Extracting Data at Scale: Machine learning at the Natural History Museum

Ben Scott[‡], Laurence Livermore[‡]

‡ The Natural History Museum, London, United Kingdom

Corresponding author: Ben Scott (b.scott@nhm.ac.uk)

Abstract

The Natural History Museum holds over 80 million specimens and 300 million pages of scientific text. This information is a vital research tool to help solve the most important challenge humans face over the coming years – mapping a sustainable future for ourselves and the ecosystems on which we depend. Digitising these collections and providing the data in a structured, computable form is a mammoth challenge. As of 2020, less than 15% of available specimen information currently residing on specimen labels or physical registers is digitised and publicly available (Walton et al. 2020). Machine learning applications can deliver a step-change in our activities' scope, scale, and speed (Borsch et al. 2020).

As part of SYNTHESYS+, the Natural History Museum is leading on the development of a cloud-based workflow platform for natural science specimens, the Specimen Data Refinery (SDR) (Smith et al. 2019). The SDR will provide a series of Machine Learning (ML) models, ranging from semantic segmentation to identify regions of interest on labels, to natural language processing to extract locality and taxonomic text entities from the labels, and image analysis to identify specimen traits and collection quality metrics. Each ML task is atomic, with users of the SDR selecting which model would best extract data from their digitised specimen images, allowing the workflows to be used in different institutions worldwide. It also solves one of the key problems in developing ML-based applications: the rapidity at which models become obsolete. New ML models can be introduced into the workflow, with incremental changes to improve processing, without interruption or refactoring of the pipeline.

Alongside specimens, digitised images of pages of scientific literature provide another vital source of data. Functional traits mediate the interactions between plant species and their environment and play roles in determining species' range size and threatened status. Such information is contained within the taxonomic descriptions of species and a natural language processing library has been developed to locate and extract plant functional traits from these texts (Hoehndorf et al. 2016). The ML models allow complex

interrelationships between taxa and trait entities to be inferred based on the grammatical structure of sentences, improving the accuracy and extent of data point extraction.

These two projects, like many other applications of ML in natural history collections, are focused on the extraction of visible information, for example, a piece of text or a measurable trait. Given the image of the specimen or page, a person would be able to extract the self-same information. However, ML excels in pattern matching and inferring unknown characters from an entire corpus. At the museum, we have started exploring this space, with our voyagerAI project for identifying specimens collected on historical expeditions of scientific discovery (e.g., the voyages of the Beagle and Challenger). This process fills in the gaps in specimen provenance and identifies 'lost' specimens collected by some of the most famous names in biodiversity history. Developing new applications of ML to uncover scientific meaning and tell the narratives of our collections, will be at the forefront of our scientific innovation in the coming years. This presentation will give an overview of these projects, and our future plans for using ML to extract data at scale within the Natural History Museum.

Keywords

artificial intelligence, museums, informatics

Presenting author

Ben Scott

Presented at

TDWG 2021

Conflicts of interest

References

- Borsch T, Stevens A, Häffner E, Güntsch A, Berendsohn W, Appelhans M, Barilaro C, Beszteri B, Blattner F, Bossdorf O, Dalitz H, Dressler S, Duque-Thüs R, Esser H, Franzke A, Goetze D, Grein M, Grünert U, Hellwig F, Hentschel J, Hörandl E, Janßen T, Jürgens N, Kadereit G, Karisch T, Koch M, Müller F, Müller J, Ober D, Porembski S, Poschlod P, Printzen C, Röser M, Sack P, Schlüter P, Schmidt M, Schnittler M, Scholler M, Schultz M, Seeber E, Simmel J, Stiller M, Thiv M, Thüs H, Tkach N, Triebel D, Warnke U, Weibulat T, Wesche K, Yurkov A, Zizka G (2020) A complete digitization of German herbaria is possible, sensible and should be started now. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e50675

- Hoehndorf R, Alshahrani M, Gkoutos GV, Gosline G, Groom Q, Hamann T, Kattge J, de Oliveira SM, Schmidt M, Sierra S, Smets E, Vos RA, Weiland C (2016) The flora phenotype ontology (FLOPO): tool for integrating morphological traits and phenotypes of vascular plants. Journal of Biomedical Semantics 7 (1): 65. https://doi.org/10.1186/s13326-016-0107-8
- Smith V, Gorman K, Addink W, Arvanitidis C, Casino A, Dixey K, Dröge G, Groom Q, Haston E, Hobern D, Knapp S, Koureas D, Livermore L, Seberg O (2019) SYNTHESYS+ Abridged Grant Proposal. Research Ideas and Outcomes 5 https://doi.org/10.3897/rio.5.e46404.figure5
- Walton S, Livermore L, Bánki O, Cubey R, Drinkwater R, Englund M, Goble C, Groom Q, Kermorvant C, Rey I, Santos C, Scott B, Williams A, Wu Z (2020) Landscape Analysis for the Specimen Data Refinery. Research Ideas and Outcomes 6 https://doi.org/10.3897/rio.6.e57602