Informatics obstacles for integrating evolutionary phenotype diversity with model organism data

Hilmar Lapp[1], Jim Balhoff[1], Cartik Kothari[1], Todd Vision[1,2], 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

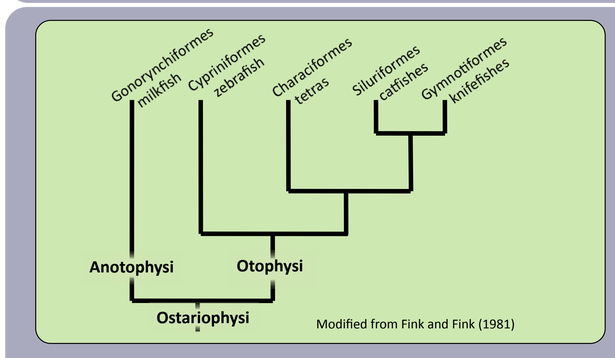
PHENOSCAPE http://phenoscape.org

National Evolutionary Synthesis Center NESCent

Motivation

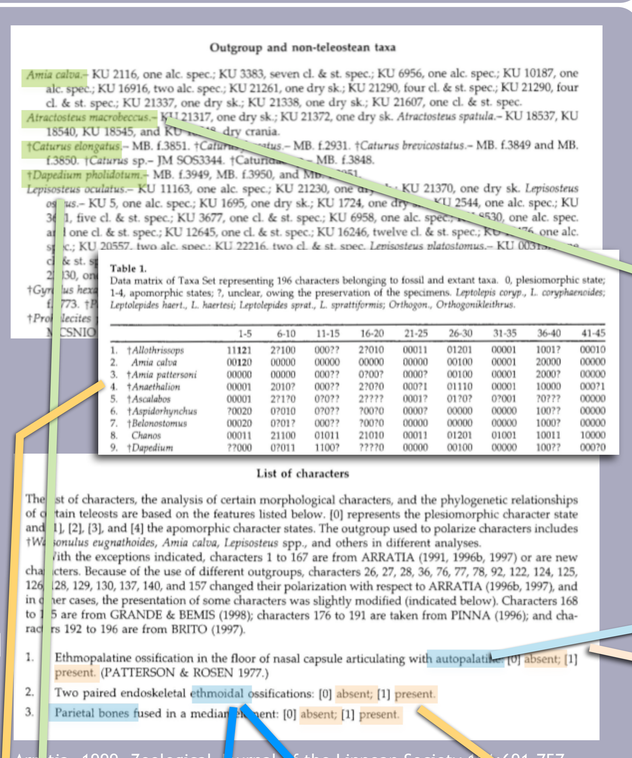
Model organism databases such as the Zebrafish Information Network (ZFIN) use terms from anatomy and quality ontologies to describe the thousands of phenotypes known from genetic studies. By contrast, the vast stores of evolutionary phenotype diversity data are typically rendered in free-form text in journal articles. This makes it very difficult 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?"

The Phenoscape project aims to develop the tools necessary to facilitate such questions to be asked on a large scale basis, starting with teleost fish. Here we illustrate the obstacles in effectively integrating the knowledge from the two domains, and the solutions we have devised.



Our initial application domain are the morphological, in particular skeletal, characters that are variable among the Ostariophysi, a clade of teleost fresh-water 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.


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 [illegible] dry crania.

†*Caturus elongatus*.– MB. f.3851. †*Caturus* [illegible]*tus*.– MB. f.2931. †*Caturus brevicostatus*.– MB. f.3849 and MB. f.3850. †*Caturus* sp.– JM SO53344. †*Caturus* [illegible]– MB. f.3848.

†*Dapedium pholidotum*.– MB. f.3949, MB. f.3950, and MB. [illegible]951.

Lepisosteus oculatus.– KU 11163, one alc. spec.; KU 21230, one dry sk.; KU 21370, one dry sk. *Lepisosteus* [illegible]*us*.– KU 5, one alc. spec.; KU 1695, one dry sk.; KU 1724, one dry sk.; KU 2544, one alc. spec.; KU 3[illegible]1, five cl. & st. spec.; KU 3677, one cl. & st. spec.; KU 6958, one alc. spec.; KU [illegible]530, one alc. spec. and one cl. & st. spec.; KU 12645, one cl. & st. spec.; KU 16246, twelve cl. & st. spec.; KU [illegible]776, one alc. spec.; KU 20557, two alc. spec.; KU 22216, two cl. & st. spec. *Lepisosteus platostomus*.– KU 00[illegible]

†*Gyrodus hexagonus*.– MB. f.773. †*Pro[illegible]decites* [illegible]CSNIO[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., Orthogoniklelithrus.*

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45
1. †*Allothrissops*	11121	27100	0007?	2?010	00011	01201	00001	1001?	00010
2. *Amia calva*	00120	00000	00000	00000	00000	00100	00001	20000	00000
3. †*Amia pattersoni*	00000	00000	0007?	0?00?	0000?	00100	00001	2000?	00000
4. †*Anaethalion*	00001	2010?	0007?	2?070	00071	01110	00001	10000	00071
5. †*Ascalabos*	00001	27170	0?0??	2????	00017	01?07	07001	70???	00000
6. †*Aspidorhynchus*	?0020	07010	0?0??	?0070	00007	00000	00000	1007?	00000
7. †*Belonostomus*	00020	0?01?	0007?	?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	1007?	00070

List of characters

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

[illegible]ith the exceptions indicated, characters 1 to 167 are from ARRATIA (1991, 1996b, 1997) or are new cha[illegible]cters. Because of the use of different outgroups, characters 26, 27, 28, 36, 76, 77, 78, 92, 122, 124, 125, 126[illegible]28, 129, 130, 137, 140, and 157 changed their polarization with respect to ARRATIA (1996b, 1997), and in [illegible]her cases, the presentation of some characters was slightly modified (indicated below). Characters 168 to [illegible]5 are from GRANDE & BEMIS (1998); characters 176 to 191 are taken from PINNA (1996); and characters 192 to 196 are from BRITO (1997).

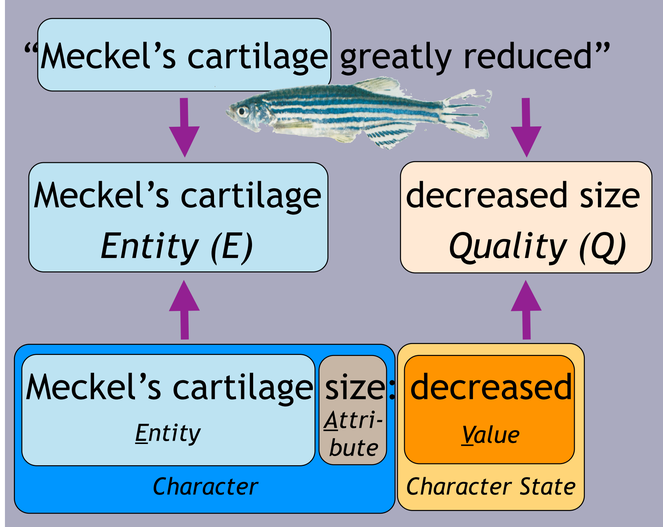
1. Ethmopalatine ossification in the floor of nasal capsule articulating with autopalatine: [0] absent; [1] present. (PATTERSON & ROSEN 1977.)
2. Two paired endoskeletal ethmoidal ossifications: [0] absent; [1] present.
3. Parietal bones fused in a median element: [0] absent; [1] present.

Arratia, 1999. Zoological Journal of the Linnean Society 1[illegible]691-757.

Our initial application domain are the morphological, in particular skeletal, characters that are variable among the Ostariophysi, a clade of teleost fresh-water fishes containing zebrafish and approximately 9,000 other species (~28% of all known fish species).

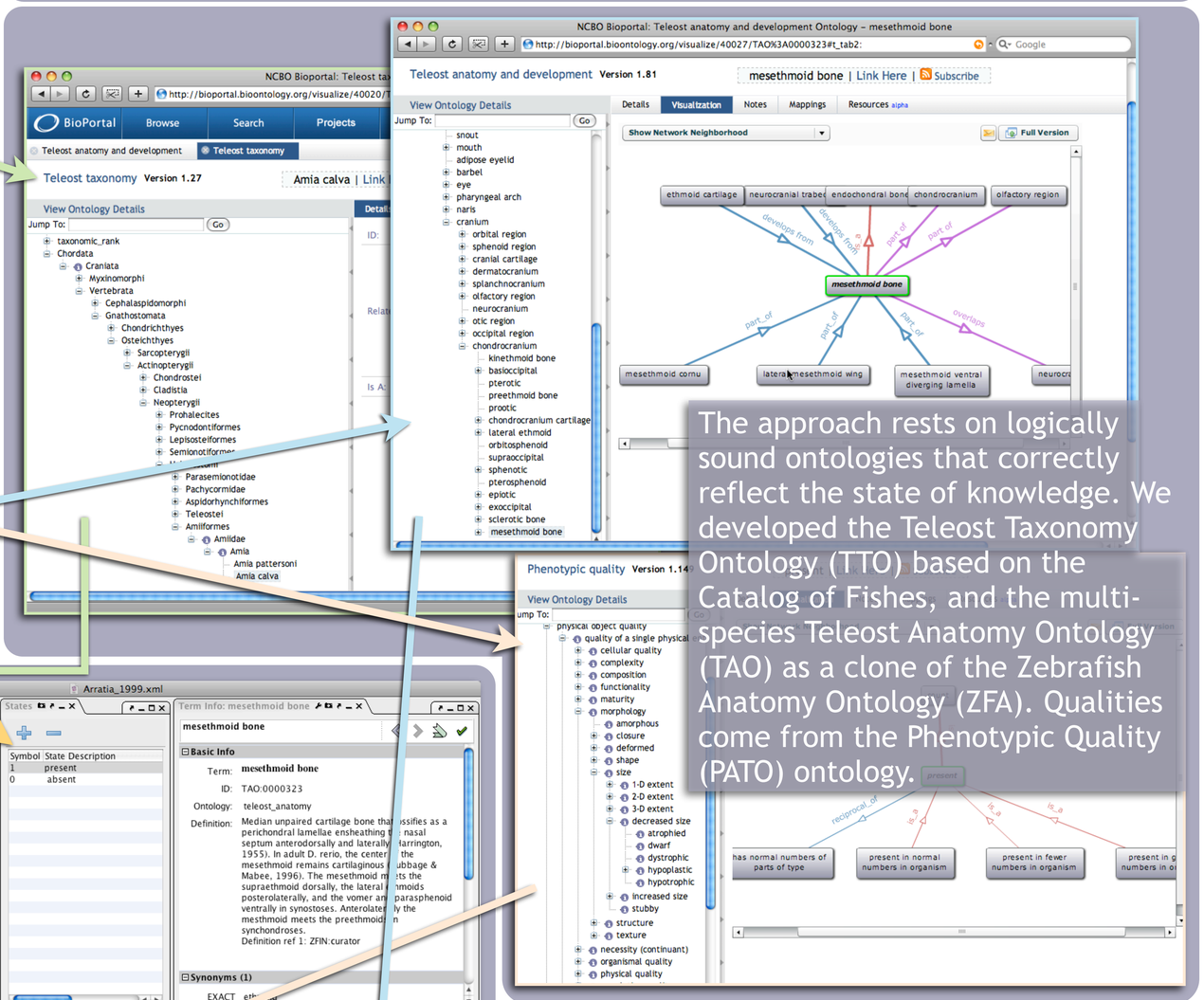
Such phenotypes are traditionally reported in free text form as states of comparative characters in data matrices published in the systematics literature.

Semantic integration through computable assertions

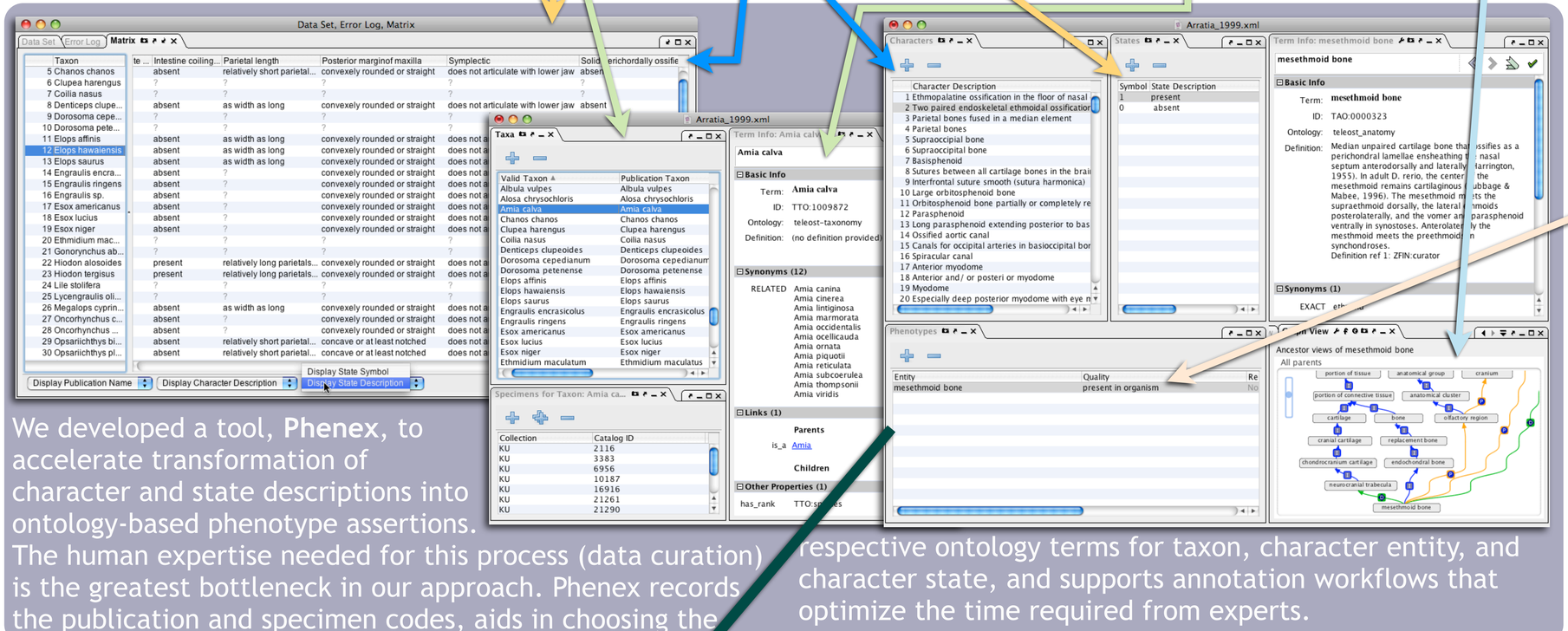


Model organism databases have achieved interoperable phenotype annotations by expressing these formally in Entity-Quality (EQ) syntax, and using ontology terms for entity as well as quality terms. An EQ statement asserts that the quality Q *inheres in* the bearer entity E. A logic reasoner can infer additional statements implied by a body of such assertions and the respective ontologies.

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



The approach rests on logically sound ontologies that correctly reflect the state of knowledge. 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). Qualities come from the Phenotypic Quality (PATO) ontology.



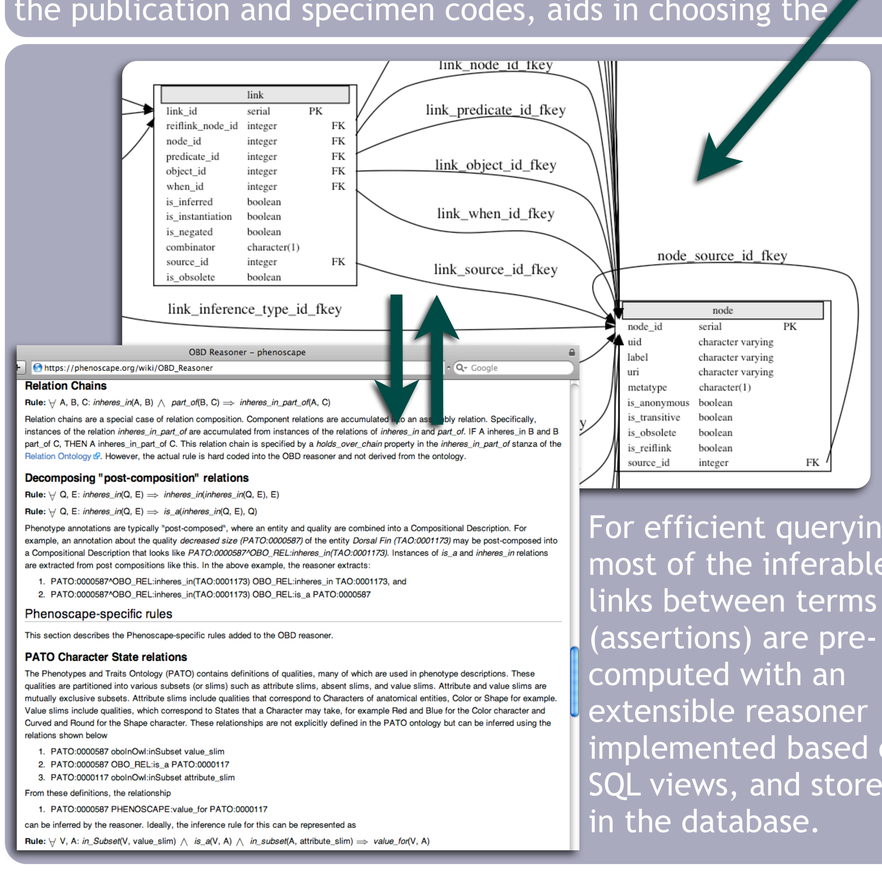
We developed a tool, **Phenex**, to accelerate transformation of character and state descriptions into ontology-based phenotype assertions. The human expertise needed for this process (data curation) is the greatest bottleneck in our approach. Phenex records the publication and specimen codes, aids in choosing the respective ontology terms for taxon, character entity, and character state, and supports annotation workflows that optimize the time required from experts.

The website allows searching by entity term, taxon, and by gene.

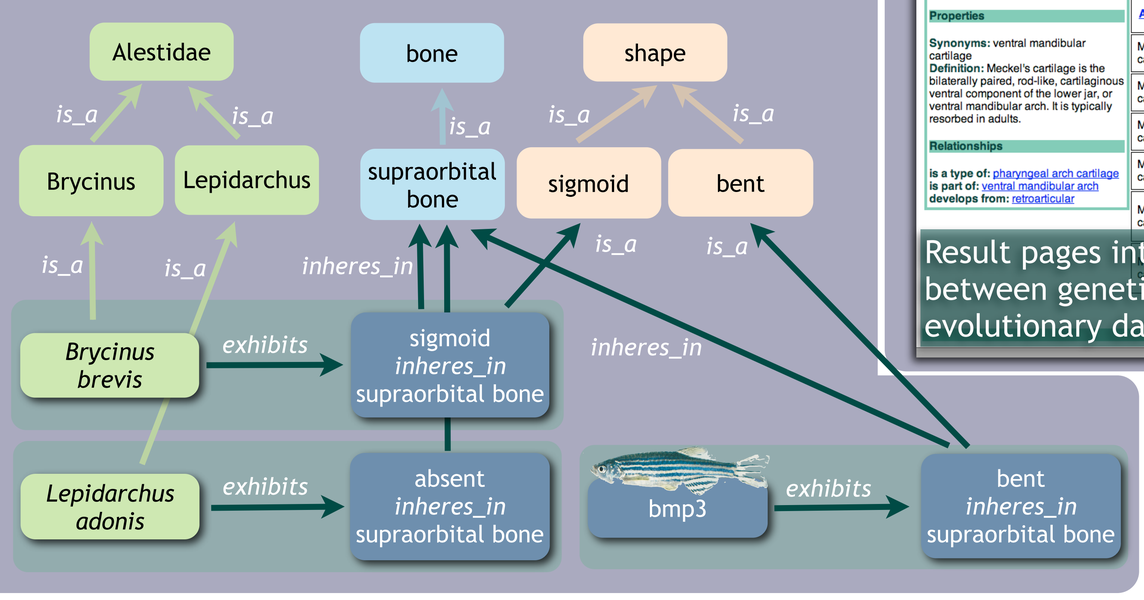
Result pages integrate between genetic and evolutionary data.

Knowledge Base

All ontologies (in OBO format), phenotype assertions (in NeXML files), and the genetic data and phenotypes from ZFIN (tabular) are integrated in a semantic web-like database (OBD) that aims to be a scalable triple-store implementation in SQL.



For efficient querying, most of the inferable links between terms (assertions) are pre-computed with an extensible reasoner implemented based on SQL views, and stored in the database.



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 and OBD work in the Berkeley Bioinformatics & Ontologies Project.