Biodiversity Information Science and Standards 2: e26222
doi: 10.3897/biss.2.26222

Conference Abstract

Digitizing EPICC Data: Trials and Tribulations in Translating 100 Year Old Data

Holly Little[‡], Anna K Leary[‡], Alexandra L Cano[‡], Adam Mansur[‡]

‡ Smithsonian National Museum of Natural History, Washington, DC, United States of America

Corresponding author: Holly Little (littleh@si.edu)

Received: 26 Apr 2018 | Published: 15 Jun 2018

Citation: Little H, Leary A, Cano A, Mansur A (2018) Digitizing EPICC Data: Trials and Tribulations in Translating 100 Year Old Data. Biodiversity Information Science and Standards 2: e26222. https://doi.org/10.3897/biss.2.26222

Abstract

The Smithsonian National Museum of Natural History (NMNH) Department of Paleobiology recently completed the first segment of a mass digitization project in support of the Eastern Pacific Invertebrate Communities of the Cenozoic (EPICC) thematic collections network. In collaboration with the Smithsonian Institution Digitization Project Office (DPO), the team imaged and transcribed labels from a portion of the Cenozoic Mollusca Collection. Once the labels were transcribed further processing was required to clean and enhance that specimen data. We sought to ensure high quality data for this project through:

1. the development of clear guidelines for documentation and treatment of specific data points;
2. updating records to match current taxonomic, lithostratigraphic, and chronostratigraphic information; and
3. create iterative workflows to maintain extensibility and to capture uncertainty in the data.

A significant challenge for any large collections digitization project is transcribing and cleaning analog information from specimen labels. Often these labels are unstructured with varying levels of data quality and quantity, making interpretation of the data difficult. These problems are compounded for a large scale project combining specimens from multiple collectors or research projects. During this digitization project, we developed methods for accounting for possibly unverified, poorly documented, or sparse analog data; for selecting tools and procedures to efficiently transform this data into standardized vocabularies and structures while ensuring data quality; and for maintaining transparency by clearly documenting the decisions and interpretations made by catalogers. To improve the efficiency of the process, we also used technologies such as Python scripting and OpenRefine to help clean and standardize the data. These steps enabled us to face these challenges of translating analog collections data of over a hundred years old into modern standards for biodiversity information.

Keywords

Digitization, Paleontology, Data Standards, Transcription

Presenting author

Holly Little