Biodiversity Information Science and Standards 3: e37647
doi: 10.3897/biss.3.37647

Conference Abstract

Specimen Data Refinery: A landscape analysis on machine learning, computer vision and automated approaches to capture specimen metadata

Laurence Livermore[‡], Robert W. N. Cubey[§]

[‡] The Natural History Museum, London, United Kingdom
[§] Royal Botanic Garden Edinburgh, Edinburgh, United Kingdom

Corresponding author: Laurence Livermore (l.livermore@nhm.ac.uk)

Received: 24 Jun 2019 | Published: 02 Jul 2019

Citation: Livermore L, Cubey R (2019) Specimen Data Refinery: A landscape analysis on machine learning, computer vision and automated approaches to capture specimen metadata. Biodiversity Information Science and Standards 3: e37647. https://doi.org/10.3897/biss.3.37647

Abstract

Capturing data from specimen images is the most viable way of enriching specimen metadata cheaply and quickly compared to traditional digitisation. Advances in machine learning and computer vision-based tools, and their increasing accessibility and affordability, are greatly increasing the potential to take automated measurements and capture other data from specimens themselves, as well as to transcribe label data.

More sophisticated segmentation of images allows us to find parts of interest: particular labels; individual specimens on a slide; or barcodes. Following segmentation, there is the potential to use colour analysis of specimens to perform conditional checking, such as looking for bad cases of verdigris in pinned insects or discoloration of gum-chloral mountant. Automating measurements and landmark analysis of specimens can be used to create trait datasets, all of which will enrich our knowledge of specimens. Segmentation of labels can allow us to cluster similar labels based on their visual properties including colour, shape and patterns—this in turn can be used to make optical character recognition, handwriting recognition and manual transcription much more efficient. Atomising, validating and resolving label data will create structured label data that can be more easily stored, searched and linked to other datasets.

© Livermore L, Cubey R. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

We present a landscape analysis on the approaches, summarising previous work, and outline our plan to build future tools and systems in the SYNTHESYS+ Project as part of the Specimen Data Refinery. This will cover the sharing of tools, reducing barriers to access, integrating workflow engines into a software architecture that allows the components to be re-used and re-purposed with provenance data for repeatability, and conforms with the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles (Wilkinson et al. 2016).

Keywords

machine learning, digitisation, automation, natural history specimens

Presenting author

Laurence Livermore

Presented at

Biodiversity_Next 2019

Funding program

The Specimen Data Refinery is part of SYNTHESYS+ and funded from the RIA - Research and Innovation action in the H2020-EU.1.4.1.2. Programme - Integrating and opening existing national and regional research infrastructures of European interest.

References

- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, ’t Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 https://doi.org/10.1038/sdata.2016.18