# Cloud AI: A comparison of specimen image data extraction processes

Ben Scott ‡

‡ Natural History Museum, London, United Kingdom

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

## Abstract

The Natural History Museum (NHM) of London has embarked on an ambitious programme to digitise the 80 million specimens in its collection, releasing them through the NHM data portal and the global biodiversity research community. As part of the digitisation process, data is transcribed from specimen labels to capture the vital taxonomic and collection event data. Accurate human transcription is slow and the NHM, like many institutions, has been exploring machine learning (ML) for automated specimen analysis and label data capture. This process requires many different models, chained in series: semantic segmentation to identify specimen and label regions of interest; optical character recognition to identify text on labels; natural language processing to extract entities from the text.

As part of SYNTHESYS+, the NHM has been building the Specimen Data Refinery (SDR) ( Smith et al. 2019) - a workflow engine for chaining ML models, each performing one atomic task in the data extraction process. The SDR is now in public beta, and we present evaluation metrics from our initial testing. Alongside the SDR project, the NHM has been exploring cloud-based artificial intelligence tools for specimen digitisation, using Google and Amazon technologies. We present an analysis of these different approaches, comparing the results from third-party AI services with models developed specifically for the biodiversity and natural history collection domains. With large corporates providing comparatively low-cost access to AI compute resources and models transferrable to many specimen image digitisation tasks, is developing bespoke solutions still required?

## Keywords

machine learning, cloud computing, digitisation, natural history collections, specimen data

## Presenting author

Ben Scott

## Presented at

TDWG 2022

## Conflicts of interest

## References


- 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