terms with the anatomy ontology maintained by the Zebrafish Information Network (zfin.org), we will be able to apply reasoning to queries over zebrafish mutant phenotypes and "evolutionary" phenotypes across the Ostariophysi.

Examples of such queries include finding candidate geners underlying the evolution of morphological characters and searches to discover similar phenotypes among different taxa.



Above: Graph representing similarity of phenotypes between the species *Brycinus brevis* and the Bmp3 mutant.

Ontologies

Teleost Anatomy Ontology

The Teleost Anatomy Ontology (TAO) is a multi-species anatomy ontology for teleost fishes that we initiated in September 2007. It started as a clone of the zebrafish anatomy ontology with which we still synchronize regularly. Development of the TAO is currently centered on the skeletal system, since it is often the focus of evolutionary studies in ichthyology as well as genetic studies in zebrafish. Since inception the number of skeletal terms has grown from 253 to 618.

Phenotype Ontology (PaTO)

The Phenotype and Trait Ontology (PaTO) is an ontology of attributes and qualities that is used by multiple OBO projects to annotate phenotypes.

Other OBO Ontologies

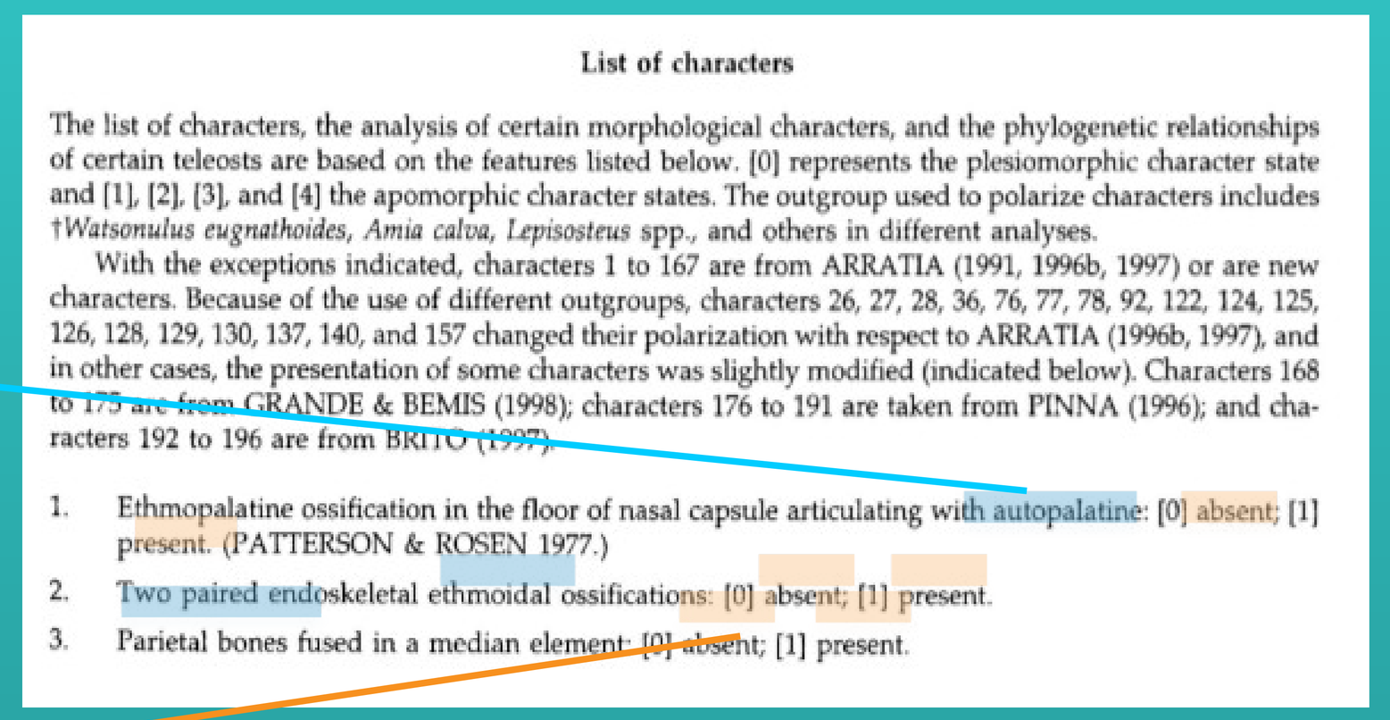
Phenoscape uses several other OBO foundry ontologies, including the relations ontology (OBO_REL), the spatial ontology (BSPO), the evidence codes ontology (ECO), to which we added codes for inferred homology, and the biological process portion of the Gene Ontology.

Teleost Taxonomy Ontology (TTO)

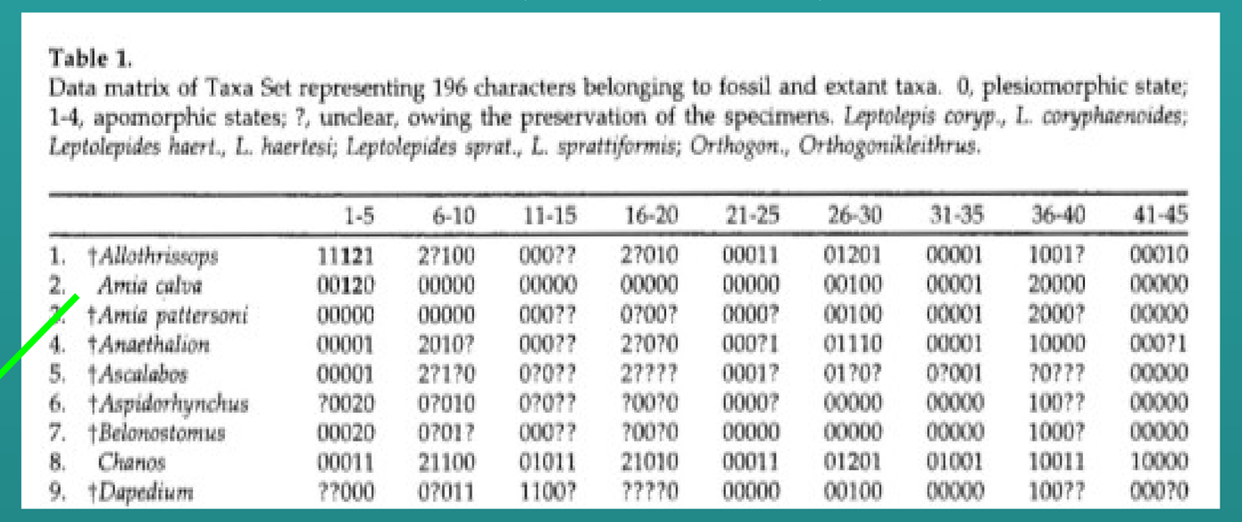
The Teleost Taxonomy Ontology (TTO) is a tree of taxonomic terms, connected by the relation *is_a*, and a set of rank terms (e.g., order, family, etc.), which are associated with taxonomic terms with a special *'has_rank'* relation. This ontology is generated from tables exported from the Catalog of Fishes (CoF), an expert database. We have also added, via manual curation, taxonomic updates from several area experts in Siluriformes, Characiformes, and Cypriniformes as well as several fossil taxa. This ontology currently contains 30,804 species and 36,785 total terms.

Curation of evolutionary literature

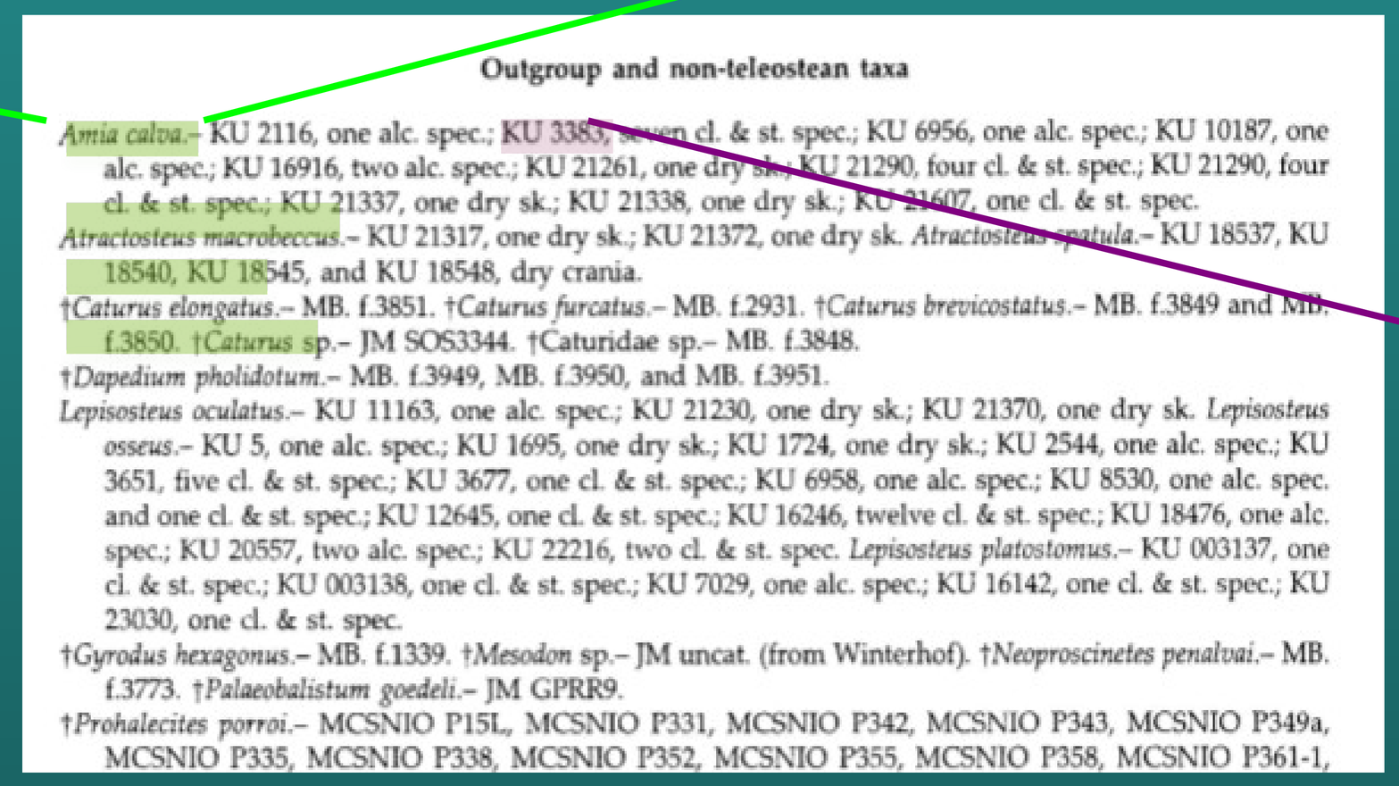
Evolutionary phenotypes (Characters and states)



Phenotype to taxon mapping (matrix)

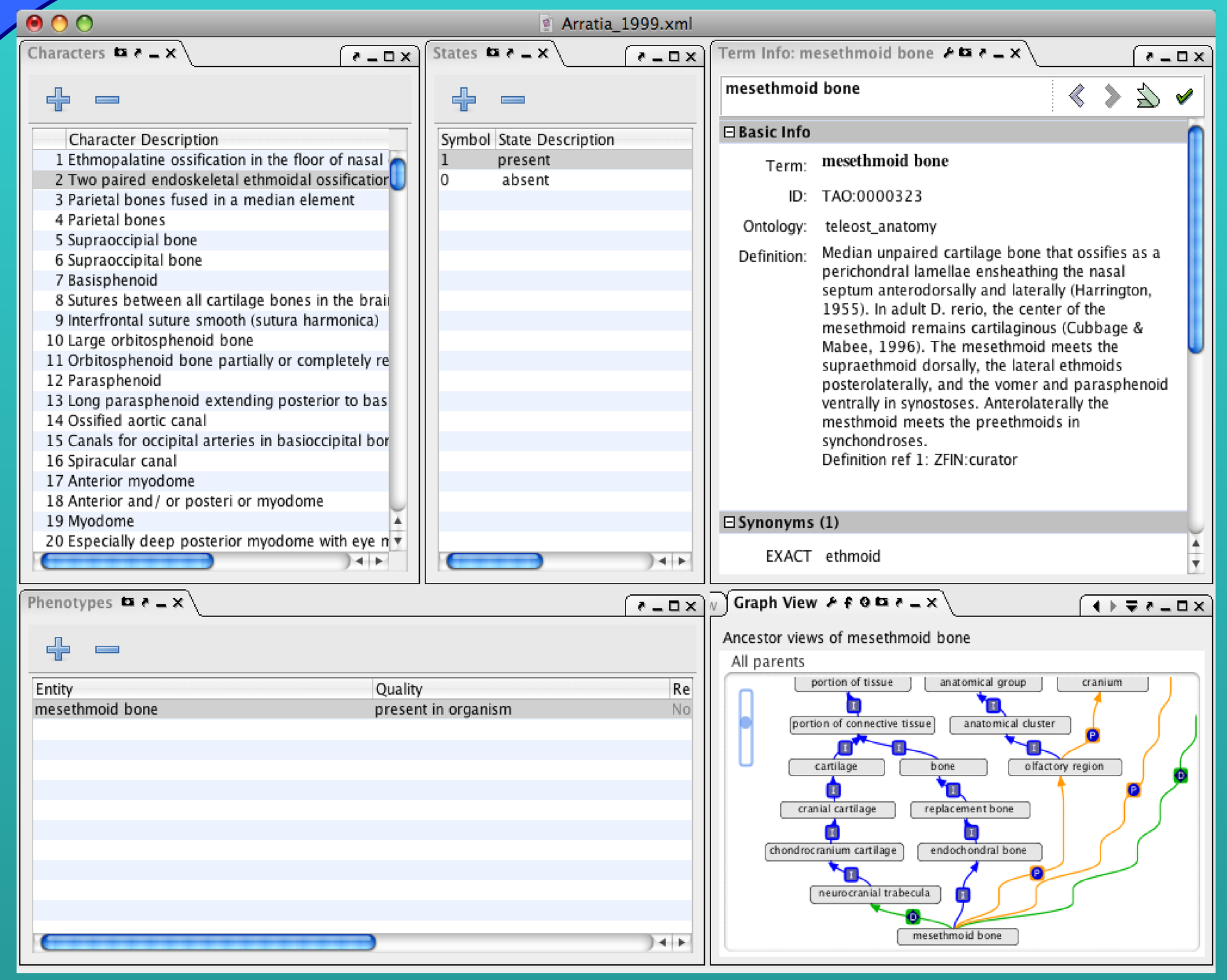


Taxa (extant and extinct)

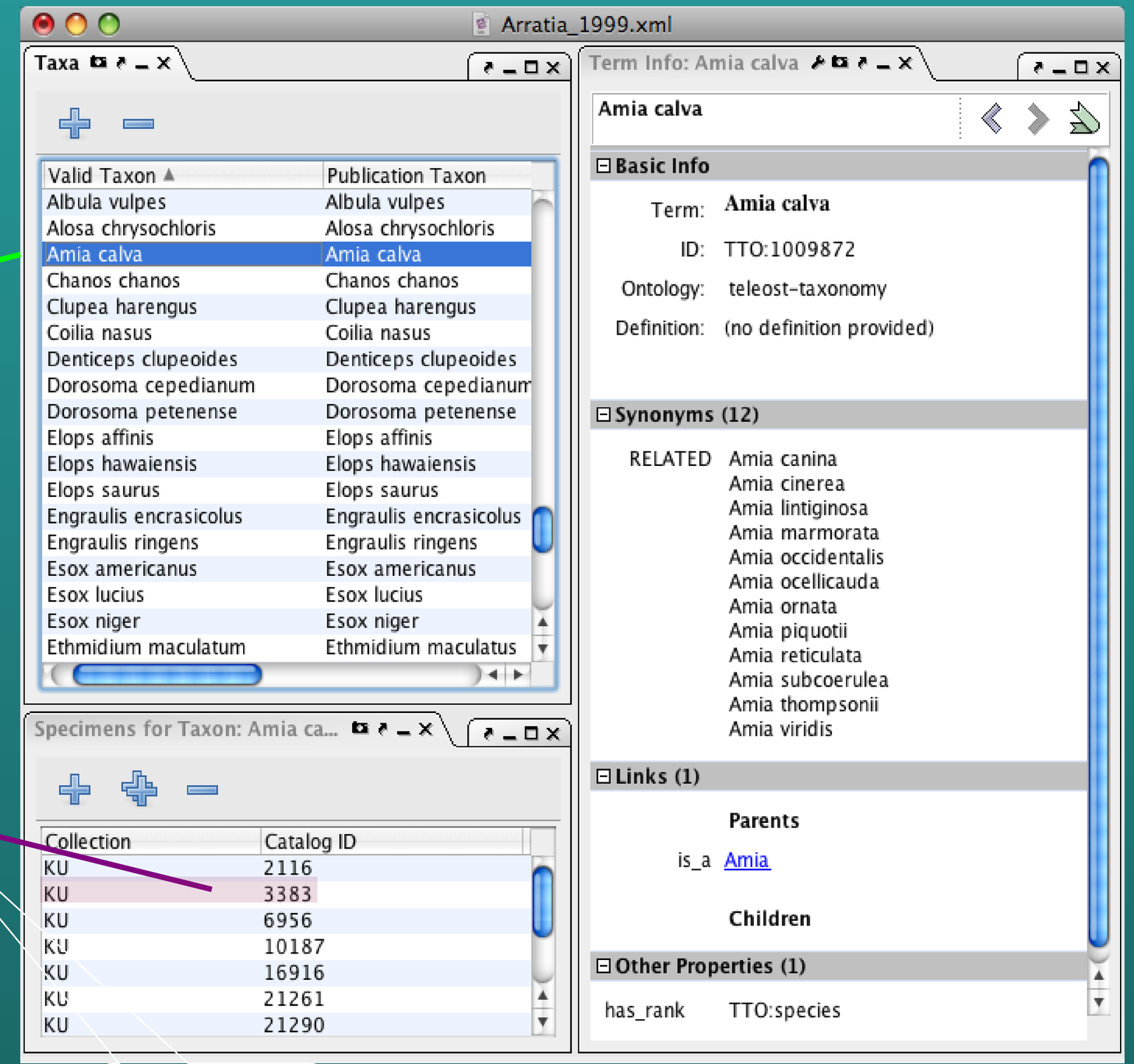


Arratia, 1999
Zoological Journal of the Linnean Society 151:691-757.

Phenex: a tool for evolutionary data curation



Above: Entering EQ phenotypes to describe a character state in Phenex



Links to physical 'voucher' in a collection

Above: Editing taxa in Phenex by selecting from a taxonomy ontology

The unique challenges faced in annotating these evolutionary data have required the development of new resources, tools and extensions to existing ones. Phenex is a software application developed by Phenoscape and hosted by the OBO Project for the curation of evolutionary data using ontologies. Data annotated with Phenex is based on the Entity/Quality (EQ) model for representing phenotypes. Phenex provides facilities for loading ontologies and easily selecting terms.

Character/character state-oriented Phenex interface

- Browse the list of characters, character states and taxa directly from a character matrix file
- Compose EQ phenotype descriptions - ontological descriptions of character states - using anatomy and phenotype ontologies
- Choose taxa directly from a taxonomy ontology

Embedding EQ into traditional character data via NeXML

- NeXML is a new XML-based standard for the exchange of character matrices and trees. It is intended to supercede the commonly used NEXUS format. Besides following XML standards, it allows annotations which can be extracted to emerging web standard formats such as RDFa. These annotations can include custom XML extensions such as our ontology-based EQ phenotypes
- EQ data are stored using the PhenoXML format, embedded within NeXML
- Taxon TTO identifiers are stored as custom annotations to taxa in the taxa block
- Mesquite is a popular tool for editing and performing comparative analysis of data from character matrices and trees. We are testing a plug-in that allows Mesquite to view our ontology annotations to characters in Nexml files. In the future, this plugin may support round trip editing in Mesquite as well as Phenex.

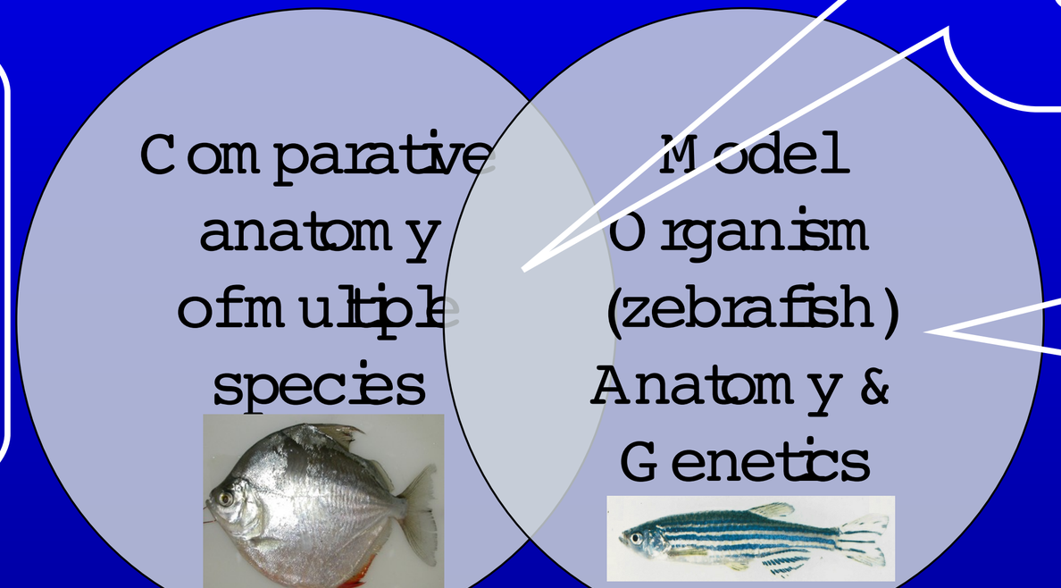
Phenoscape knowledge-base:

This week we released our database and a web interface that supports queries to our database and enables cross-domain queries to the zebrafish (ZFIN) database

Taxon	# Papers	# Species	# Characters
Cypriniforms	8	676	794
Siluriforms	20	1724	2,110
Characiforms	10	754	1,156
Gymnotiforms	1	116	231
Gonorynchiforms	3	41	467
Clupeiforms	5	200	439
Euteleosts	3	145	582
TOTAL	**51**	**3656**	**5,779**

Over 400,000 phenotypes (character-taxon pairs) from the comparative morphology literature

Candidate genes for evo-devo



Over 19,000 mutant anatomical phenotypes (zfin.org)

Sharing standards across domains is critical in order to meet the needs of cross-domain queries such as those studies relating evolution and development ("evo-devo"). Such queries might e.g. identify candidate genes for evolutionary phenotypes, candidate taxa for mutant phenotypes, and identify correlations across data types

Phenoscape database and web services

The knowledge-base is stored in a relational database using the OBD schema – which combines the ontologies and annotations into a unified set of statements. Semantics in the ontologies allow the OBD reasoner to generate additional implied annotations.

The web user interface uses publicly available web services to access the knowledge-base.

Web services

OBD Java API

OBD database

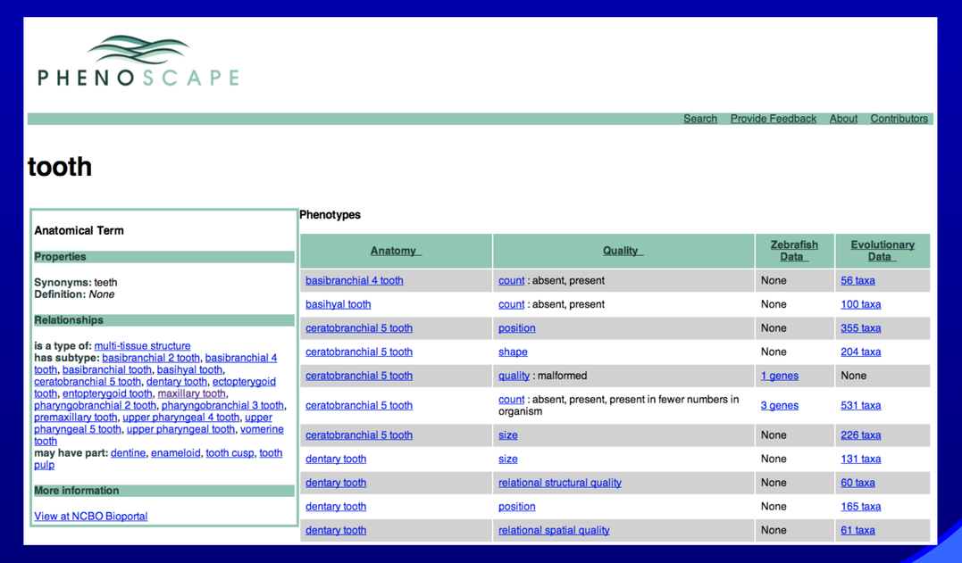
Phenoscape web user interface http://kb.phenoscape.org

The web user interface is now available!

Above: tag cloud of ontology terms from search page

Here you can search and browse annotations in the knowledge-base, via anatomical part, taxon, gene, or publication.

Corresponding evolutionary and model organism phenotypes are linked due to use of related anatomical and quality terms.



For more information:

Homepage: http://www.phenoscape.org
Blog: http://blog.phenoscape.org
Email list: phenoscape-discuss@lists.sourceforge.net

We thank NSF DBI0641025 and the National Evolutionary Synthesis Center (NESCent) NSF #EF-0423641 for funding and the many contributors to data, character curation, ontology development and ideas (see: http://kb.phenoscape.org/contributors/)