Semantic linking of phenotypes and environments: A review

Anne E Thessen, Daniel E Bunker, Pier Luigi Buttigieg, Laurel D Cooper, Wasila M Dahdul, Sami Domisch, Nico M Franz, Pankaj Jaiswal, Carolyn J Lawrence-Dill, Peter E Midford, Christopher J Mungall, Martín J Ramírez, Chelsea D Specht, Lars Vogt, Rutger Aldo Vos, Ramona L Walls, Jeffrey W White, Guanyang Zhang, Andrew R Deans, Eva Huala, Suzanna E Lewis, Paula M Mabee

Understanding the interplay between environmental conditions and phenotypes is a fundamental goal of biology. Unfortunately, data that include observations on phenotype and environment are highly heterogeneous and thus difficult to find and integrate. One approach that is likely to improve the status quo involves the use of ontologies to standardize and link data about phenotypes and environments. Specifying and linking data in this manner will allow researchers to increase the scope and flexibility of large-scale analyses aided by modern computing methods. Investments in this area would advance diverse fields such as ecology, phylogenetics, and conservation biology. While several biological ontologies are well-developed, using them to link phenotypes and environments is rare because of gaps in ontological coverage and limits to interoperability among ontologies and disciplines. In this review, we present 1) use cases from diverse disciplines to illustrate questions that could be answered more efficiently using a robust linkage between phenotypes and environments, 2) two proof-of-concept analyses that show the value of linking phenotypes to environments in fishes and amphibians, and 3) two proposed example data models for linking phenotypes and environments using the extensible observation ontology (OBOE) and the Biological Collections Ontology (BCO) that can serve as a starting point for the development of a data model linking phenotypes and environments.

Semantic Linking of Phenotypes and Environments: A Review

Anne E. Thessen[1,2]*
Daniel E. Bunker[3]
Pier Luigi Buttigieg[4]
Laurel D. Cooper[5]
Wasila M. Dahdul[6]
Sami Domisch[7]
Nico M. Franz[8]
Pankaj Jaiswal[5]
Carolyn J. Lawrence-Dill[9]
Peter E. Midford[10]
Christopher J. Mungall[11]
Martín J. Ramírez[12]
Chelsea D. Specht[13]
Lars Vogt[14]
Rutger Aldo Vos[15]
Ramona L. Walls[16]
Jeffrey W. White[17]
Guanyang Zhang[8]
Andrew R. Deans+[18]
Eva Huala+[19]
Suzanna E. Lewis+[11]
Paula M. Mabee+[6]
[1]The Data Detektiv, Waltham, MA USA annethessen@gmail.com 443.225.9185
[2]The Ronin Institute for Independent Scholarship, Monclair, NJ USA
[3]Department of Biological Sciences, New Jersey Institute of Technology Newark, NJ USA thedbunker@gmail.com
[4]Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung, Germany pbuttigi@mpi-bremen.de
[5]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR USA jaiswalp@science.oregonstate.edu and cooperl@science.oregonstate.edu
[6]Department of Biology, University of South Dakota, Vermillion, SD USA Wasila.Dahdul@usd.edu and

Paula.Mabee@usd.edu
[7]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT USA
sami.domisch@yale.edu
[8]Arizona State University, Tempe, AZ USA nico.franz@asu.edu and gzhang44@asu.edu
[9]Departments of Genetics, Development and Cell Biology and Agronomy, Iowa State University, Ames, IA USA triffid@iastate.edu
[10]peter.midford@gmail.com
[11]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA cjmungall@lbl.gov and suzi@berkeleybop.org
[12]Division of Arachnology, Museo Argentino de Ciencias Naturales – CONICET, Buenos Aires, Argentina ramirez@macn.gov.ar
[13]Departments of Plant and Microbial Biology & Integrative Biology, University and Jepson Herbaria, University of California, Berkleley, Berkeley, CA, USA cdspecht@berkeley.edu
[14]Universität Bonn, Institut für Evolutionsbiologie und Ökologie, Bonn, Germany lars.m.vogt@googlemail.com
[15]Naturalis Biodiversity Center, Leiden, The Netherlands rutger.vos@naturalis.nl
[16]iPlant Collaborative, University of Arizona, Tucson, AZ USA rwalls@iplantcollaborative.org
[17]USDA-ARS, US Arid Land Agricultural Research Center, Maricopa, AZ USA Jeffrey.White@ARS.USDA.GOV
[18]Department of Entomology, Pennsylvania State University, University Park, PA, USA adeans@psu.edu
[19]Phoenix Bioinformatics, Redwood City, CA USA huala@phoenixbioinformatics.org

*corresponding author
+co-PIs of the Phenotype RCN grant

Abstract

Understanding the interplay between environmental conditions and phenotypes is a fundamental goal of biology. Unfortunately, data that include observations on phenotype and environment are highly heterogeneous and thus difficult to find and integrate. One approach that is likely to improve the status quo involves the use of ontologies to standardize and link data about phenotypes and environments. Specifying and linking data in this manner will allow researchers to increase the scope and flexibility of large-scale analyses aided by modern computing methods. Investments in this area would advance diverse fields such as ecology, phylogenetics, and conservation biology. While several biological ontologies are well-developed, using them to link phenotypes and environments is rare because of gaps in ontological coverage and limits to interoperability among ontologies and disciplines. In this review, we present 1) use cases from diverse disciplines to illustrate questions that could be answered more efficiently using a robust linkage between phenotypes and environments, 2) two proof-of-concept analyses that show the value of linking phenotypes to environments in fishes and amphibians, and 3) two proposed example data models for linking phenotypes and environments using the extensible observation ontology (OBOE) and the Biological Collections Ontology (BCO) that can serve as a starting point for the development of a data model linking phenotypes and environments.

Introduction

Phenotype is the expression of interactions between genotype and environment. This relationship is fundamental to a wide range of biological research, from large-scale questions about the effect of climate change on global ecosystems to small-scale questions involving disease presentation in individual organisms. The urgency of such questions, coupled with the "data deluge" (Hey, Tansley & Tolle, 2009), has motivated scientists to explore more efficient ways to aggregate and explore life science data. Traditional methods of data dissemination, publication, and deposition in stand-alone databases do not support the rapid, automated, and integrative methods of data exploration needed to efficiently address pressing research priorities.

Two important barriers to understanding interactions between environment and phenotype are the heterogeneity of terms and their imprecise definitions in data sets and manuscripts. An ontology has the potential to tame this heterogeneity and allow researchers to more efficiently query and manipulate, large-scale data sets (Fig. 1); however, several challenges must be overcome before their benefits are realized. Historically, bio-ontologies first came into popular use in the biomedical and model organism communities (Côté & Robboy, 1980; Spackman, Campbell & Côté, 1997; Ashburner et al., 2000), but they are now being applied to address much broader, comparative problem complexes (Mabee et al., 2007; Deans et al., 2015; Dececchi et al., in press). A shift towards representing and reasoning over taxonomically diverse phenotypes in an ontological framework has occurred in a period of less than 10 years and, not surprisingly, brings about new semantic, computational, and even social challenges (Gerson, 2008). In this paper, we explore the difficulties of automated linking of environments and phenotypes, review the current state-of-the-art, present use cases, and propose solutions to frequently encountered problems.

Clearly representing the natural language descriptions of phenotypes and environments with a set of ontologies is difficult, because natural language, while highly expressive, is often semantically ambiguous and reliant on context (Sasaki and Putz, 2009; Seltmann et al., 2013). Despite successes in developing standards within specific disciplines (e.g., Taylor et al., 2008), standard vocabularies are rare and seldom widely or consistently used (Enke et al., 2012). Individual scientists often have preferred terms with undocumented and highly nuanced meanings (Chang & Schutze, 2006; see discussion in Huang et al., 2015). Further, there is a co-evolution between natural language and ontologies, which can complicate the recording of provenance and backwards-compatibility (Seppälä, Smith & Ceusters, 2014; Ochs et al., 2015). Thus, as it stands, a researcher wishing to perform a meta-analysis has to manually integrate data sets, which often requires discussions with data providers to clarify meaning.

In addition to the intricacies of natural language used to describe phenotypes and environments, the ontological representation of environments requires additional considerations. Environments are often described using a collection of semantically complex (and often ambiguous) terms that are applied differently across disciplines. Semantic representation of terms such as “environment”, “ecosystem”, “habitat”, “ecozone”, “bioregion”, and “biome” must account for variable biological, ecological, geographic, geopolitical, and historical usage. As a result of this complexity, many specialized environmentally-themed terms such as “Afrotropical”, which are central to fields such as zoogeography and floristic science, are not yet included in any ontology. Data about species interactions and behavior can be an essential component in defining an organism’s environment, but current ontological structures do not include behavior regulation classes that can be tied to ecological processes (e.g., negative regulation of foraging behavior by predation pressure). Environments can also be defined using a more data-driven approach where a specific environment is defined as the intersection of different factors, e.g., defining a desert as having a specific annual precipitation, temperature range, and solar irradiation. In the field of plant science, some consider field management practices (including, e.g., frequency irrigation or fertilizer application) to be a component of environment, whereas others consider field management to differ from the environment because these changes to conditions are not a part of the natural environment. Because of differences in perspective, data are organized in different ways across multiple resources. These inconsistencies in describing environments complicate analyses that might identify associations among occurrences, phenotypes and environmental conditions.

In spite of these challenges, understanding relationships between the semantics of environment and phenotype is fundamental for data integration and scientific progress in the fields of conservation, agriculture, disease control, organismal development, and numerous others in biology. Thus, there is a need for a more developed, flexible, and interlinked ontology framework representing environments, phenotypes, and their interplay. This framework for environments and phenotypes can allow automated inferencing over large, aggregated data sets, as demonstrated for gene functions and biological processes (Ashburner et al., 2000). Below, we present use cases that illustrate research questions that would benefit from semantically linking environments and phenotypes and describe existing efforts working toward this goal.

Background

Within the field of informatics, classification strategies range from flat lists of terms, to vocabularies, and ontologies. For example, a vocabulary might merely contain the terms “bone”, “leg”, “femur” and their definitions. An ontology would further define these as classes and with respect to their biological relationships by asserting that a “femur” is a type of “bone” and part of the “leg”. Moreover, such assertions are encoded in a standardized, machine-readable form. Thus ontologies empower computers to reliably interpret and reason over these logical relationship graphs. A well-known example of the technology’s potential is provided by IBM’s Watson (Gliozzo et al., 2013). Ontologies are typically recorded in a syntax format such as the Web Ontology Language (OWL; W3C OWL Working Group, 2012) or the Resource Description Framework (RDF; http://www.w3.org/RDF/) that can be readily distributed and exchanged by computers, thereby facilitating knowledge integration within a scientific community. For an ontology to actually be useful to scientists, these same scientists must mutually agree upon, develop, and nurture a shared collection of ontologies and the processes for maintaining it.

Over 40 ontologies and vocabularies have been created to describe environment and phenotype (Table 1). Some of these resources are extended and refined through incorporation of user and developer requests for new terms and cross-referencing terms to existing vocabularies. Like software development, an essential aspect of ontology development is constant evaluation through active use: describing data sets and asking key biological questions. To this end, a number of groups are making the first inroads in the use of ontologies for studying the impact of environment on phenotypes (e.g., The Encyclopedia of Life; Pafilis et al., 2015). The Minimum Information for any Sequence (MIxS; Yilmaz et al., 2011) metadata checklist, a product of the Genomic Standards Consortium (GSC; Field et al., 2011), does not specifically link phenotypes to environments, but does include fields for describing environments using the Environment Ontology (ENVO; www.environmentontology.org; Buttigieg et al., 2013) as well as the host phenotype as part of the host-associated genome/metagenome environmental package. Although MIxS recommends using terms from the Phenotypic Quality Ontology (PATO; Gkoutos et al., 2004) in the host phenotype field, host phenotypes can be complex and could be described via a mixture of phenotype ontologies (such as the Human Phenotype Ontology (Köhler et al., 2014) or Mouse Phenotype Ontology (Gkoutos et al., 2004). The International Consortium for Agricultural Systems Applications (ICASA) has built an infrastructure for combining genotype, environment, and management data in agricultural analyses using a hierarchical data dictionary (Hunt, White & Hoogenboom, 2001; White et al., 2013). This infrastructure is being integrated in crop and climate modeling efforts, notably through the Agricultural Model Intercomparison and Improvement Project (AgMIP), which promotes efforts to "simulate yield response to climatic factors, abiotic factors, and genotypic variables" (http://research.agmip.org/). Oellrich et al. (2015) recently developed a standardized method for describing and analyzing the phenotypes associated with characterized mutant genes across species that includes environmental terms from the Plant Environment Ontology (EO). Despite this progress, the available environment and phenotype ontologies still contain major gaps in the coverage of their respective domains, and significant investment is needed before data integration and analytics can be accomplished on a large scale.

Use Cases

To communicate the importance of investing in environment and phenotype ontologies, we present use cases drawn from several life science domains. These use cases represent the types of research questions that either cannot currently be answered or can only be answered with great difficulty.

Using Phenotype and Environment Ontologies in Ecology

Coping with Climate Change in Conservation, Management, and Agriculture

Example Question: Which species or crop varieties are projected to do well in my locality over the next century?

Background:

Climate change is anticipated to affect environmental conditions with unprecedented speed. Knowledge concerning the responses of ecological communities to these changes is very limited: adaptation and migration are among numerous possibilities that must be considered. Conservation and agricultural resources are also limited, so identifying and focusing interventions on taxa that are less able to adapt can be very helpful. Besides commonly-used projections based on species distributions models, another strategy for identifying at-risk species is to assess their vulnerability based on their traits (i.e., phenotypes). By linking phenotypes to specific environmental conditions, taxa or strains that are likely to thrive (or not) under those conditions can be identified. For example, cataloguing phenotypes that are more prevalent among organisms that live in hot and wet environments and detecting their presence in organisms whose environments are warming and becoming more humid, allows some bearing on the later organisms' ability to cope with such climate change. In agriculture, this can be used to identify crop varieties that are likely to give higher or more stable yields under specific conditions or wild relatives of crops that may possess useful traits. In conservation, similar analyses can identify species at risk of extinction (Thormann et al., 2014). One system that hints at performing this type of analysis currently is SemanticWildNET (Henderson, Khan & Hunter, 2007), which links data about birds and snakes to environmental conditions in Australia.

Current Workflow:

Steve works for a seed company that serves the southern Great Plains in the USA. General Circulation Models (GCMs) project that over the next 25 to 30 years farmers in the southern Great Plains will experience drier, warmer, and longer summers. His company wants to start breeding sorghum hybrids that will perform well in these future conditions. Steve's company has developed and phenotyped a wide range of parental lines that differ in yield response and phenology under different environmental conditions, much of which is proprietary data. He is able to find additional data sets that link geolocations and associated environmental conditions to taxon phenotypes for crops (e.g., GRIN, the Germplasm Resources Information Network; http://www.ars-grin.gov/npgs/) and link taxa to phenotypes (e.g. the TRY Plant Trait Database, https://www.try-db.org/; Kattge et al., 2011). Environmental data sets that include information about weather (NOAA), soil (USGS) and climate projections on a 1 km spatial grain (WorldClim data set, Hijmans et al., 2005) are publicly available through government sources.

Steve decides that the best strategy for finding the top hybrid for a specific future habitat is to manually link phenotypes to environmental conditions using the taxon name and location (geographic coordinates) as a metadata bridge (Fig. 2). To work with the data, Steve must download the files to a local machine. Because he does not have programming skills, he must manually locate the specific data of interest from each data source and then make decisions about appropriate integration using written documentation from the data provider. The data preparation and integration takes six months of full time work.

When Steve finally has his data ready to analyze he must pick a statistical workflow and software package that can identify phenotypes and environmental conditions that are observed together. The next step would be to look at the climate projections to find the projected environmental conditions his customers are likely to be facing and use these to identify the ideal suite of phenotypes for that future climate regime. The final step would be to identify the taxa that have the phenotypes in question.

Future Workflow: Agriculture

Steve works for a seed company that serves the USA. GCMs project that over the next 25 to 30 years farmers in the southern Great Plains will experience drier, warmer, and longer summers. His company wants to start breeding sorghum hybrids that will perform well in these future conditions. Steve's company has developed and phenotyped a wide range of parental lines that differ in yield response and phenology, and these lines have been annotated using ontology terms for traits (e.g., TO and PATO, Table 1) and the corresponding growth conditions (EO, Table 1). Additionally, the habitat of each line (or its source organisms) is described by classes from an environment ontology (ENVO, Table 1)

Steve queries his company's internal, semantically aware database for annotated records corresponding to the lines his company has developed that have high yields under warmer climatic conditions and when subject to drought. This gives him a list of candidate lines (i.e., phenotypes and genotypes) to use in development of new hybrids for the region. If the list of these candidates does not provide sufficient resistance to high temperatures and drought, he may choose to query a database containing information on the wild relatives of sorghum, along with average rainfall and temperature data from the natural habitat of each species and/or annotations describing their habitat using classes from an environment ontology. If necessary he will be able to introgress genetic material encoding drought or high temperature tolerance from a wild relative of sorghum into his breeding lines.

Future Workflow: Wildlife Conservation

Lupita is a park ranger that manages a coastal wildlife sanctuary. Some of the species in her sanctuary are listed as threatened by the IUCN. According to the latest climate change projections, her sanctuary is going to be hotter and wetter in 50 years. She has limited resources to maintain the biodiversity in her sanctuary for the long term. After some thought, she decides to identify at-risk species by comparing the traits of the organisms in her park with traits of organisms that do not do well in hot, wet, coastal environments. Lupita searches a semantically aware, publicly-available biology database and finds a list of traits for vertebrates whose habitats do not include hot, wet, coastal environments and a list of traits for vertebrates with habitats that do. Comparing differences between the two data sets gives a list of candidate traits which suggest a taxon would be vulnerable to the projected climate regime. Searching for these traits across the species in her sanctuary, Lupita identifies two species that are highly likely to fare poorly in the projected climate, and she devotes resources to their conservation.

Challenges Today:

A large proportion of phenotype and environment data are part of the “long tail of dark data” (Heidorn, 2008) that are not currently digital or discoverable. Although some phenotype, environment, genotype, and climate projection datasets are available, they are difficult to find and interrelate. These types of datasets can be cross-linked using space, time, or taxon, but the formats of the different datasets can pose a challenge to integration (e.g., Reed, White & Brown 2003). In addition, metadata across disciplines, data types, and time periods are rarely consistent. Key data items used for integration, such as taxonomic names, change over time and lead to poorly linked data (Edwards et al., 2011; Giles, 2011; Page, 2008; Franz et al., 2015).

Using Phenotype and Environment Ontologies in Taxonomy

Connecting Specimen Phenotypes to Environment

Example Question: Which traits are common to beetles collected from deserts?

Background:

Natural history collections worldwide contain approximately two billion specimens across various taxonomic groups (Ariño, 2010). Tens of millions of these specimens have their phenomes at least partly described in the form of published taxonomic descriptions and may have environmental data recorded on the specimen label or in a field notebook. Much of these data have not been digitized and can be difficult to find and use. Connecting specimen-based phenotype data to environmental information that describes where the specimen was isolated can support predictive modeling of diversity and distribution.

The current massive digitization effort applied to collections is primarily done manually, though efforts are being made to automate where possible (Barber, Lafferty & Landrum, 2013). The environmental data associated with a specimen, typically a note on a specimen label, is typically transcribed verbatim. If a curator wants to annotate a specimen with an environment or habitat type or other metadata, the process of reading the information and relating it to an ontology class is entirely manual. This is a very time-consuming workflow that involves reconciling synonyms and disambiguating homonyms. Ideally, much of the manual labor of reconciliation, disambiguation, and assignment would be shifted to a machine with curators intervening only periodically.

A semantic model for representing specimen phenotypes has been developed (Balhoff, Yoder & Deans, 2011) and applied to taxonomic descriptions (Mullins et al., 2012; Balhoff et al., 2013). This model applies phenotype statements directly to specimens. Each specimen, residing in an institutional collection, is associated with collecting event data, including where, when, how, and by whom it was collected. The “where”-data could be connected to environment types and other environmental data through semantic annotation using environment ontologies.

Current Workflow:

Kate wants to annotate insect specimens in a research collection with phenotypic and environmental terms. All labels, published reports, and field notebooks concerning this collection have been digitized and processed via optical character recognition (“OCRed”). She logs in as an editor into the museum collections database that allows her to virtually access each specimen and any associated documents. Kate begins working on the first specimen. The interface brings up the label, an image of the specimen, the published description, and the relevant field notebook page. An NLP-assisted algorithm within the interface reads the OCRed documents and highlights environment-related terms in the text. She quickly reads the label and notebook near the automated highlights, then searches ENVO to find the class that most accurately describes the environmental conditions described by the collector. The environment listed in the notebook does not match an existing ENVO class. She creates an issue in the ENVO issue tracker (https://github.com/EnvironmentOntology/envo-p/issues/) requesting a new class that more accurately describes the specimen environment. She does research on what the class should be called and on the suggested definition. Kate will have to wait until someone at ENVO can consider her request before completing the annotation.

After submitting the issue to ENVO, Kate reads the published description and goes back and forth between PATO and the relevant insect anatomy ontology to find the classes she needs to describe the specimen phenotypes and double-checks the classes by looking at the specimen image (when possible). This process is very time consuming. When Kate is finished she adds the relevant classes to the specimen database.

When Kate finally finishes her annotations, any user can query her museum website for specimens that meet specific phenotypic and environmental constraints.

Future Workflow:

Kate wants to annotate insect specimens with phenotypic and environmental terms. All materials concerning this collection have been digitized. Her museum has the cyberinfrastructure that allows her to virtually access each specimen, bring up the related documents, and assign relevant phenotypic and environmental terms through a point-and-click interface. Kate begins working on the first specimen. The interface brings up the label, an image of the specimen, the published description, and the relevant field notebook page. A text-mining tool highlights relevant information in these sources and suggests classes from appropriate phenotype and environment ontologies (e.g., ENVO). Kate agrees with the suggestions and clicks “approve”. For the next specimen, Kate agrees with the proposed phenotype classes, but does not see any appropriate ENVO classes. She rejects the suggestions and is taken to a window that allows her to browse ENVO for more appropriate classes. Still not satisfied, another click takes her to a window that allows her to submit a request for a new class for which she suggests a definition and relevant references. Kate is then taken back to the working environment and is presented with the next specimen. Kate again agrees with the proposed phenotype classes, but the source contains only a latitude and longitude for environment. The granularity offered by a lat/long query does not capture microhabitats the insect may have been exposed to, but she decides that a high-level description of the environment (by using biome or environmental feature classes in ENVO) is preferable to providing no metadata. Kate opens a lat/long query window where she is asked for a latitude, longitude, and date. Altitude is optional and depth is required for a lat/long over water. This query returns environmental data relevant to the date that the specimen was collected and the system suggests ENVO classes, some of which Kate agrees with. When Kate finishes her annotations, any user can query her museum website for specimens that meet specific phenotypic and environmental constraints.

Challenges Today:

Inconsistencies in geographic metadata associated with specimens are a major roadblock in connecting phenotypes and environments (Vollmar, Macklin & Ford, 2010). Specimen metadata are filled with ambiguous and synonymous terms with inconsistent granularity. For example, the Plant Bug Inventory project database (http://research.amnh.org/pbi/; Schuh, 2012) uses thousands of habitat names to describe the localities where insect species were collected, including “cloud forest”, “cloud forest with bamboo” and “cloud forest: oak trees, fern” (G. Zhang pers. comm.). The documentation required to relate these terms to each other is currently absent. In addition, high-level (but imprecise) locality information (e.g., “State College, Penn.”) is quite common for museum specimens and cannot be associated with fine-grained environment types. Further, specimen labels often contain somewhat vague terms such as “neotropical” or “mesohaline” that correspond to broad ecoregional definitions. According to Wikipedia, mesohaline is defined as water that is between 5 and 18 salinity (http://en.wikipedia.org/wiki/Salinity), but it is seldom clear whether a collector has intended a precise definition such as this when writing the label. Thus, associating many specimens currently in collections with well-defined environments may not be possible.

Some specimen metadata include a latitude and a longitude or a locality name, which may be used to infer the environment, but environments change over time. For example, a specimen may have been collected from a desert, which has since been paved over in the expansion of a metropolitan area. Environments are also subject to cycles such as seasonal, diel, or tidal. All of these factors make date and time important metadata. Annotating specimens in more three-dimensional environments, such as the ocean or a mountain plateau, requires yet another piece of information – depth or altitude.

Using Phenotype and Environment Ontologies in Phylogeny

Reconstructing Ancestral Features and Habitats

Example Question: Do species that have independently reduced or lost their eyes share common environments now or in the past?

Background:

To infer the most probable features of a common ancestor given a phylogenetic tree and the phenotypes of extant species, researchers utilize several well-developed parsimony and likelihood methods. Similarly, the habitat preference of living species can be used to reconstruct evolution of ecological niches. Connecting the phenotypic data from species with their habitat and environmental data allows efficient analysis of these associations, allowing, for example, the disentangling of evolutionary adaptation from other causes of phenotypic convergence.

Current methods of ancestral reconstruction rely on the uniform identification of a limited number of environmental traits (e.g., habitats). Users have parsimony, likelihood, and Bayesian methods at their disposal for ancestral state reconstruction (e.g., Mesquite, Maddison & Maddison, 2014; BEAST, BayesTraits, and R packages such as ape). These methods allow for both discrete and continuous values. For discrete characters, ancestral states are calculated from the specific character states (e.g., environmental traits) found in the species. For example, for a clade of species that live in either “deep sea” or “underwater cavern” habitats, ancestral state reconstruction is limited to these discrete habitats, i.e., the ancestor can be hypothesized to have lived in either one or the other habitat. However, an ontology can show that “deep sea” and "underwater cavern" are both subtypes of an "aphotic marine environment", and thus this parent term reveals this as a potential ancestral state for this clade.

Current Workflow:

Jane examines museum specimens of organisms belonging to a clade of freshwater fishes which encompasses several hundred species. She discovers that the eyes vary in their level of development: completely absent in some species, reduced in others, and fully developed in most. After mapping this trait on a well-supported phylogeny, she concludes that eye reduction and loss has occurred independently several times in this clade. This leads her to hypothesize that the changes in eye development are associated with a species' habitat. She goes to the museum databases and finds that the original descriptions of the collection sites for these specimens are recorded as free text in the Darwin Core field "verbatimLocality" (http://rs.tdwg.org/dwc/terms/verbatimLocality). She enters the "verbatimLocality" data into her matrix of features mapped onto the phylogeny. Jane notices that several terms might be synonymous and begins to research the specific definitions of the terms used and does her own research into conditions at each locality. After one month of reconciling locality terms, she begins to notice that species with reduced or absent eyes are all from subterranean environments. She proceeds with her study, now examining other environmental factors or phenotypic traits that might play a role in their shared habitat type.

Future Workflow:

While examining several hundred museum specimens of organisms belonging to a clade of freshwater fishes, Jane discovers that the eyes vary in their level of development. Mapping this trait on a well-supported phylogeny shows that eye reduction and loss has occurred independently several times in this clade. This leads her to hypothesize that the changes in eye development are associated with a species' habitat. She goes to the museum databases and finds that the original description of the place from where these specimens were collected was recorded as free text in the Darwin Core field "verbatimLocality" (http://rs.tdwg.org/dwc/terms/verbatimLocality), but the text is mapped to classes in an environmental ontology such as ENVO. She downloads these classes for all species and adds them to the matrix of features that are mapped to the phylogeny. She sees that species with reduced or absent eyes are from localities variously described as “shallow pool in cave”, “deep water well”, “deep phreatic habitat”, and “swallow hole”. A visualization tool allows her to see the ontological classes which these descriptions have been mapped to as well as any shared hierarchies or relations to other classes. She notices that these descriptions share “groundwater” and an environmental material and “subterranean” as an environmental quality. She proceeds with her study, now examining other environmental factors or phenotypic traits that might play a role in their shared habitat type.

Challenges Today:

As in the other use cases, environmental ontologies must be provisioned to include the classes relevant to a broad range of habitat types. Additionally, and similar to other use cases, phenotypes of taxa that are represented in a computational format must be readily available. The challenge unique to this use case is that methods of phylogenetic optimization that utilize ontological relationships need to be developed. This will require consideration of the hierarchy of class relationships such that the semantic similarity (Pesquita et al., 2009; Resnik, 1999) among differing ancestral states at a particular node is taken into account when calculating the appropriate assignment of a state to that node. Further, visualizations of the distribution of phenotypic and environmental features on the tree that display, e.g., the most similar ontological parent classes across nodes, need to be developed. An attempt to create an ancestral phenotype ontology has previously been made by Ramírez & Michalik (2014).

Using Phenotype and Environment Ontologies in Behavioral Ecology

Including Species Interactions in Habitat Assessments

Example Question: How will this predatory wasp affect the spider population in my vegetable garden?

Background:

Behavior is a phenotype that can be influenced by the presence or absence of other organisms. The presence of other taxa can be just as important as abiotic features for determining suitability of an environment for habitation by members of given species. An observation of a taxon exhibiting a stress behavior is very different from an observation of the same taxon exhibiting a feeding behavior. Changes in the ranges of organisms due to climate change or accidental introduction is another way by which environments change and become more or less suitable for specific phenotypes, such as feeding or courtship behaviors. These behaviors are very important and when they are interrupted, can increase or decrease abundance of the affected organism.

Current methods for retrieving behaviors that might be predictive of species interactions mostly rely on published or unpublished ethograms and incidental comments in taxonomic descriptions or experimental studies. There are databases of species interactions (Poelen, Simons & Mungall, 2014), but these reflect interactions observed and reported in the literature, without the behavioral content to make predictions about possible interactions resulting from the introduction or range expansion of one or both species. Ideally, behavioral descriptions would include specific environmental preferences as well as details of foraging, anti-predator, and courtship behavior. The ability to make predictions of interactions would be an important contribution when considering planned introductions or when setting priorities for preventing unintentional spread.

Current Workflow:

Larry depends on his vegetable garden for food and on the spiders within it for pest control. He frequently sees the jumping spider, *Phidippus clarus* Keyserling 1884, in the garden. *P. clarus* is a widespread and common spider in the Eastern US (Edwards, 2004) and has been demonstrated to be capable of controlling an experimental population of herbivorous insect pests (Hoefler, Chen & Jakob, 2006). Larry hears from his local agricultural extension office that a South American wasp that preys on spiders has been accidently introduced nearby. Should Larry be concerned that the presence of the wasp will lead to more pests in his garden? Larry takes the day off work to go to the local University library and asks a librarian to help him find information about *P. clarus* and the South American wasp. Much of the information he needs is in table format (ethogram) or in narrative text (comments in taxon descriptions and experimental studies) and is difficult to decipher. The librarian makes him aware of a database of species interactions that is easier to understand, but no data for *P. clarus* are available. At the end of the day, Larry is still uncertain about the effect of the wasps on his garden spider population.

Future Workflow:

Larry depends on his vegetable garden for food and on the spiders within it, such as the jumping spider *Phidippus clarus* Keyserling 1884, for pest control. Larry hears from his local agricultural extension office that a South American wasp that preys on spiders has been accidently introduced nearby. Should Larry be concerned that the presence of the wasp will lead to more pests in his garden? Larry checks a gardening app on his mobile device that uses a combination of ontologies and observation data to power a Q&A engine about nature in his area. Through a simple user interface, he asks the app if the wasp is likely to affect the jumping spider and whether there are additional potential consequences. Guided by the ontological structure available in its back-end, the app states that 1) *P. clarus* is known to spend large amounts of time on the tops and tips of plant shoots, and commonly lays its egg sacs near the tips of shoots (Edwards 1980; Hill 2014), and 2) the wasp searches for prey on the tops and tips of plant shoots. The inference engine used by the app are able to predict that the introduced wasp is likely interact with Larry's population of *P. clarus* spiders. Because of where eggs are laid, this would have the potential to interrupt *P. clarus* reproduction and thus reduce pressure on his garden pests. With this information, Larry spends an hour making several wasp traps out of old plastic bottles to place in his garden.

Challenges Today:

Environment ontologies currently do not explicitly incorporate species interactions in the definition of their classes; however, an ontology for describing experimental conditions (EO) does describe interactions between plants and other organisms in their environment. Many environment ontologies, as they are currently structured, may not capture features relevant to whether an environment will support specific behaviors, which can be very important data. Not all taxa will engage in important behaviors in all environments, thus for many studies, presence/absence data are not adequate. Creating a new set of ethological ontologies and developing relations from their classes to those present in environmental ontologies has great potential to address these issues, but requires significant effort to realize and maintain.

DISCUSSION

Challenges

The process of developing the Use Cases identified several major barriers to linking phenotype and environment. These fall into two categories: challenges of coverage and challenges of interoperability.

Challenges of Coverage

Variable Granularity:

Environmental data are reported with varying degrees of granularity that can take the form of nested categories (e.g., continent – country – province – township – street), intervals (e.g., ± 30 km), or significant digits (e.g., 5.236 vs 5.2). Some data sets, especially species observations, include highly granular metadata specifying the exact location or exact conditions under which a specimen was collected (such as collecting an insect from under tree bark or collecting an organism in the presence of its predator). Although existing ontologies cover many of the scales of interest, gaps prevent sufficient detail to capture all of the environmental data provided in connection with collected specimens or published studies. These are critical for some taxa, such as insects collected from under bark (Jain & Balakrishnan, 2011). Currently, such data are not discoverable due to the paucity of terms in existing ontologies and the lack of easy-to-use tools that allow for semantic annotation with multiple ontologies.

Terms and Definitions:

One of the biggest challenges in creating ontologies for application to disciplines that have a long history of published knowledge is the translation of the information in human-readable narrative into a machine-readable form. Human language is very difficult for a machine to understand largely because of its variability and nuance. Different terms (i.e., synonyms) can be used to refer to the same concept, while a single term (i.e., homonyms) can refer to multiple, different concepts. The human brain copes with this uncertainty by understanding context. One way for a machine to cope with the variability of natural language is to provide it with an ontology that includes synonymous terms; however, this can be difficult to maintain because language evolves rapidly. Homonyms are a word-sense-disambiguation problem, which requires heuristics about context to infer meaning; it is an active area of research (Zhan & Chen, 2011). A homonym example applicable to environments is the term "scrubland", which means something very different in California and South Africa. In this case, significant disambiguation could be achieved by cross-referencing terms with geo-location or place names using resources such as GAZ (Buttigieg et al., 2013), a gazetteer developed along ontological principles.

Incomplete Ontologies:

The development of ontologies in the biodiversity sciences has grown rapidly but is relatively new, thus coverage is still small (Table 1). The OBO Foundry Library (http://www.obofoundry.org/), a repository for biological ontologies, contains 22 ontologies relevant to environments and phenotypes, with a total of 136,480 classes. Of these ontologies, only one describes environments (ENVO) and one describes plant environmental conditions in experimental treatments (EO). Eleven are phenotype or anatomy ontologies that cover specific taxonomic groups, such as fungi (FYPO), animals (UBERON primarily for Chordates, with other ontologies such as PORO for specific clades like Porifera (Thacker et al., 2013)), and plants (TO) (Table 1). Many other taxa, such as the microbial eukaryotes, do not have dedicated ontologies. Furthermore, existing ontologies lack many key concepts required for application to the many facets of biodiversity. This argues for the need of "living" ontologies (actively maintained and highly-responsive to user requests) that can be updated continually and with tools and services to allow users to request new classes and update existing classes with low overhead. Ontology development is extremely time-consuming (Dahdul et al., 2015), and it must be driven by scientific requirements, not by attempts to fully provision them *a priori*. Further, provenance, i.e., the record of authorship involved in term development through persistent digital identifiers such as ORCID (orcid.org), is a poorly developed feature in most ontologies, though important for providing credit to contributors.

Challenges of Interoperability

Data Integration:

Linking environments, locations, and phenotypes will require interoperability between several data types with the varying granularity used in biodiversity and geoscience. These include data types from political and physical geography, coordinate systems, gazetteers, as well as representations of environment and habitat. GeoNames has linked political geography and some physical geography with coordinates (http://www.geonames.org/). A specimen with coordinates can easily be linked to any number of political entities using the GeoNames API. The same has not been accomplished for habitats; however, the components required to accomplish this are falling into place. For example, the LifeMapper (Prajapati, 2009) and Map of Life (Jetz, McPherson & Guralnick, 2012) projects use ecological niche modeling to map species distributions based on environmental conditions. Additionally, the Encyclopedia of Life TraitBank (http://eol.org/info/516) links taxa to their habitat type and phenotypic traits, but not to geographic coordinates (Parr et al., in press). Once greater ontological representation of the link between species and their environments is accomplished, robustly linking species' phenotypes to their environments and locations become readily achievable.

In addition to spatial variation, environments show considerable variation over time and often change over daily and seasonal cycles. This makes temporal data a key component for meaningful integration. Environmental conditions measured at 14:00 can be very different from those measured at 02:00 in the same location. The measurements made at the former, may not apply to a specimen collected at the latter. In addition, an organism is rarely only exposed to conditions measured at a single place and time. Some degree of integration is required to get a complete picture of an environment associated with a phenotype (referred to as the "exposome" in epidemiology, Wild, 2005).

Ontology Legacy Alignment:

The development of successful ontologies is often driven by a "bottom-up" community approach. While this results in a product that is relevant for users, it can also result in multiple partially overlapping ontologies, despite efforts to prevent duplication (e.g., Smith et al., 2007). For improved integration and inferencing, overlapping ontologies need to be properly aligned and those alignments need to be maintained over time. If not done properly, inferencing may be inhibited or precluded altogether. This is a general problem that is not unique to environment or phenotype ontologies (Cregan et al., 2005). A "top-down" approach to ontology development, in which classes that constitute the top levels of a new ontology come from an existing domain or upper-level ontology (e.g., CARO, UBERON, PO, BFO - Grenon & Smith, 2004; Haendel et al., 2008; Mungall et al., 2012; Cooper et al., 2014), can result in a shared structure and homogenized development across ontologies, although more specific classes will still require alignment. Aligning ontologies manually is a large task and it is difficult to know the full consequences of an alignment without testing (Ochs et al., 2015). The ability to support the provenance of alignments and re-alignments can translate into trust and continued investment. Numerous semi-automated tools for alignment have been developed (e.g., Granitzer et al., 2010; Chen et al., 2014). Challenges include setting up proper relations between classes in different ontologies such that the logical outcomes are valid and consistent (Franz & Peet 2009; Meilicke & Stuckenschmidt 2009; Jiménez-Ruiz et al., 2009; Franz et al., 2015, N. Franz unpublished data). The time and effort spent on maintaining alignments and interoperability can be eliminated if shared community resources are instead developed (Dahdul et al., 2015). For example, several independent anatomy ontologies for vertebrates [teleost (Dahdul et al., 2010); amphibian (Maglia, Leopold & Pugener, 2007); vertebrate skeletal (Dahdul et al., 2012), and vertebrate homologous organs (Niknejad et al., 2012)] were recently subsumed into UBERON, the metazoan anatomy ontology (Haendel et al., 2014), and new content and development is now focused on this single resource.

Proof of Concept Demonstrations: Linking Environments and Phenotypes

Miniaturization in Fish

Question: Has the evolution of miniaturization in fishes been driven by environmental variables?

Miniaturization is essentially the evolution of small body size and the associated set of phenotypes, typically reduction or loss of structures. Authors have related this extreme change in body size to organisms whose habitats include highly acidic waters, typical of peat bog or black water habitats (Kottelat et al., 2006). As a proof of concept, we tested the hypothesis that miniaturization is correlated with environmental variables. Using a list of miniaturized fishes and their sister taxa extracted from the literature as input, we retrieved a phenotype X taxon synthetic supermatrix from Phenoscape Knowledgebase (KB) (kb.phenoscape.org) using the Ontotrace tool (Dececchi et al., in press). Using the common phenotype ontologies as a bridge, the KB links evolutionary phenotypes of biodiverse taxa to candidate genes from model organisms. Using the taxon names as input to GBIF, we created a list of 378 georeferenced observations from museum specimen records (http://www.gbif.org/occurrence/download/0000659-150211104439307; Fig. 3). These species' latitude and longitude occurrence records were matched to the 1 km HydroSHEDS hydrography (Lehner, Verdin & Jarvis, 2008) using a horizontal distance tolerance of 3 km; they were then intersected with freshwater specific layers by Domisch et al., (unpublished data). In this data set, the watershed of each 1 km stream reach along the HydroSHEDS hydrography was delineated and then overlaid with climate (Hijmans et al., 2005), topography (Lehner, Verdin & Jarvis, 2008), land cover (Tuanmu & Jetz, 2014), and surface geology (USGS) layers. The differences in the habitat variables between miniatures and non-miniatures were explored using a two-tailed t-test (Table 2). The results showed that miniaturized fishes are found in warmer, wetter environments than their non-miniaturized counterparts. New data layers are being developed to test specific phenotypic hypotheses related to the habitats (e.g., pH, water flow) of miniaturized fishes (Domisch et al., unpublished data). Specifically, the phenotypic data from the ontology-enabled matrix can be used to examine correlations to environment with ontology-based miniaturized phenotypes (e.g., mandibular sensory canal, absent; basibranchial 2 tooth, absent).

Amphibian Reproduction

Question: Which amphibians in my neighborhood are most likely to have their breeding disrupted if a plan to drain a pond (the single source of year-round, standing freshwater) is implemented?

The Encyclopedia of Life links environments associated with a given species' habitat and phenotypes indirectly through taxon names. These data can be accessed and downloaded via TraitBank (Quintero et al., 2014; Parr et al., in press). TraitBank uses Uniform Resource Identifiers (URI), many from existing ontologies, as a controlled vocabulary for describing characters and character states to facilitate large-scale data integration (Table 3). As proof of concept, we queried TraitBank for breeding environment and developmental mode in 282 amphibian taxa. A Chi-Square Test was used to test for independence between habitat and reproductive mode. The data suggested an important reproductive difference between amphibians in aquatic and terrestrial habitats (Table 4). Ninety nine percent of the amphibians with direct development breed in a terrestrial habitat. Ninety eight percent of the amphibians with larval development (tadpoles) breed in an aquatic environment. This links the "larval development" phenotype to the "freshwater" environment and the "direct development" phenotype to the "terrestrial" environment. These data also suggest that a permanent freshwater habitat is more important to amphibians with paedomorphic development than those with indirect development.

These examples provide demonstrations of the value of linking phenotype to environment and demonstrate how these links can be made with existing tools. More complicated research questions are likely to require more nuanced linking for several reasons. First, phenotypes frequently vary within a species; one cannot assume that every member of a species has the same phenotype. In these two examples, we chose traits that were consistent across all members of a species (miniaturization and developmental mode). In the miniaturization example, this allowed addition of the GBIF query results to the Phenoscape Knowledgebase results. Second, an organism's life style (ambush predator, nocturnal frugivore, etc.) within an environment is deeply rooted to its phenotypic composition. For example, a visual predator in an environment with low-light conditions may have a large eye phenotype while a scavenger in the same environment may have a small eye phenotype. Trying to connect an eye size phenotype to this environment would have to be clarified by including the ecological role of the taxon in a given ecosystem. Third, scale can be important. Taxa of very different sizes can experience the same environment in very different ways. For example, a soil protist will experience a forest environment differently than a vascular plant. Despite these challenges, the highly simplified fish and amphibian examples above still demonstrate the results of linking phenotypes and environments with existing data and tools.

Knowledge representation

Despite the challenges of coverage and interoperability, we can demonstrate some basic models linking phenotypes to environments using existing ontologies (Figs. 4,5). The Extensible Observation Ontology (OBOE) provides a basic knowledge graph for linked measurements (Madin et al., 2007; Madin et al., 2008). This ontology has been described in detail elsewhere (Madin et al., 2007, Madin et al., 2008). Briefly, the fundamental OBOE model is built around an "Observation" class which is an observation of an "entity" and has one or more "measurements". Observations can also have a context of other observations. Phenotypes and environments can be linked by representing an organism observation with a location observation as its context (Fig. 4A). OBOE can model categorical and numerical measurements (Fig. 4B).

Thus, a geolocation, a data point, or a country code can be added to a location observation that provides context for an organism observation. OBOE allows the use of literals as instances, meaning a measurement can have as a value a string or a URI, which can be helpful when a needed URI does not exist.

Although OBOE is well suited for describing observations, it was not originally built to manage information about specimens or taxa. The Biological Collections Ontology (BCO) (Walls et al., 2014a; Walls et al., 2014b; Deck et al., 2015) offers an alternative way to link data, based on ontology design principles from the Ontology for Biomedical Investigations (OBI; Brinkman et al., 2010), but adapted for biodiversity science. A key element of BCO is the difference between a specimen collection process, which has a material entity (i.e., specimen) as output and an observing process, which has data as output. Deck at al. (2015) describes how information about locations (e.g., coordinates or environmental conditions) and taxonomy (e.g., the identification process or species name) can be linked to specimens. A similar approach can be used to link phenotypic data to observations of organisms in their environment. At its most basic, the BCO (via OBI) represents the observing process as a type of assay (an OBI class). Rather than representing taxonomic information as an observation, BCO has a class for taxonomic identification process, which, like assay, is a subclass of OBI:planned process (Fig. 5A). Fig. 5B shows how the same data from Figure 4B would be mapped to instances of BCO classes.

OBOE and BCO were developed for different uses cases and therefore have different approaches to representing observations. Nonetheless, there is significant overlap between the two ontologies (e.g., OBOE's observation corresponds closely to BCO's observing process), and ongoing efforts to align them are likely to lead to a harmonized model that can work for many different use cases.

Summary

Providing data structures that improve integration of biological data is necessary for efficiently addressing complex research questions. The link between phenotype and environment is fundamental to research in taxonomy, ecology, and phylogenetics; its relevance extends to the biomedical domain. One way to create this link is through the use of extensible ontologies designed to work across different data types, such as OBOE or BCO in combination with ENVO and other trait ontologies. Despite recent advances, significant challenges remain. We recommend the following steps to increase interoperability between phenotype and environment data:

Make it easy to contribute to existing ontologies.

The existing suite of ontologies is not adequate for linking phenotypes and environments across the tree of life. To address this, new classes need to be added to extend and improve existing taxonomy, phenotype, and envivonmental ontologies. Some ontologies have well-developed pathways for submitting new classes and editing existing classes and resources to respond to requests quickly (e.g., Gene Ontology), but frequently the social processes of validating ontologies are not a part of the ontology platform.

Georeference environments with temporal considerations.

Many taxon observations are accompanied by geographical coordinates, collection date and time, but lack adequate environmental descriptions. While services exist that can translate coordinates into a municipality, retrieving environmental information using geographic coordinates is not yet possible across the globe. In addition, because environments are dynamic, temporal information should be used to filter results. A service is needed that can take spatiotemporal information and return data concerning environmental conditions and ontology classes corresponding to environment types. Map of Life can provide some data corresponding to coordinates in some areas, but ontology classes are not yet available.

Organize research communities that share common resources.

Ontologies rely on community support, driven by scientific questions, to be relevant. Communities of experts can be organized around workshops co-occuring at conferences and funded through programs such as the National Science Foundation’s Research Coordination Network. Significant progress on discipline-specific ontologies has been made through the use of targeted workshops (e.g., Yoder et al., 2010).

Acknowledgments

The scientific meeting from which this review arose was organized by the Phenotype Research Coordination Network, which is funded by the U.S. National Science Foundation, grant number DEB-0956049. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science

Foundation. PLB's work on this project is supported through the Micro B3 project, funded by the European Union's Seventh Framework Programme (Joint Call OCEAN.2011-2: marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential) under the grant agreement no 287589. SD acknowledges funding from the German Research Foundation DFG (grant DO 1880/1-1). PJ and LC acknowledge funding from the U.S. National Science Foundation (NSF IOS:0822201, IOS:1127112, IOS:1340112). CDS acknowledges funding from the U.S. National Science Foundation grant number DEB 1208666. R. Walls was supported by the iPlant collaborative as part of the National Science Foundation Award Numbers DBI-0735191 and DBI-1265383. CJ Mungall and SE Lewis were supported by R24OD011883 and by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. PM Mabee and WM Dahdul were supported through by Phenoscape project (NSF grants DBI-1062404 and DBI-1062542).

Glossary

collecting event – The process of specimen collection that occurs at a specific time and place.

cyberinfrastructure – The technological framework of interconnected databases and computers across institutions that enable and support advanced, large-scale scientific research.

Darwin Core – A standard reference of terms related to biological diversity, in particular taxa and their occurrences. Darwin Core was created to facilitate sharing of biodiversity information.

GCM (General Circulation Model) – From Wikipedia: A general circulation model (GCM), a type of climate model, is a mathematical model of the general circulation of a planetary atmosphere or ocean and based on the Navier–Stokes equations on a rotating sphere with thermodynamic terms for various energy sources (radiation, latent heat).

genotype. The genetic makeup or set of genes of an organism.

georeferenced – Observations or specimen collection records that are associated with locality information (e.g., latitude and longitude).

human readable – Information that is presented in a format that can be understood by a human.

inferencing – Performed by software programs (“reasoners”) that deduce logically consistent statements implied by the entities and relations asserted in an ontology or database.

knowledgebase – A database of interconnected information.

machine readable – Information stored in a data format that can be understood by a computer.

Machine Learning (ML) – A type of artificial intelligence in which software programs have the ability to learn (make decisions or data predictions) without being explicitly programmed when given new data.

meta-analysis – A statistical analysis of data that is combined from independently conducted research studies.

NLP (Natural Language Processing) – Methods used in computer programs to understand and extract data from natural (human) language.

ontology – A set of defined terms (classes, concepts) and the relations between them that represent the knowledge of a particular domain. Terms in an ontology are related in a directed, acyclical graph.

OCR (Optical Character Recognition) - automated conversion of images of text into machine-readable text

OWL (Web Ontology Language) – The name encompassing the set of web-based languages used for ontology building supported by the World Wide Web Consortium (WC3) international standards body and based on the rules of formal semantics.

phenome – The entirety of an organism's phenotypic traits.

phenotype – One or more observable characteristics of an organism.

provenance - History of data and its place of origin.

RDF (Resource Description Framework) – A family of World Wide Web Consortium (WC3) specifications originally designed as a metadata model and generally used to model information in knowledge management applications

semantic – of or relating to meaning or context.

semantic annotation – The act of adding (i.e. 'tagging') information artifacts such as images, free-text anatomical descriptions, or specimen collection records, with classes from an ontology or similar resource which represents their meaning in a machine-readable fashion.

specimen – A whole organism or part of an organism preserved in a collection.

taxonomic description – Natural language description of a taxonomic group, typically includes phenotypic characters such as morphology and behavior.

URI (Uniform Resource Identifier) – A string of characters used to identify a resource that enables interactions with representations of the resource over the internet.

vocabulary – Flat list of terms that can be used to classify data. These terms are not explicitly related to one another.

References

Allen V, Batello C, Berretta E, Hodgson J, Kothmann M, Li X, McIvor J, Milne J, Morris C, Peeters A, Sanderson M, The Forage and Grazing Terminology Committee. 2011. An international terminology for grazing lands and grazing animals. *Grass and Forage Science* 66:2–28. doi: 10.1111/j.1365-2494.2010.00780.x

Ariño A. 2010. Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics* 7:82-92. http://dx.doi.org/10.17161/bi.v7i2.3991

Arnaud E, Cooper L, Shrestha R, Menda N, Nelson RT, Matteis L, Skofic M, Bastow R, Jaiswal P, Mueller L, McLaren G. 2012. Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes. In: KEOD 2012 – Proceedings of the International Conference on Knowledge Engineering and Ontology Development, 220-225. http://wrap.warwick.ac.uk/id/eprint/59831

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene Ontology: Tool for the unification of biology. *Nature Genetics* 25:25-29. doi:10.1038/75556

Balhoff JP, Yoder M, Deans A. 2011. Linking semantic phenotypes to character matrices and specimens. In *Biodiversity Information Standards (TDWG) annual meeting*. New Orleans.

Balhoff JP, Mikó I, Yoder M, Mullins PL, Deans AR. 2013. A semantic model for species description, applied to the ensign wasps (Hymenoptera: Evaniidae) of New Caledonia. *Systematic Biology* 62:639-659. doi: 10.1093/sysbio/syt028

Barber A, Lafferty D, Landrum L. 2013. The SALIX Method: A semi-automated workflow for herbarium specimen digitization. *Taxon* 62:581-590. doi: http://dx.doi.org/10.12705/623.16

Brinkman R, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone SA, Soldatova LN, Stoeckert CJ Jr, Turner JA, Zheng J, The OBI Consortium. 2010. Modeling biomedical experimental processes with OBI. *J. Biomedical Semantics* 1:S7. doi: http://www.jbiomedsem.com/content/1/S1/S7

Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE, ENVO Consortium. 2013. The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics* 4:43. doi:http://www.jbiomedsem.com/content/4/1/43

Ceusters W. 2012. An information artifact ontology perspective on data collections and associated representational artifacts. *Studies in Health Technology and Informatics* 180:68-72.

Chang JT, Schutze H. 2006. Abbreviations in biomedical text. In: *Text Mining for Biology and Biomedicine* 99–119.

Chen M, Yu S, Franz N, Bowers S, Ludäscher B. 2014. Euler/X: A toolkit for logic-based taxonomy integration. arXiv preprint arXiv:1402.1992

Cooper L, Elser JL, Preece J, Arnaud E, Stevenson DW, Todorovic S, Zhang E, Mungall C, Smith B, Jaiswal P. 2014. Common reference ontologies for plant biology (cROP): A platform for integrative plant genomics [abstract no. W819]. *Plant & Animal Genome XXII* Available at: https://pag.confex.com/pag/xxii/webprogram/Paper9799.html

Côté RA, Robboy S. 1980. Progress in medical information management: Systematized Nomenclature of Medicine (SNOMED). *Journal of the American Medical Association* 243:756-762. doi:10.1001/jama.1980.03300340032015

Cregan A, Mochol M, Vrandecic D, Bechhofer S. 2005. Pushing the limits of OWL, rules, and Protégé. In: Grau B, Horrocks I, Parsia B, eds. *Proceedings of the OWLED*05 Workshop on OWL, Experiences and Directions*. CEUR-WS.

Dahdul WM, Balhoff JP, Engeman J, Grande T, Hilton EJ, Kothari CR, Lapp H, Lundberg JG, Midford PE, Vision TJ, Westerfield M, Mabee PM. 2010. Evolutionary characters, phenotypes and ontologies: Curating data from the systematic biology literature. *PloS ONE* 5:e10708. doi: 10.1371/journal.pone.0010708

Dahdul WM, Balhoff JP, Blackburn DC, Diehl AD, Haendel MA, Hall BK, Lapp H, Lundberg JG, Mungall CJ, Ringwald M, Segerdell E, Van Slyke CE, Vickaryous MK, Westerfield M, Mabee PM. 2012. A unified anatomy ontology of the vertebrate skeletal system. *PloS ONE*. 7:e51070. doi: 10.1371/journal.pone.0051070

Dahdul WM, Dececchi TA, Ibrahim N, Lapp H, Mabee PM. 2015. Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy. *Database*. 2015:bav040. doi:10.1093/database/bav040

Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, Cooper LD, Courtot M, Csösz S, Cui H, Dahdul W, Das S, Dececchi TA, Dettai A, Diogo R, Druzinsky RE, Dumontier M, Franz NM, Friedrich F, Gkoutos GV, Haendel M, Harmon LJ, Hayamizu TF, He Y, Hines HM, Ibrahim N, Jackson LM, Jaiswal P, James-Zorn C, Köhler S, Lecointre G, Lapp H, Lawrence CJ, Le Novère N, Lundberg JG, Macklin J, Mast AR, Midford PE, Mikó I, Mungall CJ, Oellrich A, Osumi-Sutherland D, Parkinson H, Ramírez MJ, Richter S, Robinson RN, Ruttenberg A, Schulz KS, Segerdell E, Seltmann KC, Sharkey MJ, Smith AD, Smith B, Specht CD, Squires RB, Thacker RW, Thessen AE, Fernandez-Triana J, Vihinen M, Vize PD, Vogt L, Wall CE, Walls RL, Westerfeld M, Wharton RA, Wirkner CS, Woolley JB, Yoder MJ, Zorn AM, Mabee P. 2015. Finding our way through phenotypes. *PLoS Biology* 13:e1002033. doi:10.1371/journal.pbio.1002033

Dececchi TA, Balhoff JP, Lapp H, Mabee PM. in press. Toward synthesizing our knowledge of morphology: Using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Systematic Biology*. doi:10.1093/sysbio/syv031

Deck J, Guralnick R, Walls R, Blum S, Haendel M, Matsunaga A, Wieczorek J. 2015. Meeting Report: Identifying practical applications of ontologies for biodiversity informatics. *Standards in Genomic Sciences* 10:25. doi: 10.1186/s40793-015-0014-0

DiGiuseppe N, Pouchard L, Noy N. 2014. SWEET ontology coverage for earth system sciences. *Earth Science Informatics* 7:249–264. doi: 10.1007/s12145-013-0143-1

Edwards G. 2004. Revision of the jumping spiders of the genus *Phidippus* (Araneae: Salticidae). *Occasional papers of the Florida State Collection of Arthropods* 11:1–156.

Edwards G. 1980. Taxonomy, ethology, and ecology of *Phidippus* (Araneae: Salticidae) in eastern North America. D. Phil. Thesis, University of Florida. Available at: https://archive.org/details/taxonomyethology00edwa.

Edwards P, Mayernik M, Batcheller A, Bowker GC, Borgman CL. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science*. 41:667–690. doi:0.1177/0306312711413314

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology 6:R44. doi:10.1186/gb-2005-6-5-r44

Enke N, Thessen AE, Bach K, Bendix J, Seeger B, Gemeinholzer B. 2012. The user’s view on biodiversity data sharing. *Ecological Informatics* 11:25-33. http://dx.doi.org/10.1016/j.ecoinf.2012.03.004

Field D, Amaral-Zettler L, Cochrane G, Cole JR, Dawyndt P, Garrity G, Gilbert J, Glöckner FO, Hirschman L, Karsch-Mizrachi I, Klenk HP, Knight R, Kottmann R, Kyrpides N, Meyer F, Gil IS, Sansone SA, Schriml LM, Sterk P, Tatusova T, Ussery DW, White O, Wooley J. 2011. The Genomic Standards Consortium. *PLoS Biology* 9:e1001088. doi:10.1371/journal.pbio.1001088

Franz N, Peet R. 2009. Towards a language for mapping relationships among taxonomic concepts. *Systematics and Biodiversity* 7:5-20. doi: 10.1017/S147720000800282X

Franz NM, Chen M, Yu S, Kianmajd P, Bowers S, Ludäscher. 2015. Reasoning over taxonomic change: Exploring alignments for the *Perelleschus* use case. *PLoS ONE* 10:e0118247. doi:10.1371/journal.pone.0118247

Giglio M, Mungall C, Uetz P, Yin L, Goll J, Siegele D, Chibucos M, Hu J. 2009. Development of an Ontology of Microbial Phenotypes (OMP). *Nature Precedings*. doi:10.1038/npre.2009.3639.1

Giles JRA. 2011. Geoscience metadata - no pain, no gain. In: Sinha AK, Arctur D, Jackson eds. *Societal Challenges and Geoinformatics*. Geological Society of America, 29–33.

Gkoutos G, Green E, Mallon A, Hancock JM, Davidson D. 2004. Using ontologies to describe mouse phenotypes. *Genome Biology* 6:R8. doi:10.1186/gb-2004-6-1-r8

Gliozzo A, Biran O, Patwardhan S, McKeown K. 2013. Semantic Technologies in IBM Watson. In: *Proceedings of the Fourth Workshop on Teaching NLP and CL*. Sofia, Bulgaria: Association for Computational Linguistics, 85–92. Available at: http://www.aclweb.org/anthology/W13-3413.

Granitzer M, Sabol V, Onn K, Lukose D, Tochtermann K. 2010. Ontology Alignment—A Survey with Focus on Visually Supported Semi-Automatic Techniques. *Future Internet* 2:238–258. doi:10.3390/fi2030238

Grenon P, Smith B. 2004. SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition & Computation: An Interdisciplinary Journal* 4:69-104. doi:10.1207/s15427633scc0401_5

Haendel M, Neuhaus F, Osumi-Sutherland D, Mabee PM, Mejino JLV Jr, Mungall CJ, Smith B. 2008. CARO–the common anatomy reference ontology. In: *Anatomy Ontologies for Bioinformatics*. London: Springer, 327–349.

Haendel M, Balhoff JP, Bastian F, Blackburn DC, Blake JA, Bradford Y, Comte A, Dahdul WM, Dececchi TH, Druzinsky RE, Hayamizu TF, Ibrahim N, Lewis SE, Mabee PM, Niknejad A, Robinson-Rechavi M, Sereno PC, Mungall CJ. 2014. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *Journal of Biomedical Semantics* 5:21. doi:10.1186/2041-1480-5-21

Harris MA, Lock A, Bähler J, Oliver SG, Wood V. 2013. FYPO: The fission yeast phenotype ontology. Bioinformatics 29:1671-1678. doi:10.1093/bioinformatics/btt266

Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C. 2013. The ChEBI reference and ontology for biologically relevant chemistry: Enhancements for 2013. Nucleic Acids Research 41:D456-D463. doi:10.1093/nar/gks1146

Heidorn PB. 2008. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends* 57:280–299.

Henderson M, Khan I, Hunter J. 2007. Semantic WildNET: An ontology based biogeographical system, Available at: http://www.itee.uq.edu.au/eresearch/filething/files/get/projects/ecoportalqld/SemWildNET.pdf.

Hey T, Tansley S, Tolle K. 2009. *The fourth paradigm: data-intensive scientific discovery*, Seattle: Microsoft Research.

Hijmans R, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*. 25:1965–1978.

Hill D. 2014. Notes on the jumping spider *Phidippus clarus* Keyserling 1885 (Araneae: Salticidae: Dendryphantinae). *Peckhamia* 113.1:1–32.

Hoefler C, Chen A, Jakob E. 2006. The potential of a jumping spider, *Phidippus clarus*, as a biocontrol agent. *Journal of Economic Entomology.* 99:432–436. doi:http://dx.doi.org/10.1603/0022-0493-99.2.432

Huang F, Macklin J, Cui H, Cole HA, Endara L. 2015. OTO: Ontology Term Organizer. *BMC Bioinformatics* 16:47. doi:10.1186/s12859-015-0488-1

Hunt LA, White JW, Hoogenboom G. 2001. Agronomic data: advances in documentation and protocols for exchange and use. *Agricultural Systems* 70:477-492.

Jain M, Balakrishnan R. 2011. Microhabitat selection in an assemblage of crickets (Orthoptera: Ensifera) of a tropical evergreen forest in Southern India. *Insect Conservation and Diversity* 4:152–158.

Jaiswal P, Ware D, Ni J, Chang K, Zhao W, Schmidt S, Pan X, Clark K, Teytelman L, Cartinhour S, Stein L, McCouch S. 2002. Gramene: Development and integration of trait and gene ontologies for rice. *Comparative and Functional Genomics* 3:132-136. doi:10.1002/cfg.156

Jaiswal P, Avraham S, Ilic K, Kellogg EA, McCouch S, Pujar A, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Ware D, Zapata F. 2005. Plant Ontology (PO): A controlled vocabulary of plant structures and growth stages. Comparative Functional Genomics 6:388-397. doi:10.1002/cfg.496

Jetz W, McPherson J, Guralnick R. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology & Evolution* 27:151–159. doi:10.1016/j.tree.2011.09.007

Jiménez-Ruiz E, Grau B, Horrocks I, Berlanga R. 2009. Ontology Integration Using Mappings: Towards Getting the Right Logical Consequences. In: *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg, 173–187.

Kattge J, Diaz S, Lavorel S, Prentice C, Leadley P, Bonisch G, Garnier E, Westoby M, Reich PB, Wright IJ, Cornelissen JHC, Violle C, Harrison SP, van Bodegom PM, Reichstein M, Enquist BJ, Soudzilovskaia NA, Ackerly DD, Anand M, Atkin O, Bahn M, Baker TR, Baldocchi D, Bekker R, Blanco CC, Blonder B, Bond WJ, Bradstock R, Bunker DE, Casanoves F, Cavender-Bares J, Chambers JQ, Chapin FS, Chave J, Coomes D, Cornwell WK, Craine JM, Dobrin BH, Duarte L, Durka W, Elser J, Esser G, Estiarte M, Fagan WF, Fang J, Fernandez-Mendez F, Fidelis A, Finegan B, Flores O, Ford H, Frank D, Freschet GT, Fyllas NM, Gallagher RV, Green WA, Gutierrez AG, Hickler T, Higgins SI, Hodgson JG, Jalili A, Jansen S, Joly CA, Kerkhoff AJ, Kirkup D, Kitajima K, Kleyer M, Klotz S, Knops JMH, Kramer K, Kuhn I, Kurokawa H, Laughlin D, Lee TD, Leishman M, Lens F, Lenz T, Lewis SL, Lloyd J, Llusia J, Louault F, Ma S, Mahecha MD, Manning P, Massad T, Medlyn BE, Messier J, Moles AT, Muller SC, Nadrowski K, Naeem S, Niinemets U, Nollert S, Nuske A, Ogaya R, Oleksyn J, Onipchenko VG, Onoda Y, Ordonez J, Overbeck G, Ozinga WA, Patino S, Paula S, Pausas JG, Penuelas J, Phillips OL, Pillar V, Poorter H, Poorter L, Poschlod P, Prinzing A, Proulx R, Rammig A, Reinsch S, Reu B, Sack L, Salgado-Negre B, Sardans J, Shiodera S, Shipley B, Siefert A, Sosinski E, Soussana JF, Swaine E, Swenson N, Thompson K, Thornton P, Waldram M, Weiher E, White M, White S, Wright SJ, Yguel B, Zaehle S, Zanne AE, Wirth C. 2011. TRY - a global database of plant traits. *Global Change Biology* 17:2905-2935.

Keyserling E. 1884-1885. Neue Spinnen aus Amerika. VI., Verhandlungen der k.k. zoologisch-botanischen Gesellschaft in Wien, Wien: 497. Available at: http://www.biodiversitylibrary.org/part/6448

Kottelat M, Britz R, Hui T, Witte KE. 2006. *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual dimorphism, comprises the world's smallest vertebrate. *Proceedings of the Royal Society B* 273:895-899.

Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jähn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park SM, Riggs ER, Scott RH, Sisodiya S, Van Vooren S, Wapner RJ, Wilkie AO, Wright CF, Vulto-van Silfout AT, de Leeuw N, de Vries BB, Washington NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN. 2014. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research* 42:D966-D974. doi:10.1093/nar/gkt1026

Lehner B, Verdin K, Jarvis A. 2008. New global hydrography derived from spaceborne elevation data. *EOS, Transactions American Geophysical Union*. 89:93–94.

Mabee PM, Ashburner M, Cronk Q, Gkoutos GV, Haendel M, Segerdell E, Mungall C, Westerfield M. 2007. Phenotype ontologies: the bridge between genomics and evolution. *Trends in Ecology and Evolution* 22:345-350. doi:10.1016/j.tree.2007.03.013

Maddison W, Maddison D. 2014. Mesquite: a modular system for evolutionary analysis. Available at: http://mesquiteproject.org.

Madin JS, Bowers S, Schildhauer MP, Krivov S, Pennington D, Villa F. 2007. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics* 2:279–296. doi: 10.1016/j.ecoinf.2007.05.004

Madin JS, Bowers S, Schildhauer MP, Jones MB. 2008. Advancing ecological research with ontologies. *Trends in Ecology & Evolution* 23:159-168.

Maglia A, Leopold J, Pugener L. 2007. An anatomical ontology for amphibians. In: Altman R, Dunker A, Hunter L. eds. *Pacific Symposium on Biocomputing 2007*. World Scientific, 367–378.

Malone J, Holloway E, Adamsusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. 2010. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26:1112–1118. doi: 10.1093/bioinformatics/btq099

Meilicke C, Stuckenschmidt H. 2009. An Efficient Method for Computing Alignment Diagnoses. In: *Web Reasoning and Rule Systems*. Springer Berlin Heidelberg, 182–196.

Mullins P, Kawada R, Balhoff JP, Deans AR. 2012. A revision of *Evaniscus* (Hymenoptera, Evaniidae) using ontology-based semantic phenotype annotation. *ZooKeys* 223:1–38. doi:10.3897/zookeys.223.3572

Mungall C, Torniai C, Gkoutos G, Lewis SE, Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome Biology* 13:R5. doi: 10.1186/gb-2012-13-1-r5

Niknejad A, Comte A, Parmentier G, Roux J, Bastian FB, Robinson-Rechavi M. 2012. vHOG, a multispecies vertebrate ontology of homologous organs groups. *Bioinformatics*. 28:1017–1020. doi: 10.1093/bioinformatics/bts048

Ochs C, Perl Y, Geller J, Haendel MA, Brush M, Arabandi S, Tu S. 2015. Summarizing and visualizing structural changes during the evolution of biomedical ontologies using a Diff Abstraction Network. *Journal of Biomedical Informatics* 56:127-144. doi:10.1016/j.jbi.2015.05.018

Oellrich A, Walls RL, Cannon EK, Cannon SB, Cooper L, Gardiner J, Gkoutos GV, Harper L, He M, Hoehndorf R, Jaiswal P, Kalberer SR, Lloyd JP, Meinke D, Menda N, Moore L, Nelson RT, Pujar A, Lawrence CJ, Huala E. 2015. An ontology approach to comparative phenomics in plants. *Plant Methods* 11:10. doi: 10.1186/s13007-015-0053-y

Pafilis E, Frankild SP, Schnetzer J, Fanini L, Faulwetter S, Pavloudi C, Vasileiadou K, Leary P, Hammock J, Schulz K, Parr CS, Arvanitidis C, Jensen LJ. 2015. ENVIRONMENTS and EOL: Identification of Environment Ontology terms in text and the annotation of the Encyclopedia of Life. *Bioinformatics* 31:1872-1874. doi:10.1093/bioinformatics/btv045

Page RDM. 2008. Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Briefings in Bioinformatics* 9:345–354.

Parr C, Wilson N, Schulz K, Leary P, Hammock J, Rice J, Corrigan RJ Jr. in press. TraitBank: Practical semantics for organism attribute data. *Semantic Web*. Available at: http://www.semantic-web-journal.net/system/files/swj650.pdf.

Pesquita C, Faria D, Falcao A, Lord P, Couto FM. 2009. Semantic similarity in biomedical ontologies. *PLoS Computational Biology* 5:e1000443.

Poelen JH, Simons JD, Mungall CJ. 2014. Global Biotic Interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics* 24:148–159. doi:10.1016/j.ecoinf.2014.08.005

Prajapati V. 2009. LIFEMAPPER: Mapping and Predicting the Distribution of Life with Distributed Computation: The Future of Biodiversity. *Archives of Applied Science Research* 1:306–312.

Quintero E, Thessen AE, Arias-Caballero P, Ayala-Orozco B. 2014. A statistical assessment of population trends for data deficient Mexican amphibians. *PeerJ* 2:e703. doi:https://dx.doi.org/10.7717/peerj.703

Ramírez MJ, Michalik P. 2014. Calculating structural complexity in phylogenies using ancestral ontologies. *Cladistics* 30:635-649. doi:10.1111/cla.12075

Reed B, White M, Brown J. 2003. Remote sensing phenology. In: Schwartz M, ed. *Phenology: An Integrative Environmental Science*. Springer Netherlands, 365–381.

Resnik P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* 11:95–130.

Sasaki N, Putz FE. 2009. Critical need for new definitions of “forest” and “forest degradation” in global climate change agreements. *Conservation Letters* 2:226-232. doi:10.1111/j.1755-263X.2009.00067.x

Schindelman G, Fernandes JS, Bastiani CA, Yook K, Sternberg PW. 2011. Worm Phenotype Ontology: Integrating phenotype data within and beyond the *C. elegans* community. B*MC Bioinformatics* 12:32. doi:10.1186/1471-2105-12-32

Schuh R. 2012. Integrating specimen databases and revisionary systematics. *ZooKeys* 209:255–267.

Seltmann K, Pénzes Z, Yoder M, Bertone MA, Deans AR. 2013. Utilizing descriptive statements from the Biodiversity Heritage Library to expand the Hymenoptera Anatomy Ontology. *PLoS ONE* 8:e55674. doi: 10.1371/journal.pone.0055674

Seppälä S, Smith B, Ceusters W. 2014. Applying the realism-based ontology versioning method for tracking changes in the Basic Formal Ontology. In: Garbacz P, Kutz O, eds. *Formal Ontology in Information Systems*. IOS Press, 227-240.

Shrestha R, Arnaud E, Mauleon R, Davenport GF, Hancock D, Morrision N, Bruskiewich R, McLaren G. 2010. Multifunctional crop trait ontology for breeders' data: Field book, annotation, data discovery and semantic enrichment of the literature. *AoB Plants* 2010:plq008. doi: 10.1093/aobpla/plq008

Smith CL, Eppig JT. 2009. The Mammalian Phenotype Ontology: Enabling robust annotation and comparative analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 1:390-399. doi:10.1002/wsbm.44

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall C, The OBI Consortium, Leontis N, Rocca-Serra P, Ruttenber A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S. 2007. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25:1251-1255. doi:10.1038/nbt1346

Spackman KA, Campbell KE, Côté RA. 1997. SNOMED RT: A reference terminology for health care. *Proceedings of the AMIA Annual Fall Symposium* 1997:640-644.

Takhtajan A. 1986. *Floristic Regions of the World*, University of California Press.

Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T, Brazma A, Brinkman RR, Clark AM, Deutsch EW, Fiehn O, Fostel J, Ghazal P, Gibson F, Gray T, Grimes G, Hancock JM, Hardy NW, Hermjakob H, Julian RK, Kane M, Kettner C, Kinsinger C, Kolker E, Kuiper M, Le Novere N, Leebens-Mack J, Lewis SE, Lord P, Mallon AM, Marthandan N, Masuya H, McNally R, Mehrle A, Morrison N, Orchard S, Quackenbush J, Reecy JM, Robertson DG, Rocca-Serra P, Rodriguez H, Rosenfelder H, Santoyo-Lopez J, Scheuermann RH, Schober D, Smith B, Snape J, Stoeckert CJ, Tipton K, Sterk P, Untergasser A, Vandesompele J, Wiemann S. 2008. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology* 26:889–896. doi:10.1038/nbt0808-889

Thacker RW, Díaz MC, Kerner A, Vignes-Lebbe R, Segerdell E, Haendel MA, Mungall CJ. 2013. The Porifera Ontology (PORO): enhancing sponge systematics with an anatomy ontology. *Journal of Biomedical Systematics* 5:39. doi:10.1186/2041-1480-5-39

Thormann I, Parra-Quijano M, Endresen D, Rubio-Teso ML, Iriondo MJ, Maxted N. 2014. Predictive characterization of crop wild relatives and landraces. Biodiversity International. Available at http://www.bioversityinternational.org/e-library/publications/

Tuanmu M, Jetz W. 2014. A global 1km consensus land cover product for biodiversity and ecosystem modelling. *Global Ecology and Biogeography* 23:1031–1045.

USGS International Surface Geology http://certmapper.cr.usgs.gov/data/envision/index.html?widgets=geologymaps

Vos RA, Biserkov J, Balech B, Beard N, Blissett M, Brenninkmeijer C, van Dooren T, Eades D, Gosline G, Groom QJ, Hamann TD, Hettling H, Hoehndorf R, Helleman A, Hovenkamp P, Kelbert P, King D, Kirkup D, Lammers Y, DeMeulemeester T, Mietchen D, Miller JA, Mounce R, Nicolson N, Page R, Pawlik A, Pereira S, Penev L, Richards K, Sautter G, Shorthouse DP, Tahtinen M, Weiland C, Williams AR, Sierra S. 2014. Enriched biodiversity data as a resource and service. *Biodiversity Data Journal* 2:e1125. doi:10.3897/BDJ.2.e1125

Vollmar A, Macklin JA, Ford L. 2010. Natural history specimen digitization: challenges and concerns. *Biodiversity Informatics* 7:93-112. doi: http://dx.doi.org/10.17161/bi.v7i2.3992

W3C OWL Working Group, 2012. *OWL 2 Web Ontology Language*, Available at: http://www.w3.org/TR/2012/REC-owl2-overview-20121211.

Walls R, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, Bowers S, Buttigieg PL, Davies N, Endresen D, Gandolfo MA, Hanner R, Janning A, Krishtalka L, Matsunaga A, Midford P, Morrison N, Tuama EO, Schildhauer M, Smith B, Stucky BJ, Thomer A, Wieczorek J, Whitacre J, Wooley J. 2014a. Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS ONE* 9:e89606.

Walls R, Guralnick R, Deck J, Buntzman A, Buttigieg PL, Davies N, Denslow MW, Gallery RE, Parnell JJ, Osumi-Sutherland D, Robbins RJ, Rocca-Serra P, Wieczorek J, Zheng J. 2014b.

Meeting report: Advancing practical applications of biodiversity ontologies. *Standards in Genomic Sciences* 9:17. doi:10.1186/1944-3277-9-17

White J, Hunt L, Boote K, Jones JW, Koo J, Kim S, Porter CH, Wilkens PW, Hoogenboom G. 2013. Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. *Computers and Electronics in Agriculture* 96:1–12. doi:10.1016/j.compag.2013.04.003

Wild C. 2005. Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology, Biomarkers & Prevention* 14:1847. doi: 10.1158/1055-9965.EPI-05-0456

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD. 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* 29:415-420. doi:10.1038/nbt.1823

Yoder MJ, Mikó I, Seltmann KC, Bertone MA, Deans AR. 2010. A Gross Anatomy Ontology for Hymenoptera. *PLoS ONE* 5:e15991. doi: 10.1371/journal.pone.0015991

Zhan J, Chen Y. 2011. Research on Word Sense Disambiguation. *Advanced Materials Research* 181-182:337–342.

Table 1: List of resources (vocabularies and ontologies) relevant to annotating phenotypes and environments.

Name	Abbreviation	URL	Reference
AGROVOC		1	
Behavioral Ontology	NBO	2	
Chemical Entities of Biological Interest	ChEBI	3	Hastings et al., 2013
CMECS Habitat Classification		4	
Crop Ontology	CO	5	Shrestha et al., 2010
Eagle-i Resource Ontology	ERO	6	
EcoLexicon		7	
Ecological Classifications NatureServe		8	
Environment Ontology	ENVO	9	Buttigieg et al., 2013
EUNIS Habitat Classification		10	
Experimental Factor Ontology	EFO	11	Malone et al., 2010
Exposure ontology	EXO	12	
Fission Yeast Phenotype Ontology	FYPO	13	Harris et al., 2013
Flora Phenotype Ontology	FLOPO	14	Vos et al., 2014
Floristic Regions of the World			Takhtajan, 1986
Fungal gross anatomy	FAO	15	
Gazetteer	GAZ	16	
Gene Ontology	GO	17	Ashburner et al., 2000
GeoNames		18	
Getty Thesaurus of Geographic Names		19	
Global Administrative Areas	GADM	20	
Human Phenotype Ontology	HP	21	Köhler et al., 2014
Information Artifact Ontology	IAO	22	Ceusters, 2012
International Consortium for Agricultural Systems Applications standards	ICASA		White et al., 2013
IUCN Habitats Classification Scheme		23	
Mammalian phenotype	MP	24	Smith and Eppig, 2009
Mapping European Seabed Habitats	MESH	25	
NASA GCMD keyword list for locations		26	
Ontology of Biological Attributes	OBA	27	
Ontology of Biomedical Investigation	OBI	28	Brinkman et al., 2010
Ontology of Microbial Phenotypes	OMP	29	Giglio et al., 2009
Phenotype Quality Ontology	PATO	30	Gkoutos et al., 2004
Plant Environment Ontology	EO	31	
Plant Ontology	PO	32	Jaiswal et al., 2005
Plant Trait Ontology	TO	33	Jaiswal et al., 2002 Arnaud et al., 2012
Population and Community Ontology	PCO	34	
Relation Ontology	RO	35	
Semantic Web for Earth and Environmental Terminology	SWEET	36	DiGiuseppe et al., 2014
Sequence Ontology	SO	37	Elbeck et al., 2005

Terminology of Grazing Lands and Grazing Animals			Allen et al., 2011
Uber Anatomy Ontology	UBERON	38	Mungall et al., 2012; Haendel et al., 2014
Worm Phenotype	WBPhenotype	39	Schindelman et al., 2011
WWF Ecozones		40	

1 http://aims.fao.org/agrovoc#.VG4QG_nF_ng
2 https://code.google.com/p/behavior-ontology/
3 https://www.ebi.ac.uk/chebi/
4 https://marinemetadata.org/references/cmecshabitat
5 http://pantheon.generationcp.org/index.php?option=com_content&task=section&id=7&Itemid=35
6 https://www.eagle-i.net/
7 http://ecolexicon.ugr.es/en/index.htm
8 http://explorer.natureserve.org/classeco.htm
9 http://www.environmentontology.org
10 https://marinemetadata.org/references/eunishabitat
11 http://www.ebi.ac.uk/efo/
12 http://www.obofoundry.org/cgi-bin/detail.cgi?id=exo
13 http://www.pombase.org/
14 http://wiki.pro-ibiosphere.eu/wiki/Traits_Task_Group
15 http://www.yeastgenome.org/fungi/fungal_anatomy_ontology/
16 http://bioportal.bioontology.org/ontologies/GAZ
17 http://geneontology.org/
18 http://www.geonames.org/
19 http://www.getty.edu/research/tools/vocabularies/tgn/index.html
20 http://www.gadm.org/
21 http://www.human-phenotype-ontology.org/
22 https://code.google.com/p/information-artifact-ontology/
23 http://www.iucnredlist.org/technical-documents/classification-schemes/habitats-classification-scheme-ver3
24 http://www.informatics.jax.org/searches/MP_form.shtml
25 http://www.emodnet-seabedhabitats.eu/
26 https://marinemetadata.org/references/cfregions
27 http://wiki.geneontology.org/index.php/Extensions/x-attribute
28 http://obi-ontology.org/page/Main_Page
29 http://microbialphenotypes.org/wiki/index.php/Main_Page
30 http://obofoundry.org/wiki/index.php/PATO:Main_Page
31 http://planteome.org/amigo/cgi-bin/crop_amigo/browse.cgi?
32 http://www.plantontology.org/
33 http://planteome.org/amigo/cgi-bin/crop_amigo/browse.cgi?
34 https://github.com/PopulationAndCommunityOntology/pco
35 https://github.com/oborel/obo-relations
36 https://sweet.jpl.nasa.gov/
37 http://www.sequenceontology.org/
38 http://uberon.github.io/
39 http://www.wormbase.org/
40 http://wwf.panda.org/about_our_earth/ecoregions/ecoregion_list/

Table 2: Mean annual temperature and precipitation associated with miniature and non-miniature phenotypes in fishes

Variable	Type	Mean	p value	df	t Statistic	t Critical
Annual Mean Temperature (°C)	Miniature	24.8°C	0.002	227	3.128	1.970
	Non-miniature	22.6°C				
Annual Mean Precipitation (mm)	Miniature	6.9×10^7 mm	0.008	227	2.668	1.970
	Non-miniature	1.8×10^7 mm				

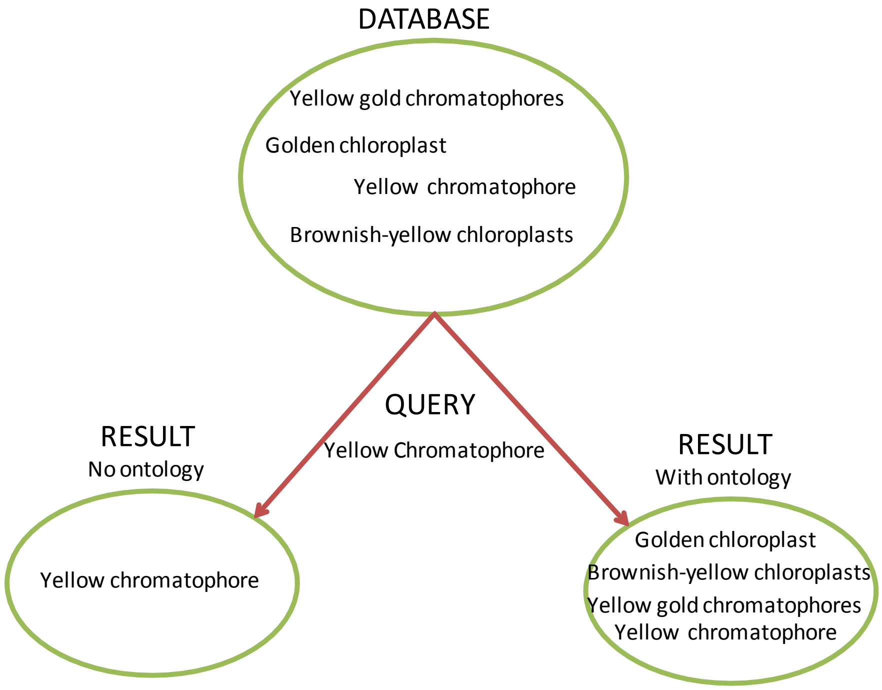
Table 3: Some of the URIs used to describe amphibian breeding and development in TraitBank

Term	URI
breeding habitat	http://eol.org/schema/terms/BreedingHabitat
development mode	http://eol.org/schema/terms/DevelopmentalMode
terrestrial habitat	http://purl.obolibrary.org/obo/ENVO_00002009
intermittent pond	http://purl.obolibrary.org/obo/ENVO_00000504
permanent pond	http://eol.org/schema/terms/permanentFreshwater
freshwater stream	http://eol.org/schema/terms/freshwaterStream
direct development	http://eol.org/schema/terms/directDeveloper
larval development	http://eol.org/schema/terms/larvalDevelopment
paedomorphic	http://purl.obolibrary.org/obo/HOM_0000029

Table 4: Breeding habitat and developmental mode for 282 species of amphibians

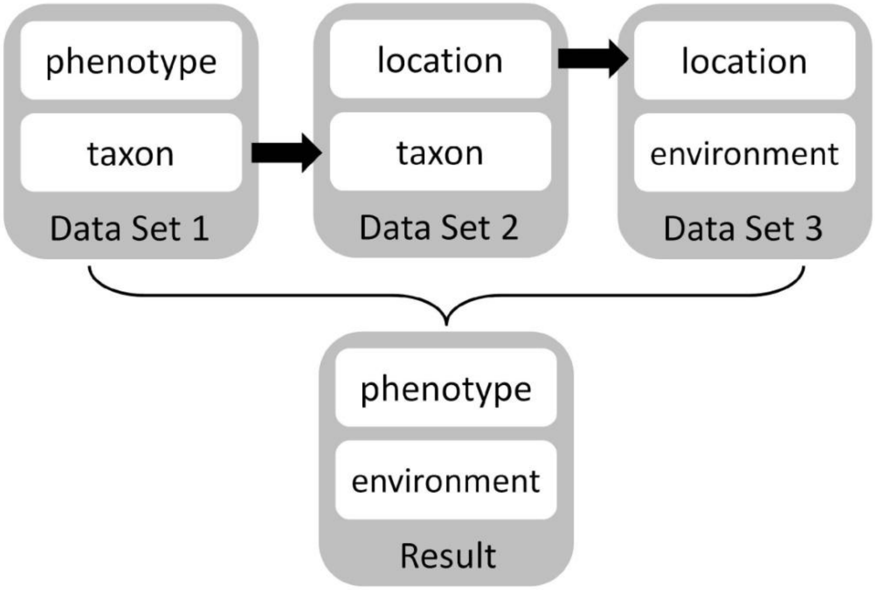
	Larval	Direct	Paedomorphic	df	test statistic	$X^2_{0.95}$
Freshwater Stream	30	0	1	6	278	1.635
Intermittent Pond	28	0	0			
Permanent Pond	59	2	3			
Terrestrial	2	166	0			

Figure 1: Ontology and the Heterogeneity Challenge



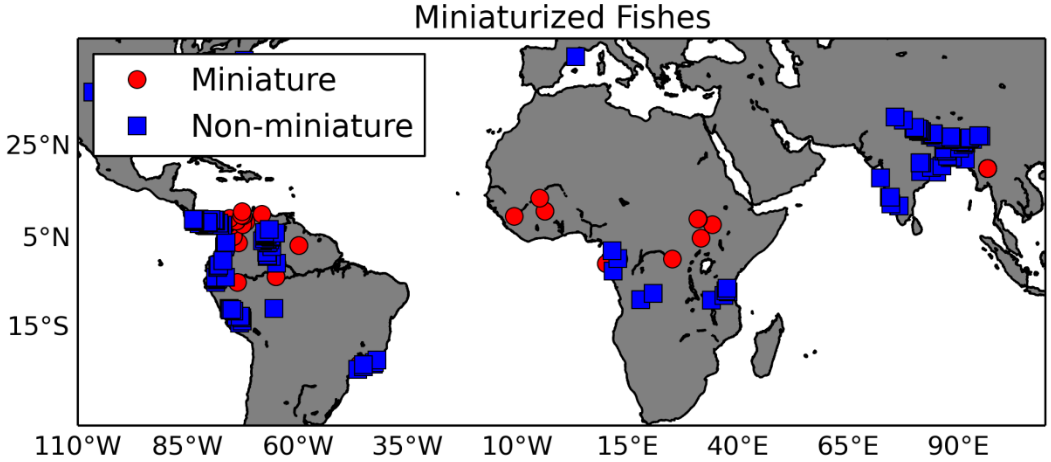
This diagram demonstrates how ontologies can solve the challenge of heterogeneous terminology. In this example, the database contains four different natural language descriptions about dinoflagellate chloroplasts harvested from text. A user needs to query the database for instances of dinoflagellates with yellow chromatophores. Without an ontology to provide the query engine information about synonomy ("chromatophore" = "chloroplast") and term relationships ("brownish-yellow", "golden", and "yellow gold" are subtypes of "yellow"), a query for "yellow chromatophore" will only yield one of the four results the user needs and would find using an ontology. Without an ontology to link closely related concepts with a common parent, and reconcile heterogeneous terms, a user would have to perform many more queries to get a desired result, which may not be tractable in a large dataset.

Figure 2: Manual Workflow Conceptual Diagram



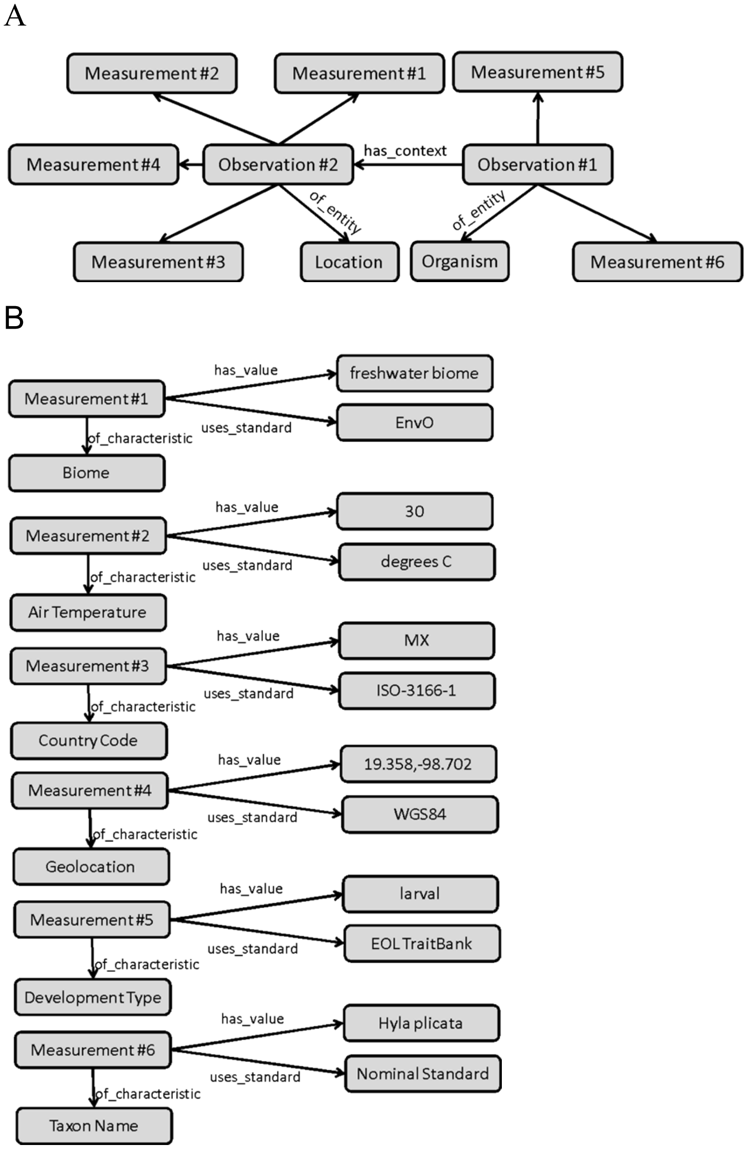
This diagram shows the manual workflow to link phenotype and environment data sets using current tools and services.

Figure 3: Map of Miniaturized Fishes and Their Non-Miniaturized Sister Taxa



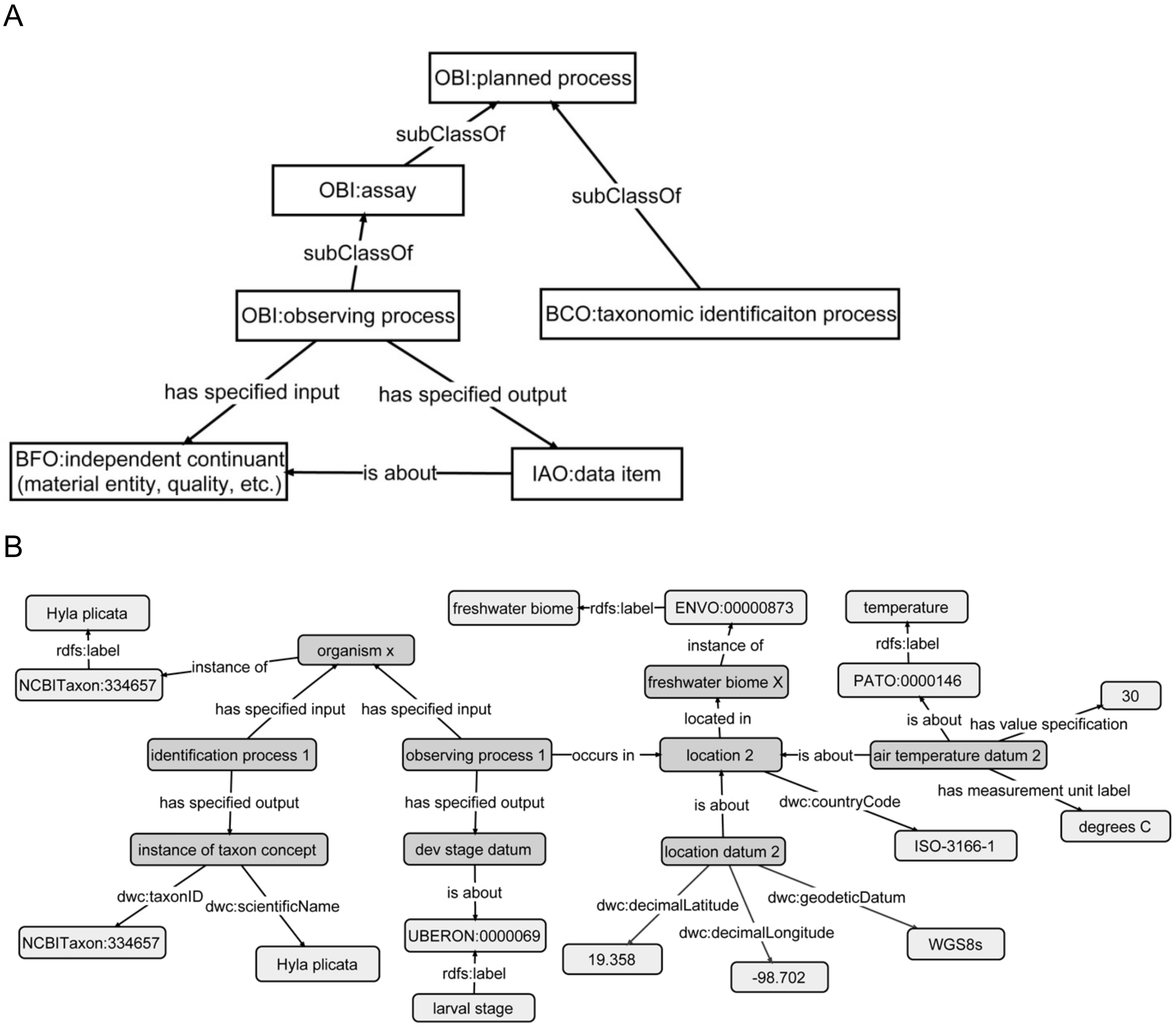
This map shows locations of fish species exhibiting the miniaturized phenotype (red circles) and their non-miniature sister taxa (blue squares). The georeferenced occurrence data were gathered from GBIF.

Figure 4: Using the OBOE Ontology to Link Phenotype and Environment



This demonstrates linking phenotype and environment using instances of the OBOE classes Entity, Observation, and Measurement. A) Links between Entity, Observation, and Measurement OBOE classes. B) Example measurements of phenotypes and environments using instances of the OBOE classes. Numbered measurement instances are consistent across A and B. This representation is simplified with regards to the taxonomic entities in play (Baskauf and Webb unpublished data).

Figure 5: Using BCO to Link Phenotype and Environment



This demonstrates linking phenotype and environment using classes and relations from the Biological Collections Ontology (BCO). A) A simple version of the classes and relations used to describe observations in the BCO, with classes imported from OBI (Ontology for Biomedical Investigations), IAO (Information Artifact Ontology), and BFO (Basic Formal Ontology). B) Links among organism, phenotype, and environment, using the BCO model, using the same data as in Fig. 3. Light grey boxes represent either literal values (e.g., Hyla plicata), or instances of classes from external ontologies (ENVO – Environment Ontology, UBERON – Uber Anatomy Ontology, PATO – Phenotye Quality Ontology). Properties with a dwc prefix are imported directly from Darwin Core.