Community content building for evolutionary biology: Lessons learned from LepTree and Encyclopedia of Life

Cynthia Parr, Encyclopedia of Life, National Museum of Natural History, Smithsonian Institution parrc@si.edu

Dana Campbell, LepTree, University of Maryland, College Park, danac@umd.edu

John Park, LepTree, University of Maryland, College Park, mailto:park@umiacs.umd.edu

Online resources to aid large-scale ecological and evolutionary biology are beginning to take root, only a decade behind fields such as genomics and molecular biology. One barrier has been a long tradition, in evolutionary biology at least, of work by individuals on the order of a few hundred of species rather than the thousands or hundreds of thousands necessary to understand the general evolutionary or ecological processes that explain species characteristics and distributions. Advances in collaborative and semantic software offer promise – it should be possible to develop high quality online species-level datasets for comparative analyses and even to integrate, via machine reasoning, across highly customized datasets. In this talk we will compare and contrast two approaches to assembling the data. The LepTree project (http://www.leptree.net) has undertaken large scale molecular phylogenetics to determine deep structure in the very large (>180,000) Lepidopteran clade. Parallel with the molecular work, we built online, open-source largely Drupal-based tools to facilitate assembly of morphological and life history characteristics, and terminology for the superfamilies, families, and subfamilies in the group. By storing these data as highly atomized, ontology-driven triples we aim to facilitate machine reasoning to reconcile divergent vocabularies about morphology, permit inference to assign character states at the leaves of the tree, and aid discovery of links between genetic diversity and morphological diversity. Encyclopedia of Life (http://www.eol.org, open source under MIT license at http://github.com/eol) has taken a very different initial approach, relying on mash-up technology using a simple data standard, TDWG's Species Profile Model, and sophisticated names management to assemble unstructured text and image content from partner databases for much of the 1.9 million species known. Tools for expert curation of this content allow us to include sources like Flickr and Wikipedia. Both projects have challenges making good use of expert time to act as editors. Though neither project has yet generated large quantities of comparative data, frameworks are built and they are both poised to grow. The LepTree approach cautions us that effort spent "semanticizing" data up front is probably not worthwhile, but that scientists are willing to categorize using standard vocabularies and write brief text when the rewards are clear. The EOL team has learned that even brute force efforts require significant human input, but scaling is possible. The next step is to marry these approaches – foster integration by EOL of increasingly atomized, semantic data that is either gathered by specialist communities such as LepTree or generated by crowd-sourcing or data extraction and then reviewed by specialists on EOL. LepTree data now flows to EOL and will serve as one of several case studies for exploring the integration and subsequent serving of atomized data suitable for large scale comparative evolutionary studies.

Slides are available at Slideshare: http://slidesha.re/cbbkmx