Machine Learning as a Service for DiSSCo’s Digital Specimen Architecture

Jonas Grieb[‡], Claus Weiland[‡], Alex Hardisty[§], Wouter Addink[|,¶], Sharif Islam[|,¶], Sohaib Younis[#], Marco Schmidt[¤]

‡ Senckenberg - Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main, Germany
§ School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom
| Naturalis Biodiversity Center, Leiden, Netherlands
¶ Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands
# Department of Mathematics and Computer Science, Philipps-University Marburg, Marburg, Germany
¤ Palmengarten der Stadt Frankfurt, Frankfurt am Main, Germany

Corresponding author: Claus Weiland (cweiland@senckenberg.de)

Abstract

International mass digitization efforts through infrastructures like the European Distributed System of Scientific Collections (DiSSCo), the US resource for Digitization of Biodiversity Collections (iDigBio), the National Specimen Information Infrastructure (NSII) of China, and Australia’s digitization of National Research Collections (NRCA Digital) make geo- and biodiversity specimen data freely, fully and directly accessible.

Complementary, overarching infrastructure initiatives like the European Open Science Cloud (EOSC) were established to enable mutual integration, interoperability and reusability of multidisciplinary data streams including biodiversity, Earth system and life sciences (De Smedt et al. 2020).

Natural Science Collections (NSC) are of particular importance for such multidisciplinary and internationally linked infrastructures, since they provide hard scientific evidence by allowing direct traceability of derived data (e.g., images, sequences, measurements) to physical specimens and material samples in NSC.

To open up the large amounts of trait and habitat data and to link these data to digital resources like sequence databases (e.g., ENA), taxonomic infrastructures (e.g., GBIF) or environmental repositories (e.g., PANGAEA), proper annotation of specimen data with rich (meta)data early in the digitization process is required, next to bridging technologies to facilitate the reuse of these data.

This was addressed in recent studies (Younis et al. 2018, Younis et al. 2020), where we employed computational image processing and artificial intelligence technologies (Deep Learning) for the classification and extraction of features like organs and morphological traits from digitized collection data (with a focus on herbarium sheets).

However, such applications of artificial intelligence are rarely—this applies both for (sub-symbolic) machine learning and (symbolic) ontology-based annotations—integrated in the workflows of NSC's management systems, which are the essential repositories for the aforementioned integration of data streams.

This was the motivation for the development of a Deep Learning-based trait extraction and coherent Digital Specimen (DS) annotation service providing "Machine learning as a Service" (MLaaS) with a special focus on interoperability with the core services of DiSSCo, notably the DS Repository (nsidr.org) and the Specimen Data Refinery (Walton et al. 2020), as well as reusability within the data fabric of EOSC.

Taking up the use case to detect and classify regions of interest (ROI) on herbarium scans, we demonstrate a MLaaS prototype for DiSSCo involving the digital object framework, Cordra, for the management of DS as well as instant annotation of digital objects with extracted trait features (and ROIs) based on the DS specification openDS (Islam et al. 2020).

Source code available at: https://github.com/jgrieb/plant-detection-service

Keywords

FAIR Digital Object, Distributed System of Scientific Collections, plant organ detection, deep learning, region-based convolutional neural network, image annotation

Presenting author

Jonas Grieb

Presented at

TDWG 2021

Funding program

H2020-INFRADEV-2019-2020 – Grant Agreement No. 871043

Deutsche Forschungsgemeinschaft (DFG) - Project number 316452578

Conflicts of interest

References

- De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8 (2). https://doi.org/10.3390/publications8020021
- Islam S, Hardisty A, Addink W, Weiland C, Glöckler F (2020) Incorporating RDA Outputs in the Design of a European Research Infrastructure for Natural Science Collections. Data Science Journal 19 https://doi.org/10.5334/dsj-2020-050
- 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
- Younis S, Weiland C, Hoehndorf R, Dressler S, Hickler T, Seeger B, Schmidt M (2018) Taxon and trait recognition from digitized herbarium specimens using deep convolutional neural networks. Botany Letters 165: 377-383. https://doi.org/10.1080/23818107.2018.1446357
- Younis S, Schmidt M, Weiland C, Dressler S, Seeger B, Hickler T (2020) Detection and annotation of plant organs from digitised herbarium scans using deep learning. Biodiversity Data Journal 8 https://doi.org/10.3897/bdj.8.e57090