Phenoscape: Linking phenotypic biodiversity to model organism genetics using ontologies

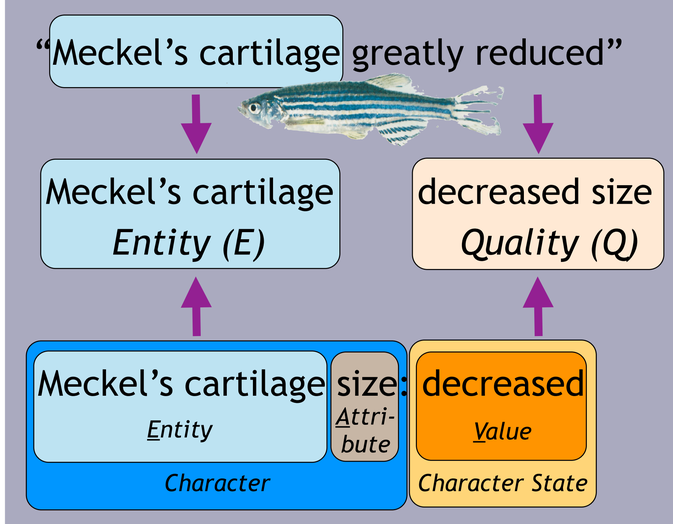
Todd Vision[1,2], Jim Balhoff[1], Cartik Kothari[1], Hilmar Lapp[1], Wasila Dahdul[3], Paula Mabee[3], John Lundberg[4], Peter Midford[5], Monte Westerfield[6]

(1) US National Evolutionary Synthesis Center, Durham, North Carolina, USA, (2) University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, (3) University of South Dakota, Vermillion, South Dakota, USA, (4) Academy of Natural Sciences, Philadelphia, Pennsylvania, USA, (5) University of Kansas, Lawrence, Kansas, USA, (6) Zebrafish Information Network and University of Oregon, Eugene, Oregon, USA

Abstract

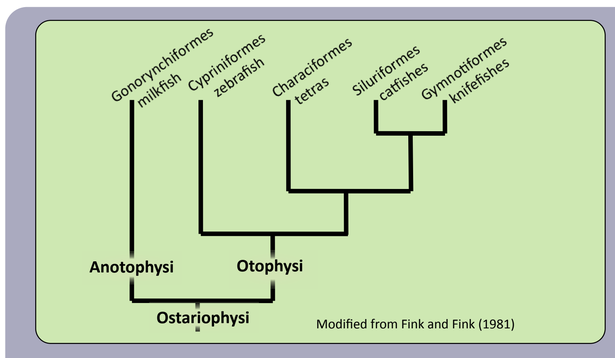
Model organism databases such as the Zebrafish Information Network (ZFIN) use terms from anatomy and quality ontologies to describe the thousands of phenotypes known from genetic studies. By contrast, the vast stores of phenotypic biodiversity data are typically rendered in free-form text in journal articles. Thus, there is no informatic tool for researchers to ask questions that cut across these two domains of knowledge, such as "what sister clades in nature differ phenotypically in the same way a given mutant genotype differs from the wildtype?" Developing such a tool is the aim of the Phenoscape project. We illustrate here the components necessary to reach this goal and how they interact to enable new questions to be jointly asked of phenotypic biodiversity and genomic data.

Comparing phenotypes among genetic mutants of model organisms: the EQ formalism



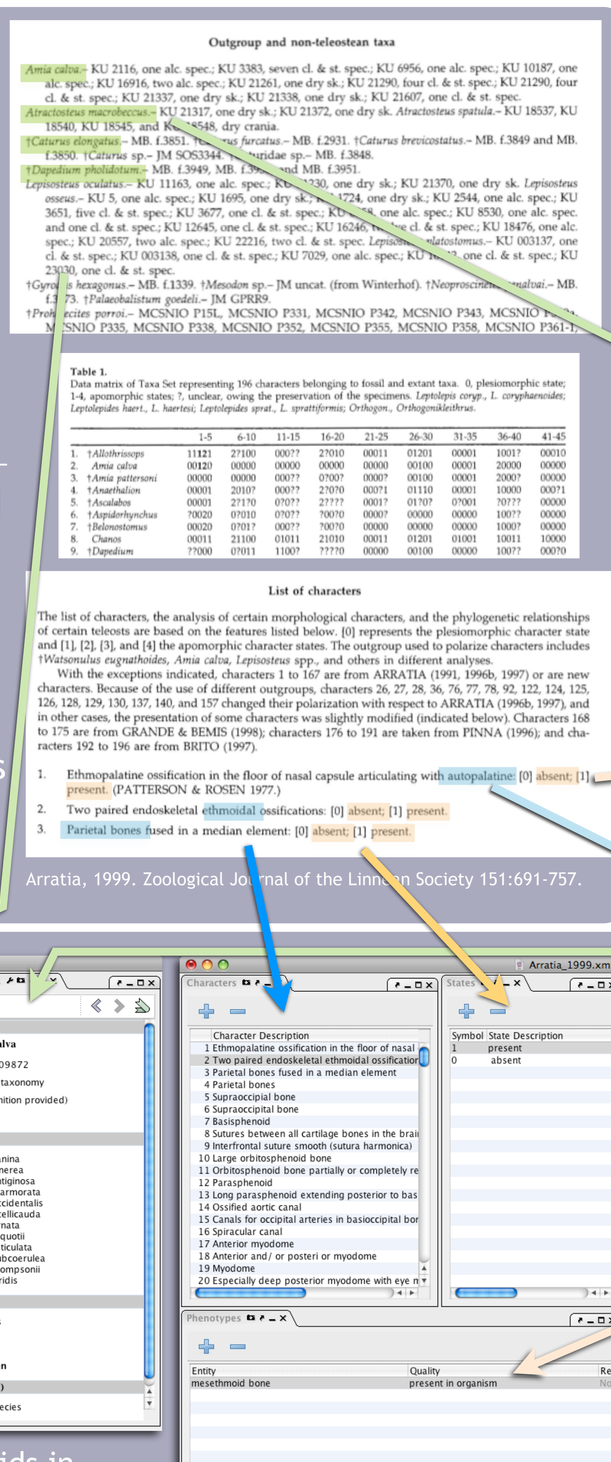
Model organism databases have achieved interoperability between phenotype annotations by expressing these formally in Entity-Quality (EQ) syntax, and using ontology terms for entity as well as quality terms. An EQ statement asserts that the quality Q *inheres in* the bearer entity E.

The same formalism can be applied to the descriptions of phenotype biodiversity in the systematics and taxonomic literature, thereby making it possible to link evolutionary phenotypes to their genetic underpinnings.



Our initial application domain is the evolutionary phenotype diversity for Ostariophysi, a clade of teleost freshwater fishes containing zebrafish and approximately 9,000 other species (~28% of all known fish species).

Such phenotypes are traditionally reported in free text form as states of comparative characters in data matrices published in the systematics literature, typically as a result of analyzing natural history collection specimens.



Outgroup and non-teleostean taxa

Amia calva.– KU 2116, one alc. spec.; KU 3383, seven cl. & st. spec.; KU 6956, one alc. spec.; KU 10187, one alc. spec.; KU 16916, two alc. spec.; KU 21261, one dry sk.; KU 21290, four cl. & st. spec.; KU 21290, four cl. & st. spec.; KU 21337, one dry sk.; KU 21338, one dry sk.; KU 21607, one cl. & st. spec.

Atractosteus macrobeccus.– KU 21317, one dry sk.; KU 21372, one dry sk. *Atractosteus spatula*.– KU 18537, KU 18540, KU 18545, and KU 18548, dry crania.

†*Caturus elongatus*.– MB. f.3851. †*Caturus furcatus*.– MB. f.2931. †*Caturus brevicostatus*.– MB. f.3849 and MB. f.3850. †*Caturus* sp.– JM SOS3344. †Caturidae sp.– MB. f.3848.

†*Dapedium pholidotum*.– MB. f.3949, MB. f.3[illegible] and MB. f.3951.

Lepisosteus oculatus.– KU 11163, one alc. spec.; KU [illegible]230, one dry sk.; KU 21370, one dry sk. *Lepisosteus osseus*.– KU 5, one alc. spec.; KU 1695, one dry sk.; KU [illegible]724, one dry sk.; KU 2544, one alc. spec.; KU 3651, five cl. & st. spec.; KU 3677, one cl. & st. spec.; KU [illegible]58, one alc. spec.; KU 8530, one alc. spec. and one cl. & st. spec.; KU 12645, one cl. & st. spec.; KU 16246, [illegible] cl. & st. spec.; KU 18476, one alc. spec.; KU 20557, two alc. spec.; KU 22216, two cl. & st. spec. *Lepisosteus platostomus*.– KU 003137, one cl. & st. spec.; KU 003138, one cl. & st. spec.; KU 7029, one alc. spec.; KU [illegible]2, one cl. & st. spec.; KU 23030, one cl. & st. spec.

†*Gyrosteus hexagonus*.– MB. f.1339. †*Mesodon* sp.– JM uncat. (from Winterhof). †*Neoproscinetes penalvai*.– MB. f.3[illegible]73. †*Palaeobalistum goedeli*.– JM GPRR9.

†*Prohalecites porroi*.– MCSNIO P15L, MCSNIO P331, MCSNIO P342, MCSNIO P343, MCSNIO P[illegible] MCSNIO P335, MCSNIO P338, MCSNIO P352, MCSNIO P355, MCSNIO P358, MCSNIO P361-1[illegible]

Table 1.
Data matrix of Taxa Set representing 196 characters belonging to fossil and extant taxa. 0, plesiomorphic state; 1-4, apomorphic states; ?, unclear, owing the preservation of the specimens. *Leptolepis coryp., L. coryphaenoides; Leptolepides haert., L. haertesi; Leptolepides sprat., L. sprattiformis; Orthogon., Orthogonikleithrus.*

		1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45
1.	†*Allothrissops*	11121	2?100	000??	2?010	00011	01201	00001	1001?	00010
2.	*Amia calva*	00120	00000	00000	00000	00000	00100	00001	20000	00000
3.	†*Amia pattersoni*	00000	00000	000??	0?00?	0000?	00100	00001	2000?	00000
4.	†*Anaethalion*	00001	20107	000??	2?070	00071	01110	00001	10000	00071
5.	†*Ascalabos*	00001	27170	0?0??	2????	0001?	01?0?	07001	?0???	00000
6.	†*Aspidorhynchus*	?0020	07010	0?0??	?0070	0000?	00000	00000	100??	00000
7.	†*Belonostomus*	00020	07017	000??	?0070	00000	00000	00000	1000?	00000
8.	*Chanos*	00011	21100	01011	21010	00011	01201	01001	10011	10000
9.	†*Dapedium*	??000	07011	1100?	????0	00000	00100	00000	100??	00070

List of characters

The list of characters, the analysis of certain morphological characters, and the phylogenetic relationships of certain teleosts are based on the features listed below. [0] represents the plesiomorphic character state and [1], [2], [3], and [4] the apomorphic character states. The outgroup used to polarize characters includes †*Watsonulus eugnathoides, Amia calva, Lepisosteus* spp., and others in different analyses.

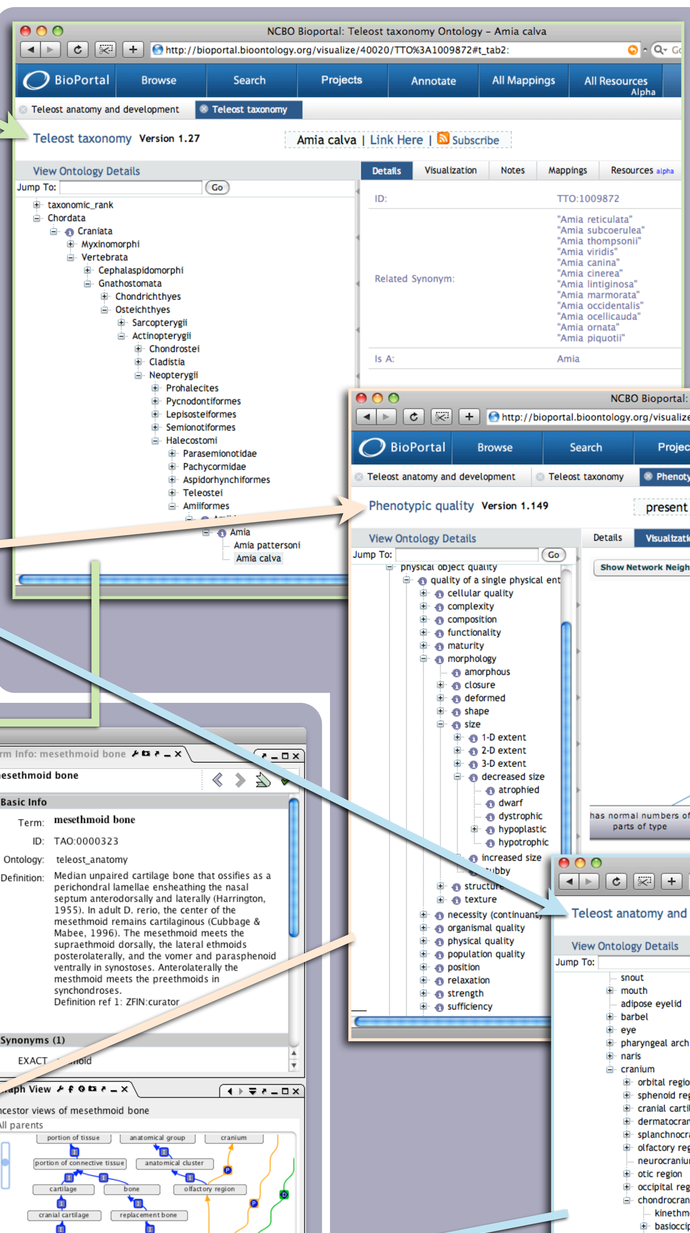
With the exceptions indicated, characters 1 to 167 are from ARRATIA (1991, 1996b, 1997) or are new characters. Because of the use of different outgroups, characters 26, 27, 28, 36, 76, 77, 78, 92, 122, 124, 125, 126, 128, 129, 130, 137, 140, and 157 changed their polarization with respect to ARRATIA (1996b, 1997), and in other cases, the presentation of some characters was slightly modified (indicated below). Characters 168 to 175 are from GRANDE & BEMIS (1998); characters 176 to 191 are taken from PINNA (1996); and characters 192 to 196 are from BRITO (1997).

1. Ethmopalatine ossification in the floor of nasal capsule articulating with autopalatine: [0] absent; [1] present. (PATTERSON & ROSEN 1977.)
2. Two paired endoskeletal ethmoidal ossifications: [0] absent; [1] present.
3. Parietal bones fused in a median element: [0] absent; [1] present.

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

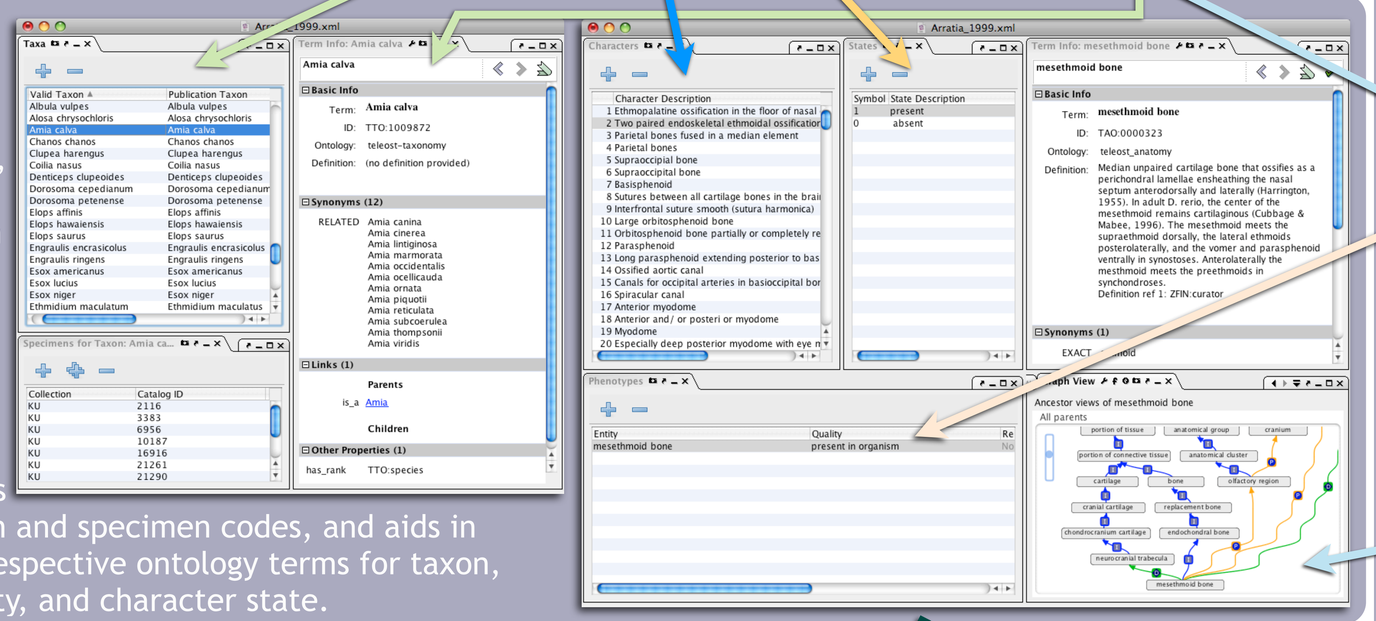
Ontologies

Logically sound ontologies that correctly reflect the state of knowledge are key to this approach. We developed the Teleost Taxonomy Ontology (TTO) based on the Catalog of Fishes, and the multi-species Teleost Anatomy Ontology (TAO) as a clone of the Zebrafish Anatomy Ontology (ZFA).



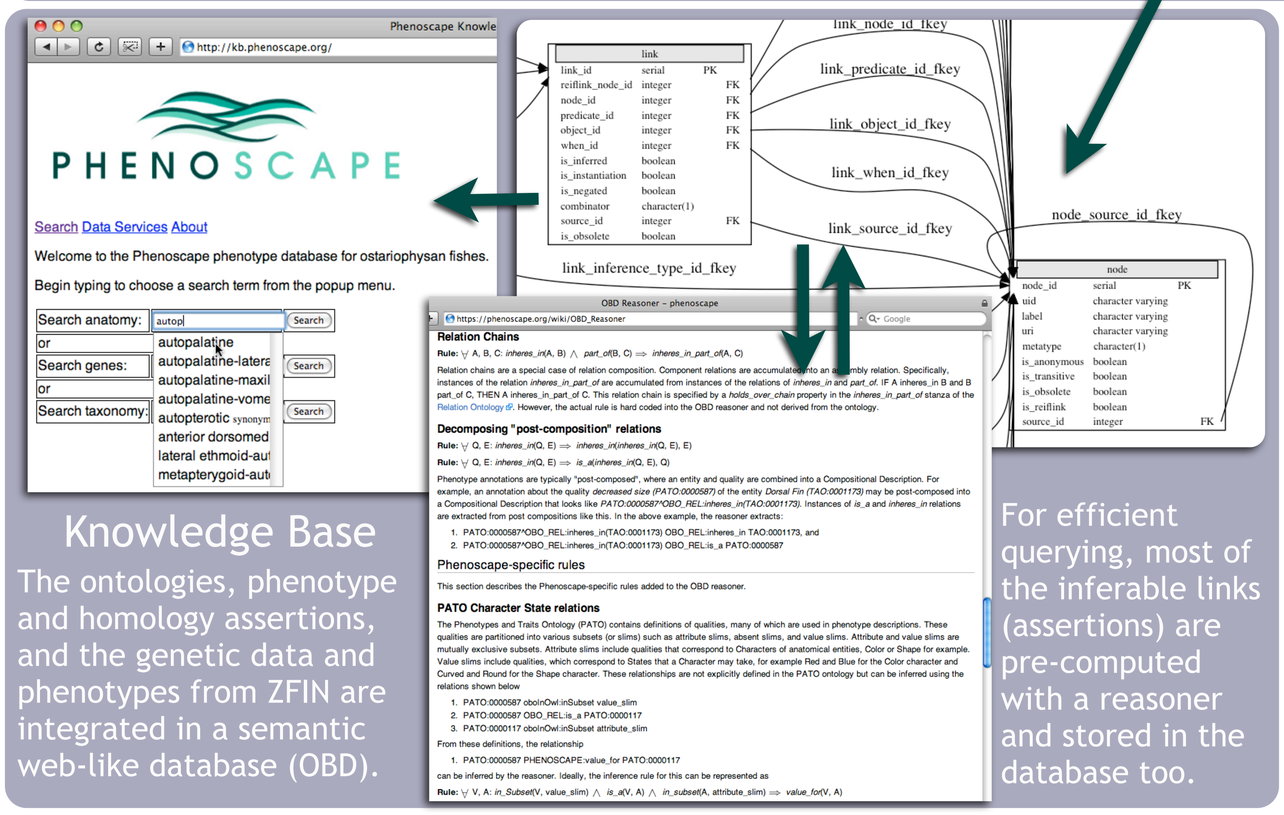
Data Curation

We developed a tool, Phenex, to accelerate transformation of character and state descriptions into formal phenotype assertions. Phenex records the publication and specimen codes, and aids in choosing the respective ontology terms for taxon, character entity, and character state.



Knowledge Base

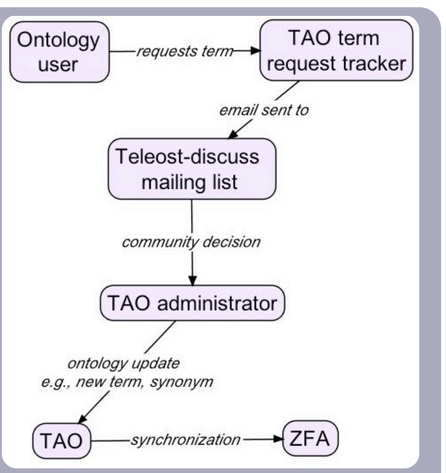
The ontologies, phenotype and homology assertions, and the genetic data and phenotypes from ZFIN are integrated in a semantic web-like database (OBD).



For efficient querying, most of the inferable links (assertions) are pre-computed with a reasoner and stored in the database too.

Community involvement

The ontologies involved in this effort have to evolve constantly in response to curation. A multi-step process ensures that new terms can be proposed by anyone, and are reviewed by the community of domain experts.



Homology

Terms in the multi-species TAO are structurally defined, and hence do not imply homology, not even with themselves when applied to different species. Rather, homology is a hypothesis requiring attribution and evidence, the latter also coming from an ontology. This remains an ongoing challenge.

Publication	Entity 1	Taxon 1	Entity 2	Taxon 2	Evidence
Britz and Hoffman 2006	supradorsal	Teleostei	claustrum bone	Otophysi	inferred from developmental similarity
Britz and Hoffman 2006	supradorsal	Teleostei	claustrum bone	Otophysi	inferred from positional similarity

Acknowledgments. We thank NSF DBI 0641025, NIH HG0002659, and the National Evolutionary Synthesis Center (NESCent), NSF #EF-0423641, for funding. We are also indebted to our curators, and the semantic data integration experts in the Berkeley Bioinformatics & Ontologies Project (BBOP).