Planetary Knowledge Base: Semantic Transcription Using Graph Neural Networks

Qianqian Gu‡, Ben Scott‡, Vincent S. Smith‡

‡ Natural History Museum, London, United Kingdom

Corresponding author: Qianqian Gu (qianqian.gu@nhm.ac.uk)

Abstract

The Natural History Museum, London (NHM), in collaboration with Amazon Web Services (AWS), has embarked on a project to build the Planetary Knowledge Base (PKB), a comprehensive graph network comprising data on all specimens, collectors, and localities. In the initial prototype, we have concentrated on botanical specimens, using all plant taxa and specimens within the Global Biodiversity Information Facility (GBIF), combined with geographic data from GeoNames and biographic data from WikiData, Bionomia, Harvard Index of Botany, TL2, and Tropicos. Development of the PKB is a huge undertaking—our first proof of concept has more than 100 million nodes.

The primary application of this knowledge graph (KG) is powering the automated transcription of specimen labels. Using Graph Convolutional Neural Networks, textual information from labels can be aligned to the entities in the graph, creating structured semantic data from the raw text. Text is extracted from images using services from the AWS ecosystem, including Optical Character Recognition and Natural Language Processing to identify the units of information, creating a high-throughput auto-digitisation workflow for extracting structured data.

The PKB graph network enables new ways to interrogate collections. It can help identify species that may require re-examination or re-identification due to taxonomic updates or inconsistencies. It can also flag potential discrepancies or conflicts in the data, such as cases where the same species is recorded under different names or classifications across various sources. Moreover, the PKB can detect possible errors and outliers in the knowledge graph and point out specimens that could represent new species misidentified within the collection. By cross-validating species with the International Union for Conservation of Nature (IUCN) Red List, it can also assist in analysing species populations with insufficient data.

The PKB is being developed as a cloud service, so researchers and other institutions can experiment with this transformative technology, using it to support their own digitisation efforts.

Keywords

knowledge graph, knowledge base, machine learning, cloud service

Presenting author

Qianqian Gu

Presented at

TDWG 2023

Acknowledgements

AWS (Amazon Web Services) Global Impact Computing Team

Conflicts of interest

The authors have declared that no competing interests exist.