BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25698
doi: 10.3897/biss.2.25698

OPEN ACCESS

Conference Abstract

Phenomap - Challenges and Successes in Bringing Together Multiple Data Projects to Build New Visualizations of Phenotypic Information and Specimen Records

Matthew Collins[‡,§], Rebecca Tarvin[|], Martha Kandziora[¶], Wasila Dahdul[#], Deborah Paul[¤]

‡ University of Florida, Gainesville, United States of America
§ iDigBio, Gainesville, United States of America
| University of Texas at Austin, Austin, United States of America
¶ University of California, Merced, Merced, United States of America
# University of South Dakota, Vermillion, United States of America
¤ Florida State University, Tallahassee, United States of America

Corresponding author: Matthew Collins (mcollins@acis.ufl.edu)

Received: 11 Apr 2018 | Published: 21 May 2018

Citation: Collins M, Tarvin R, Kandziora M, Dahdul W, Paul D (2018) Phenomap - Challenges and Successes in Bringing Together Multiple Data Projects to Build New Visualizations of Phenotypic Information and Specimen Records. Biodiversity Information Science and Standards 2: e25698. https://doi.org/10.3897/biss.2.25698

Abstract

Connecting biodiversity data across databases is not as easy as one might think. Different databases use different identifiers and taxonomies and connecting these data often results in loss of information and precision. Here we present some of the challenges we faced with integrating multiple biodiversity data sets, including specimen data from the scientific collections, during a hackathon hosted by the Phenoscape project in December of 2017. The hackathon brought together a diverse group of participants, including biologists and software developers, to explore ways of using the computable phenotype data in the Phenoscape Knowledgebase (KB) (Edmunds et al. 2015). The KB contains ontology-annotated data that links evolutionary phenotypes from the comparative literature to model organism phenotypes enabling, e.g., the retrieval of candidate genes for evolutionary phenotypes and the generation of synthetic supermatrices of presence/absence characters. During this hackathon, our team explored how to link phenotype data in the KB to museum specimen

data in iDigBio (Matsunaga et al. 2013) with the hope of creating visualizations including world maps showing species distributions with different character states and their phylogenetic relationships. We visualized lineage relationships by querying the Open Tree of Life (OT) (Hinchliff et al. 2015) website using data integrated by another group at the hackathon that linked KB and OT taxonomic identifiers.

Phenoscape uses terms from anatomy, quality, and taxonomy ontologies to annotate characters and taxonomic information from the phylogenetic literature along with specimen information. When populating the KB, specimen identifiers such as occurrence identifiers, collector's number, and catalog numbers were preserved if present in the literature. We found that these identifiers, although standard in the biodiversity domain, were mostly insufficient to uniquely identify the source specimen in iDigBio. As an alternative, we instead mapped all the occurrences of taxa using string matches of the genus and species from Vertebrate Taxonomy Ontology identifiers. Without specimen identifiers that are consistent across databases, we lost the ability to explore spatial and temporal variation of characters within genera and were only able to explore phenotypes and geographic distributions among genera. We look forward to discussing these issues with the collections community represented at this meeting by the Society for the Preservation of Natural History Collections (SPNHC).

We developed an R Shiny application that integrates characters and taxa from Phenoscape with specimen records from iDigBio and phylogenies from OT, to visualize phenotypic characters and taxon distributions in three interactive panels. The app allows a user to visualize OT phylogenies and place presence/absence character data on the tree. Specifically, users can: select taxa or specific characters to visualize their geographic distributions, navigate a phylogeny browser which displays character and specimen data available for taxa under consideration, and view a heatmap of characters available for character and taxon combinations. Because of our challenges joining data, our distribution map leaves users with the impression that all individuals in a genus exhibit a character whereas the KB was populated with data describing individuals. We hope that with improved data standards and their use by more people, constructing applications like ours will become easier.

Keywords

phenoscape, idigbio, phylogeny, trait, linked data

Presenting author

Matthew Collins

Presented at

Biodiversity Information Standards (TDWG) 2018, Dunedin, NZ

References

- Edmunds R, Su B, Balhoff J, Eames BF, Dahdul W, Lapp H, Lundberg J, Vision T, Dunham R, Mabee P, Westerfield M (2015) Phenoscape: Identifying Candidate Genes for Evolutionary Phenotypes. Molecular Biology and Evolution 33 (1): 13-24. https://doi.org/10.1093/molbev/msv223
- Hinchliff C, Smith S, Allman J, Burleigh JG, Chaudhary R, Coghill L, Crandall K, Deng J, Drew B, Gazis R, Gude K, Hibbett D, Katz L, Laughinghouse HD, McTavish EJ, Midford P, Owen C, Ree R, Rees J, Soltis D, Williams T, Cranston K (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences* 112 (41): 12764-12769. https://doi.org/10.1073/pnas.1423041112
- Matsunaga A, Thompson A, Figueiredo R, Germain-Aubrey C, Collins M, Beaman R, MacFadden B, Riccardi G, Soltis P, Page L, Fortes JB (2013) eScience (eScience), 2013 IEEE 9th International Conference. 2013 IEEE 9th International Conference on e-Science https://doi.org/10.1109/escience.2013.48