Organization: Academy of Natural Sciences (ANSP) of Drexel University
Primary mentor: Steve Dilliplane, ANSP Biodiversity Informatics Support
Supporting mentors: Jennifer Vess, ANSP Brooke Dolan Archivist & Vaughn Shirey, Georgetown University Biodiversity Informatics PhD student
Project title: From Natural History Literature to Linked Open Data Biodiversity Knowledge Graph

The Academy of Natural Sciences of DREXEL UNIVERSITY

Description: Biodiversity research depends on physical, temporal and geospatial context for understanding species occurrence, abundance, and distribution as well as for understanding organism behavior and lifecycle. Historic and contemporary taxonomic and related literature are important sources for references to what species occurred where and when and are often, along with museum specimens, the only extant evidence for species characteristics and occurrence. This project will focus on automating the identification and disambiguation of specimen description relationships within biodiversity literature texts found in the Biodiversity Heritage Library (https://biodiversitylibrary.org) and, where possible, present the resulting biodiversity knowledge (sub-)graph for visual browsing.

Problems: Though there has been research done on concept extraction from full-text digital libraries, no projects have systematically linked corresponding ontological concepts across a variety of important facets of the biodiversity knowledge graph. Challenges to be addressed include: 1) identifying specific knowledge gaps and opportunities presented by various selections of natural history literature, 2) introducing standard terms for "extended specimen" data and relationships, and 3) establishing metrics for accuracy and precision of extracted links.

Techniques: Within the biodiversity and information management communities, several natural language processing, text mining, semantic analysis, and georeferencing tools and resources are already available.

Data: Proceedings of the Academy of Natural Sciences of Philadelphia dating back to 1841, digitized through the Biodiversity Heritage Library, contain a wealth of descriptions of natural history specimens of interest to studies in taxonomy, biogeography, morphology and phenology. This data is typical of BHL resources available to supplement and extend specimen collection records data available through the Global Biodiversity Information Facility (GBIF) and Integrated Digitized Biocollections data aggregators.

Outcome: This project seeks to augment existing biodiversity specimen collection data by linking related descriptions found in the rich tradition of natural history literature. The research will assess the accuracy and precision of the automated processes mining, cross-referencing and representing multiple sources of biodiversity information.