EO: THE EVOLUTION ONTOLOGY

Adam M. Goldstein

Darwin Digital Library of Evolution ¶ http://darwinlibrary.amnh.org
American Museum of Natural History ¶ 79th & Central Park West, New York, NY 10024
agoldstein@iona.edu ¶ http://www.iona.edu/faculty/agoldstein

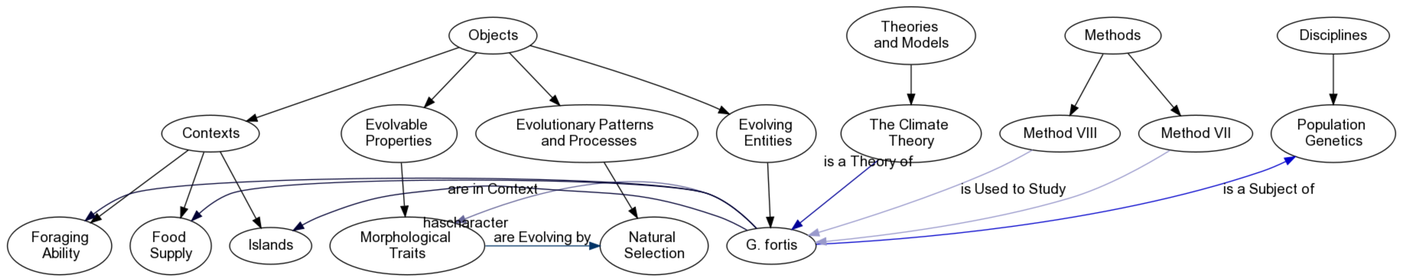

Figure 1: *Classification by EO of a work about Galápagos finches [4]. The top 4 classes are essential to EO, as are the second level under* `Objects`*. Deeper levels are provided by Endler [3, 144] or from the larger context of work on Galápagos finches.* `Method VII` *and* `Method VIII` *are methods for detecting natural selection described by Endler;* `The Climate Theory` *is the view that fluctuations in rainfall are responsible for evolution of the finches morphology.* `Contexts` *contains what are often thought of as environments. The work being categorized is* about *all those classes at the bottom-most level and their interrelations. See sec 2, "Semantics," below.*

Abstract

Existing ontologies model components of evolution, but none synthesize them, or describe the broader framework of ideas used to conceptualize and study evolution. EO aims to do just this. EO is intended as a metadata framework for data and literature curation and data mining. Current projects include organizing literature in a growing, comprehensive bibliographic database about evolution, and characterizing Darwin's works.

1. Motivations

Metadata schemes for describing processes of evolution and the means of studying and conceptualizing them are notably lacking. Excellent resources exist for constructing phylogenies and identifying genes or physiological processes similar in structure or function [1] [5]. But there is no way of organizing literature and data about phylogeny, genes, and the like that illuminates *how and why they got to be the way they are*. Consider Petren and colleagues' [6] study of Galápagos finches. They believe that their most important conclusions concern the relative roles of geography, random drift and gene flow, natural selection, genetic diversity, and phenotypic diversity in an adaptive radiation. Yet the paper's MeSH keywords primarily concern genetics and physiology, with no mention of evolutionary processes. Sequence information about the loci under study were submitted to GenBank; a researcher can place them in perspective across taxa and as a part of physiological processes—but no information is provided about Petren et. al.'s discovery of how and why they evolved. EO is intended to fill this lacuna.

2. Semantics

Most development work on EO has aimed at generating a correct description of evolution and the scientific practices used to study it. Implementation in machine-readable form is wasted if the subject is not characterized properly. EO is comprised of four central hierarchies: One for objects of study, three for scientific practices used by evolutionists.

Table 1: *Predicates for the object classes*

Source of x	Source of y	Pxy	Example
`Evolving entities`	`Contexts`	x is in context y	x: Arctic charr y: Post-glacial lake
`Evolving entities`	`Evolvable properties`	x has character y	x: *G. fortis* y: Shallow beak
`Evolvable properties`	`Processes & patterns`	x evolves by y	x: Genotype ratio y: Heterosis

Objects of study This hierarchy contains classes for *objects in nature* studied by evolutionary scientists: Taxa; processes of evolution such as natural selection; physiological and molecular traits and processes; ecological entities and processes; and others at many different levels of biological hierarchy. There are four subclasses of `Objects of Study` describing the essential properties of an evolutionary process.

Evolving entities An entity that persists in time across generations or analogous cycles of propagation; e.g., species.

Evolvable properties Of an evolving entity, a heritable property having alternative states, the level of variation of which can change in time; e.g., allele frequency.

Contexts An event or state causing change in the level of variation of an evolvable property across generations or analogous cycles; the causal background to an evolutionary change. In the case of organisms, "context" as intended here is often identified with the organisms' environment; e.g., disease, predation.

Evolutionary patterns and processes A *pattern* is a signature sequence of changes through time in an evolvable property; e.g., punctuated equilibria—changes in evolvable properties responsible for taxonomic differences from other species arise at a species' origin, after which those properties remain stable for the remaining lifetime of the species; no particular type of context is implied. A *process* is a signature sequence in time in an evolvable property, in a signature context; e.g., the process of natural selection, which requires conditions or agents inducing differences in adaptedness among variants of an evolvable property.

Together with three predicates, "x is in context y", "x has character y", and "x evolves by y", these classes describe the essence of evolution: Some member of `Evolving entities` that *has some character* that is an `Evolvable Property` is *in some context* in `Contexts` and *evolves by* some member of `Patterns and processes` (Table 1).

Table 2: *Predicates that link hierarchies*

Source of x	Source of y	Pxy	Example
`Objects of study`	`Disciplines`	x is a subject of y	x: Allele freq. y: Pop'n Genetics
`Methods`	`Objects of study`	x is used to study y	x: Pop'n census y: H-W equilibrium
`Theories and models`	`Objects of study`	x is a theory of y	x: Neutral th'y y: Mean fitness

Scientific Practice Three classes describe practices of scientific investigation useful for organizing data and literature.

Disciplines A group of scientists organized around a particular object of study, method, or conceptual framework; e.g., Population Genetics, Paleontology.

Methods Work-flows and processes used to produce or gather data about evolution; e.g., F-statistics; electrophoresis.

Models and theories General claims about, or models of, entities described in `Objects of Study`, which explain or predict the behavior of the latter; e.g., the shifting balance theory, the neutral theory.

Predicates taking entities from these three classes as arguments connect these classes with `Objects of Study` (Table 2). A discipline has, as its aim, the study of certain types of objects, *taking them as a subject*; a method *is used to study* evolutionary phenomena; and a theory about objects of evolution *is a theory of* those phenomena. For example, an information resource is *about* an individual in `Objects of study`, and that individual *is a subject of* an individual in `Disciplines`. Figure 1 shows the classification of a work referred to by Endler [3] using the entire range of EO classes described above.

3. Projects & Implementation

Projects Development of EO is driven by its intended use in the Darwin Digital Library of Evolution (DDLE). The DDLE aims to publish high-quality transcriptions of Charles Darwin's manuscripts, and to serve as an Internet portal to the literature on evolution published by Darwin's precursors and his inheritors up to and including the present day. A database of bibliographic works, the Literature of Evolution (LE), is currently under construction (≈ 3.5K records today) intended to serve the latter purpose. EO is being used to classify the works in this database, and as a framework for text mining Darwin's works to classify subjects he writes about across his *corpus*. LE, its records classified by EO, will serve as an entry point for research on evolution to the already vast but still growing Biodiversity Heritage Library (BHL). Upon completion, the BHL will contain digital copies of almost all works held by an international consortium of libraries of museums of natural history. Absent some subject access point to this literature, works about evolution are all but lost to the researcher.
Implementation EO will be represented in `OWL` (Web Ontology Language) with the intention of integrating with the many already existing ontologies in relevant subject areas. These include ontologies for biological taxa, phenotypes and genotypes, anatomy, physiology, and other elements of `Objects of Study`. Other relevant subject areas include ecology, geology, geography, information resources [2], and descriptions of scientific practice.

References

[1] 2007. The OBO foundry. http://obofoundry.org/. Accessed 18 July 2009.

[2] D'ARCUS, BRUCE and FRÉDÉRICK GIASSON. 2008. Bibliographic ontology specification. http://bibliontology.com/. Accessed 18 July 2009.

[3] ENDLER, JOHN A. 1984. *Natural Selection in the Wild*. Princeton: Princeton University Press.

[4] GRANT, P. R., B. R. GRANT, J. N. SMITH, I. J. ABBOTT, and L. K. ABBOTT. 1976. Darwin's finches: Population variation and natural selection. *Proceedings of the National Academy of Sciences, USA* 73: 257–61.

[5] NCBI. 2009. Entrez, the life sciences search engine. http://www.ncbi.nlm.nih.gov/. Revised, 18 June 2009; accessed 18 July, 2009.

[6] PETREN, K., P. R. GRANT, B. R. GRANT, and L. F. KELLER. 2005. Comparative landscape genetics and the adaptive radiation of Darwin's finches: The role of peripheral isolation. *Molecular Ecology* 14: 2943–57.

International Conference on Biomedical Ontology, 24 July 2009, Buffalo NY