BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 7: e112466
doi: 10.3897/biss.7.112466

OPEN ACCESS

Conference Abstract

High Throughput Information Extraction of Printed Specimen Labels from Large-Scale Digitization of Entomological Collections using a Semi-Automated Pipeline

Margot Belot‡, Leonardo Preuss‡, Joël Tuberosa‡, Magdalena Claessen‡, Olha Svezhentseva‡, Franziska Schuster‡, Christian Bölling‡, Théo Léger‡

‡ Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin, Germany

Corresponding author: Margot Belot (margot.belot@mfn.berlin)

Received: 09 Sep 2023 | Published: 11 Sep 2023

Citation: Belot M, Preuss L, Tuberosa J, Claessen M, Svezhentseva O, Schuster F, Bölling C, Léger T (2023) High Throughput Information Extraction of Printed Specimen Labels from Large-Scale Digitization of Entomological Collections using a Semi-Automated Pipeline. Biodiversity Information Science and Standards 7: e112466. https://doi.org/10.3897/biss.7.112466

Abstract

Insects account for half of the total described living organisms on Earth, with a vast number of species awaiting description. Insects play a major role in ecosystems but are yet threatened by habitat destruction, intensive farming, and climate change. Museum collections around the world house millions of insect specimens and large-scale digitization initiatives, such as the digitization street digitize! at the Museum für Naturkunde, have been undertaken recently to unlock this data. Accurate and efficient extraction of insect specimen label information is vital for building comprehensive databases and facilitating scientific investigations, sustainability of the collected data, and efficient knowledge transfer. Despite the advancements in high-throughput imaging techniques for specimens and their labels, the process of transcribing label information remains mostly manual and lags behind the pace of digitization efforts.

In order to address this issue, we propose a three step semi-automated pipeline that focuses on extracting and processing information from individual insect labels. Our solution is primarily designed for printed insect labels, as the OCR (optical character

recognition) technology performs well for printed text while handwritten texts still yield mixed results. The pipeline incorporates computer vision (CV) techniques, OCR, and a clustering algorithm. The initial stage of our pipeline involves image analysis using a convolutional neural network (CNN) model. The model was trained using 2100 images from three distinct insect label datasets, namely AntWeb (ant specimen labels from various collections), Bees & Bytes (bee specimen labels from the Museum für Naturkunde), and LEP_PHIL (Lepidoptera specimen labels from the Museum für Naturkunde). The first model enables the identification and isolation of single labels within an image, effectively segmenting the label region from the rest of the image, and crops them into multiple new, single-label image files. It also assigns the labels to different classes, i.e., printed text or handwritten, with handwritten labels sorted out from the printed ones. In the second step, labels classified as “printed” are then parsed by an OCR engine to extract the text information from the labels. Tesseract and Google Vision OCRs were both tested to assess their performance. While Google Vision OCR is a cloud-based service with limited configurability, Tesseract provides the flexibility to fine-tune settings and enhance its performance for our specific use cases. In the third step, the OCR outputs are aggregated by similarity using a clustering algorithm. This step allows for the identification and formation of clusters that consist of labels sharing identical or highly similar content. Ultimately, these clusters are compared against a curated database of labels and are assigned to a known label or highlighted as new and manually added to the database.

In order to assess the efficiency of our pipeline, we performed benchmarking experiments using a set of images similar to those the models were trained on, as well as additional image sets obtained from various museum collections. Our pipeline offers several advantages, streamlining the data entry process, and reducing manual extraction time and effort, while also minimizing potential human errors and inconsistencies in label transcription. The pipeline holds the promise of accelerating metadata extraction from insect specimens, promoting scientific research and enabling large-scale analyses to achieve a more profound understanding of the collections.

Keywords

artificial intelligence, natural history, OCR, convolutional neural network, clustering, insects, museum

Presenting author

Margot Belot

Presented at

TDWG 2023

Conflicts of interest

The authors have declared that no competing interests exist.