BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e138512
doi: 10.3897/biss.8.138512

OPEN ACCESS

Conference Abstract

Unscrambling the Eggs: Automated Data Extraction from Structured Record Cards

Arianna Salili-James‡, Ben Scott‡, Douglas Russell§, Ken Norris‡, Sanson T. S. Poon‡, Vincent S. Smith‡

‡ Natural History Museum, London, United Kingdom
§ Natural History Museum, Tring, United Kingdom

Corresponding author: Arianna Salili-James (arianna.salili-james@nhm.ac.uk), Ben Scott (b.scott@nhm.ac.uk), Douglas Russell (d.russell@nhm.ac.uk), Ken Norris (k.norris@nhm.ac.uk)

Received: 03 Oct 2024 | Published: 04 Oct 2024

Citation: Salili-James A, Scott B, Russell D, Norris K, Poon STS, Smith VS (2024) Unscrambling the Eggs: Automated Data Extraction from Structured Record Cards. Biodiversity Information Science and Standards 8: e138512. https://doi.org/10.3897/biss.8.138512

Abstract

The Natural History Museum in the UK (NHM) is home to more than 80 million objects spanning 4.5 billion years of history. Each of these contain a wealth of data, whether on specimen labels, index cards, registers and/or diaries. Transcribing and categorising this information can help unlock crucial research potential. To do this at scale, we turn to computer vision (CV) and Machine Learning (ML) techniques to automate this work.

Over a million of the museum's specimens are ornithological, including one of the largest and most comprehensive egg collections in the world. Representing 52% of known bird species, with over 300,000 clutches (where a clutch defines the total group of eggs laid in a nest), collected over the last 200 years, arguably make this the most important archive of avian environmental change data in existence(Norris et al. 2023). The eggs were historically catalogued using index cards, containing key information such as identification, collection date, locality and clutch size. A proportion of these egg cards have now been imaged and this led to the start of this project, focusing on a sample of 15,000 photographed egg cards (example seen in Fig. 1).

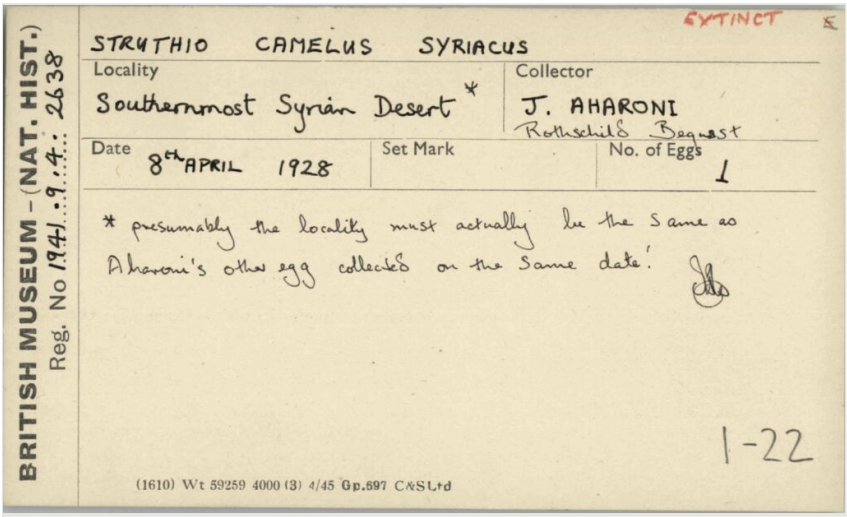

Figure 1.

An example of an egg index card from the NHM collection at NHM Tring. We are interested in extracting the information from all categories. Here, these categories are split into boxes within the card. While many index cards today are printed, most examples in the collection are handwritten, and sometimes contain multiple handwritings, as seen in this example.

Our initial approach used Google Vision to perform Optical Character Recognition (OCR) to transcribe all text with the egg cards. By focusing on textboxes around key terms (e.g., “Collector”), and using CV tools, we approximated boxes around every key category. Finally, each text segment was associated to a category box, followed by minor post-processing in order to extract (i.e., transcribe and categorise) the data. Here we successfully extracted the data within the sample, with a 98.6% average accuracy. Although our methods worked well for our sample, they did rely on consistency within the structures of cards.

To expand the project further, and to mitigate the reliance on consistent structures within cards, we turned to Large Language Models (LLMs). This allowed us to explore automatic data extraction from different types of cards and labels, despite variation in the card structure, and even handle unknown categories of text. Consequently, the scope of the data collected was widened, such as adding ornithological specimen data (e.g., skins), as well as external datasets through collaboration with the British Trust for Ornithology, who manage the Nest Record Scheme (Crick et al. 2003), which holds decades of vital information on the progress of monitored nests in the UK.

This index-card data-extraction project is just the beginning. As we expand our data extraction capabilities, our aim is to develop a novel pipeline that can be applied not just to avifauna-related cards, but any structured textual data, with the potential to unlock invaluable insights.

Keywords

digitisation, OCR, LLMs, ornithology

Presenting author

Arianna Salili-James

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Crick HP, Baillie S, Leech E (2003) The UK Nest Record Scheme: its value for science and conservation. Bird Study 50: 254-270. https://doi.org/10.1080/00063650309461318
- Norris K, Bond A, Cooper J, Adams M, van Grouw H, White J, Stervander M, Russell DD, Loader S (2023) Unlocking avian museum collections to enable and advance environmental change research. Ibis 166 (1): 315-322. https://doi.org/10.1111/ibi.13271