BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e138060
doi: 10.3897/biss.8.138060

OPEN ACCESS

Conference Abstract

Development of an Automated Label Data Entry System from Herbarium Specimen Images at Hyogo Herbarium (HYO)

Atsuko Takano[‡,§], Yasuhiko Horiuchi[|], Hajime Konagai[¶], Chung-Kun Lee[#,¤], Hiromune Mitsuhashi[«]

‡ University of Hyogo, Sanda, Japan
§ Museum of Nature and Human Activities, Hyogo, Sanda, Japan
| The NPO Field, Takatsuki, Japan
¶ Functiontales, Kyoto, Japan
# Museum of Nature and Human Activities, Sanda, Japan
¤ University of Hyogo, Hyogo, Japan
« University of Hyogo, Museum of Nature and Human Activities, Hyogo, Sanda, Japan

Corresponding author: Atsuko Takano (takano@hitohaku.jp)

Received: 28 Sep 2024 | Published: 30 Sep 2024

Citation: Takano A, Horiuchi Y, Konagai H, Lee C-K, Mitsuhashi H (2024) Development of an Automated Label Data Entry System from Herbarium Specimen Images at Hyogo Herbarium (HYO). Biodiversity Information Science and Standards 8: e138060. https://doi.org/10.3897/biss.8.138060

Abstract

We would like to introduce our recently developed systems for taking images of herbarium specimens and for the automatic extraction of data from specimen labels at the Herbarium of the Museum of Nature and Human Activities, Hyogo, Japan (HYO).

Firstly, we designed a low-cost, but high-quality specimen imaging system for non-professional photographers to obtain images rapidly (Takano et al. 2019). Our system uses a mass-produced, mirrorless single-lens reflex (SLR) camera (SONY ILCE6300) with a zoom lens (Samyang Optics SYIO35AF-E35 mm F/2.8). We made a photo stand by ourselves to reduce costs. In addition, we have adopted an LED (light-emitting diode) lighting system with high color rendering. This imaging system has been introduced, with some improvements or adjustments for available space, to various herbaria in Japan (e.g., University of Tokyo (TI), Kyoto University (KYO)), contributing to the digitization of herbarium specimens across Japan.

© Takano A et al. 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.

Next, we developed a system to extract label information from specimen images. The specimen image was uploaded to Google OCR and data were extracted in the form of text. Uploading the whole specimen image decreased the reading accuracy of the software because the plant images behaved as OCR (Optical Character Reader) noise. Therefore, the label part was cut out from the whole specimen image by using D-Lib*[1] and uploaded to tesseract OCR*[2] for OCR extraction of the label information (Aoki 2019, Takano et al. 2020). When installing this system for HYO, we designed it as an application accessible externally via the internet, which proved very useful during the coronavirus pandemic: part-time workers checked and conducted label data input from home.

Finally, we decided to develop a system that would automatically label the text data extracted by OCR and input them into the appropriate cells of the database. Even though the text data could be extracted from specimen images, it needed a human to input them into the database. Therefore, we adopted Named Entity Recognition (NER), a system that extracts named entities such as place names, identifying proper nouns from unstructured text data. It enables information recorded in herbarium specimens to be tagged as named entities. We tried text matching at first, but the result was not satisfactory, so we started to use machine learning instead. We compared three natural language libraries for Japanese: BERT (Bidirectional Encoder Representations from Transformers), Albert (A Lite version of BERT), and SpaCy. Despite BERT and SpaCy returning similarly high f-scores (indicating good performance), we decided to use SpaCy because it runs better on ordinary PCs or servers. With sufficient machine learning after the creation of a text corpus (a specialised dataset) specific to labels on herbarium specimens, we successfully developed the application. The project files are available on GitHub*[3] (Takano et al. 2024).

We then examined whether this system could be applied to non-plant specimen images, i.e., fishes or birds, and found that it could efficiently extract data. Therefore, we decided to publicize this system on the cloud server and share it with other natural history museums in Japan*[4]. Curators can obtain a unique ID and password and upload specimen images from their collection to extract label data. The digitization of natural history collections in Japan has been long behind other countries, and this system will help to accelerate it.

The system mentioned above is specialized for the natural history collections of Japan, but we believe it is possible to build similar programs in other countries, and we hope our experience will contribute to the mobilization of the world's natural history collections.

Keywords

named entity recognition, optical character recognition, digitization

Presenting author

Atsuko Takano

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Aoki K (2019) Automatic recognition and extraction of specimen labels in digital imaging herbarium specimens. "On the subject of the Shoei Collection". Doshisha University Graduate School of Culture and Information Science. [In Japanese].
- Takano A, Horiuchi Y, Fujimoto Y, Aoki K, Mitsuhashi H, Takahashi A (2019) Simple but long-lasting: A specimen imaging method applicable for small- and medium-sized herbaria. PhytoKeys 118: 1-14. https://doi.org/10.3897/phytokeys.118.29434
- Takano A, Horiuchi Y, Aoki K, Fujimoto Y, Mitsuhashi H (2020) Developing new methods for digitization of herbarium specimens and electronic data capture adjustable Japanese herbaria. Jour. Phytogeogr. Taxon 68 (2): 103-119. [In Japanese]. https://doi.org/10.18942/chiribunrui.0682-05
- Takano A, Cole TH, Konagai H (2024) A novel automated label data extraction and data base generation system from herbarium specimen images using OCR and NER. Scientific Reports 14 (1). https://doi.org/10.1038/s41598-023-50179-0

Endnotes

*1 http://dlib.net/

*2 https://github.com/tesseract-ocr

*3 https://github.com/HajimeKonagai/HitohakuA-Lalabel

*4 https://innovatemuseum.net/