Extracting Expedition Log Data Found in the Biodiversity Heritage Library

The Biodiversity Heritage Library (BHL) has amassed a large collection of expedition logs and archival field notes. The data contained in these materials are non-standard, unreadable by machines, and unvetted by humans.

Introduction

Species occurrence data contained in expedition logs can reveal snapshots of past ecological states and illuminate human impacts on species habitats. Extracting this data from BHL’s historic texts will increase humanity’s understanding of environmental change through time at hyperlocal and global scales.

Methodology

Piloted the application of Handprint, a Handwriting Text Recognition (HTR) algorithm, to images of analog field notes and used bounding box outputs to sort text into a data table.

Analysis

Handprint, an HTR program developed at CalTech Library, automatically groups together identified text items into lines, which roughly correspond to cells in a data table. We were then able to identify groups of boxes that overlapped along the x- and y-axes to create columns and rows.

Because the handwritten table was neatly spaced, the pipeline was largely successful. This process will be less effective for messier data tables.

We Need Feedback!

If you are a computational researcher, bioinformatician, museum professional, or environmental policy analyst interested in making use of historic species data, please:

→ Click here to answer a brief survey

Objective

BHL staff would like to deposit invaluable historic species occurrence data with big data biodiversity aggregators like the Global Biodiversity Information Facility (GBIF) and information brokers like Wikidata. Once deposited, this data could then be leveraged by climate change researchers, developers of global species monitoring platforms, and environmental policy makers.

Results

Handwriting Text Recognition (HTR) is still unable to reliably transform handwriting into machine readable text, but it is improving. It is now possible to sort text output into tables using bounding box coordinates.

The Sample Set

Birds observed, banded, and collected as part of the Pacific Ocean Biological Survey Program (POBSP), 1961-1973. The Program deployed over 40 Smithsonian Institution employees to conduct biological surveys of plants and animals that occurred on the islands and atolls. A major focus was determining migration, distribution, and populations of seabirds.

Conclusion

Once images of occurrence records are transformed into machine readable data, species names can be validated, identifiers can be added, datasets can be deposited to biodiversity data aggregators, and catalog records of literature in BHL can be better connected to specimen records in natural history museums.

Data & Materials


Handprint Handwriting Text Recognition: https://github.com/caltechlibrary/handprint

Transformation code available at: https://github.com/kmika11/bhl_unlocking_datatables

Poster Presented at SPNHC 2022 | Liberating Natural History Collections Data in Biodiversity Literature