様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 １１ 日現在

機関番号：２４５０６
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ０６８３２

研究課題名（和文）機械学習とOCRを用いた植物標本画像からのラベル情報自動取得プログラムの開発

研究課題名（英文）Development of Automatic Label Information Acquisition Program from Plant Specimen Images Using Machine Learning and OCR

研究代表者
高野　温子（Takano, Atsuko）

兵庫県立大学・自然・環境科学研究所・教授

研究者番号：２０３４４３８５

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：　標本画像からのラベルデータ自動抽出法の開発を掲げた当初の目的通り、光学文字認識（OCR）と固有表現抽出（NER）の手法を用いて、ラベルデータをOCRでテキスト抽出し、抽出テキストを採集日や採集者、植物の学名等に分割し、CSVファイル形式で出力するシステムを開発した。また本研究に先立って開発していた標本画像の撮影装置普及にも努め、東京大学、京都大学等の日本国内の各研究機関に標本画像撮影装置が導入されて、日本国内の植物標本デジタル化作業の加速化に寄与することができた。

研究成果の学術的意義や社会的意義
世界に約30億ある自然史標本のデジタル化とウェブ公開は、標本へのアクセシビリティと流動性を高め、生物多様性保全とその研究の促進に大いに貢献する。本研究により開発されたラベルデータ自動読み取りシステムは、現状は植物標本に特化したシステムではあるが、他の標本ラベル情報自動読み取りシステムも今回と同じ方法で開発可能であり、自然史標本全般の流動性を高めることに寄与しうる。2022年に改正された博物館法により所蔵資料のデジタルアーカイブ作成と公開が博物館の努力義務となった今、低コストで実現可能な資料デジタルアーカイブ作成手法を全国の博物館が求めており、本研究はその一つの答えを提供している。

研究成果の概要（英文）：As per the original aim of developing an automatic label data extraction method from herbarium specimen images, a system was developed successfully using Optical Character Recognition (OCR) and Named Entity Recognition (NER: a sort of natural language processing technology). The system extracts label data in text using OCR from specimen image, split and recognize the extracted text into collection dates, collectors, scientific names of plants, etc. by NER, and output the data in CSV file format. Efforts were also made to disseminate the specimen image photographing equipment that had been developed prior to this research, and specimen image photographing equipment was introduced to various research institutions in Japan, including the University of Tokyo and Kyoto University, thereby contributing to accelerating the digitisation process of herbarium specimens in Japan.

研究分野：植物分類学

キーワード：標本デジタル化　OCR　NER　自然言語処理

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

植物・昆虫・岩石鉱物などの自然史標本は世界に 30 億あるとも言われている（Wheeler et al. 2012）。植物や昆虫、岩石等の自然史標本は、何より実物であるという価値と、それらが採集された場所や日にち、環境、採集者の情報などがセットされているという点で、ある時ある場所にそれらの生物や鉱物・岩石がいたという「実在の証拠」としての価値がある。これらは自然史博物館やそれに類する研究機関に保管されデジタル化も進みつつあるが、いまだ十分とはいえない。特に世界の自然史博物館や研究機関と比較して、我が国における自然史資料のデジタル化の歩みは遅い。自然史資料のデジタル化促進のため、まず我々は簡便かつ迅速に高品質なデジタル標本画像を撮影できる装置を開発した（Takano et al. 2019）。本研究では次のステップとして、標本ラベル情報入力の自動化に取り組むことにした。自然史標本は一般に数が多い。例えば兵庫県立人と自然の博物館には昆虫標本約 108 万点、植物（維管束の他、藻類・蘚苔類・菌類含む）標本約 63 万点ある。標本ラベルには生物名の他、採集地名、採集日、採集者名、採集者番号、博物館 ID 等多くの情報が掲載されている。これまでは標本のラベルを熟知したアルバイトにデータ入力を依頼していたが、標本点数の多さに反比例するような限られた予算では、入力作業は滞りがちであった。もし実際の標本を見ながらデータ入力を行うより標本画像を使ったデータ入力の方が楽なら、標本画像からラベルデータを自動取得する仕組みができれば、それは標本デジタル化のインセンティブとして働くことだろう。もうひとつ標本デジタル化のインセンティブ創出の方法として、標本画像の研究や教育普及活動への活用方法を提案することが挙げられる。資料整理関係の業務量が増えるというデメリットを凌駕するメリットがなければ、デジタル化はいつまでも遅々として進まないだろう。そのような問題意識と背景をもって本研究を行った。

２．研究の目的

本研究では、

①標本画像から標本ラベル及びアノテーション（再同定）ラベル部分を機械学習により抽出するプログラムの作成

②ラベル画像から OCR ソフトを用いたテキスト抽出

③標本デジタル画像の研究利用として、画像からの自動種判別システムの開発

④テキストデータの種類わけ・フィールド振り分けプログラム作成

⑤標本画像撮影法及びデジタルアーカイブ作成フリーソフト（Survey Data Collector）の普及

の５つを目的として研究を行った。

３．研究の方法

1. 青木滉太氏（当時：同志社大学大学院）と共同で、人と自然の博物館植物標本画像 2000 枚と、DLib（フリーの画像処理ライブラリ）を用いて、標本画像からラベル部分だけを切り出すプログラムを作成した。ラベルは通常植物採集者が作成するが、人によってサイズが微妙に異なる。また通常は標本台紙の右下にラベルを貼り付けるが、植物体のサイズや形状により、ラベルを貼る場所が変わる場合もある。そういった標本毎のばらつきを考慮しつつ、初めにラベル画像を 100 枚準備して DLIb に学習させた。その後、ラベルを切り出す位置や切り出す画像サイズ

の最適条件を検討した。
2. 1 の成果に基づいて切り出したラベル画像をもちいて、3 つのフリーOCR ソフト（Tesseract OCR, google ドライブ, quick OCR）を比較検討した。
3. 標本デジタル化の更なるインセンティブを創出するため、植物標本画像からの AI 種自動判定システムの構築に関わった。14 万点近い植物標本画像をシステム構築の教師データ・テストデータとして共同研究者の島根大学秋廣高志氏に送り、結果の解釈と新たな実験の提言などを行った。
4. 人と自然の博物館（HYO）所蔵の標本 2 万点の画像、人力で作成した標本 2 万点分の完全なラベル情報の入力データ、（HYO）で使用している日本産維管束植物の学名データベース、日本の郵便番号データ(https://www.post.japanpost.jp/zipcode/download.html
）と紐づいた住所情報を用意した。固有表現認識のラベルは産地（日英)、採集者、採集日等できるだけの情報を拾えるように設定した。

はじめに各種辞書を使用したテキストマッチングの手法を試した。入力テキストから、辞書の項目に適合する箇所を、適合した辞書の項目のクラスとして抽出するという手法である。標本画像 1 万件を用い、①標本画像からラベル部分を検出、google OCR でテキスト抽出、テキスト情報をあるカラムに保存したのち　③学名辞書や地名辞書に一致する部分をテキスト内で検索する作業を行ったところ、一致する箇所が半分以下と少なかった。そこで方針を変え、機械学習による固有表現抽出を行うこととした。

機械学習による分類に基づく固有表現抽出を行うため、マニュアルでテストデータ及び教師データに用いる植物標本ラベルのコーパス作成を行った。戦略として、HYO の標本画像 2 万点中にみられる標本ラベルのバリエーションを網羅することを最優先とした。通常、標本ラベルは植物を採集した人間が作成する。標本ラベルに書く情報は植物名、採集地、採集者名、採集日などでおおよそ共通しているが、ラベルのフォーマットは採集者によって異なる。しかしそれぞれの採集者は、同じフォーマットでラベルをつくり続けることが多い。そこで採集者ごとにデータをグループ化し、それぞれランダムに順位付けし、順位の 1〜3 番目のものを、学習データの作成対象とした。効率的なコーパス作成のため、学習データ作成ツール(アノテーションツール)を制作した。また緯度、経度、標高にもアノテーションを付けることとし、地名については日本語と英語の両方をアノテーションすることにした。前処理として、テキストの一致検索で取得できた固有表現については、事前に Entity(抽出済み固有表現) として登録し、それ以外の学習データの作成を人の手によって行った。採集者名の人数や HYO と標本交換を行っているハーバリウムの数から、2 万点中にはおよそ 1,000 種のラベルフォーマットがあると推定されたため、アノテーションツールを用いて固有表現にアノテーションをつけ、人力で約 1,000 点の教師データを作成した。

人力でアノテーションをつける方法は労力がかかるため、その補完として以下の方法でサンプリングデータを 4,000 件生成した。各項目のサンプリング方法は 以下の通りである。科名（日本語、英語)、学名と日本名については、HYO の植物学名 DB からランダム抽出を行い、地名の各データ（Prefecture,　City, etc.）は Japan Post の全国住所 DB と、HYO の 2 万件の入力済みデータから半々の確率でランダム抽出、採集日については、ランダムな日付を生成し、フォーマットは以下の通り様々な形をとる設定とした（e.g., March.13.1988, 10. Oct. 2003, 16. March, 1997, May.16.1972, 1979/4/1, 1. Aug. 2005, 24. May, 2013, 12. January,1975, Mar. 28. 2009, December. 30, 1988)。採集者名は HYO のデータからランダム抽出を行い、採集者番号は 1〜999999 の間でランダムな番号を生成し、緯度経度は、フォーマットは以下の通り様々なフォー

マットで（e.g., -15° 03'1.6, -70 度 43 分, 24' 03'45, -152 48, 23° 33' 09, 34 度 44 分, 67° 48, -131' 21, -86' 35'26, 70'53'28）ランダムな緯度・経度を生成した。標高については 0‐8000 の間でランダムに数字を生成し、１つだけの表記のものや xxx ～ xxx m など範囲のものも生成した。メモについては多様過ぎるため、生成しなかった。
上述した 2 通りの方法で準備した 2 種類のデータを、１の手作業コーパスデータ → 学習用データ 893 件、テストデータ 77 件、２のサンプリングデータ →学習用データ 4855 件、テストデータ 50 件に振り分けた。テストデータを 2 種準備したのは、実際に作業者が欲しいデータ(手作業によるコーパス)と未知のデータ(サンプリングデータ)を使用したテストを両方行うためである。
Spacy, BERT, Albert の３つの AI モデルに事前学習済み日本語データをロードした後、以下の 4 通りの方法で AI に学習させ、手作業コーパスデータのデータ 77 件とサンプリングデータからのデータ 50 件を、独立のテストデータとして用いて結果を比較した。計測には F 値を使用し、最も高いものを運用時のモデルデータとして採用することにした。4 通りの方法とはそれぞれ、1.手作業コーパスデータのみの学習、2.サンプリングデータのみの学習、3.手作業・サンプリングデータを一緒にした学習、4.「サンプリングデータのみで学習」の後、最良結果をロードし手作業のコーパスデータを学習させる。である。４つの学習方法、 2 種類のテストデータを用いて 20 エポックずつ学習させた。比較した３つの AI モデル、事前学習済みデータ、モデルのパラメータ保存時の容量は以下のとおりである。1. Spacy： ja-ginza (5.55MB), 2．BERT : 'cl-tohoku/bert-base-japanese-whole-word-masking' (419MB). 3．Albert (BERT の軽量モデル) : 'ken11/albert-base-japanese-v1-with-japanese-tokenizer(43.4MB). 動作環境は、HP Z4 G4 Workstation、プロセッサ Intel(R) Xeon(R) W-2223 CPU @ 3.60GHz 3.60 GHz, 実装 RAM 32.0 GB (31.7 GB 使用可能), システムの種類：64 ビット オペレーティング システム、x64 ベース プロセッサ, エディション Windows 10 Pro for Workstations, ストレージ NVMe, GPU NVIDIA RTX A4500 であった。

5．標本画像撮影法及び開発プログラムの普及

NPO 法人西日本自然史博物館ネットワークの協力を得て、2022 年 12 月 3 日に兵庫県立人と自然の博物館で標本撮影装置および Survey Data Collector の使用講習会を行った。また本研究で得た成果を、兵庫県博物館協会（2023 年 2 月）、関東地区博物館協会（2022 年 10 月）、日本植物分類学会　講演会（2022 年 12 月）、同学会大会ポスター発表等にて発表し、2020 年には日本植物分類学会和文誌の植物地理・分類研究に論文を発表した。

４．研究成果

①と②
各種フリーOCR ソフトを比較検討した結果、Google ドライブの OCR 機能が優れていることがわかった。そこで、人と自然の博物館のウェブサーバー上で、ラベル切り出し画像を google ドライブにアップロード→OCR 結果と DB 入力フィールド 標本画像をブラウザ上に表示→標本画像を見ながら OCR 読み取り結果を修正し、各データ項目に割り付けて保存→管理者チェックを行った後 csv ファイルで出力。いう植物標本 OCR 読みとりシステムを構築した。システム構築は所属館の維持運営費で行ったが、原理は本研究の成果を応用している。本システムは Google Chrome で動くブラウザベースのアプリという形で実現したのだが、コロナ禍で資料整

理のアルバイトさんに在宅勤務をお願いせざるを得ない時期は、自宅でネットにつないでアプリを立ち上げ、標本画像を見ながら OCR 抽出データを確認して入力作業を行うことが可能になり、自身の研究成果に大いに助けられた。

③国立科学博物館をはじめ日本国内各植物標本庫が所蔵する標本画像 50 万点余りを用いて、日本産植物の 2171 分類群を 91%の確率で種判定できるシステムを構築した。本研究の結果、虫食いの少ないきれいな標本画像を 50 点以上用意出来れば高い判定確率をだせること、人間が間違いを起こしやすい分類群については AI も判定を迷うことがわかった。また Grad-CAM 解析により、AI がどこを見て植物の判定をしているかを調べたところ、人間が判定につかう場所をよく見ていることがわかった。一連の成果は Scientific Reports に掲載された。またバイオサイエンスとインダストリー誌に和文抄訳を掲載した。開発した植物標本自動種判定システムは島根大学の HP 上で公開されている（http://tayousei.life.shimane-u.ac.jp/ai/index_all.php）。

④2 種類のテストデータセット、3 つの AI モデルで 4 通りの学習方法を行った結果を Table 3 に示し、「3.手作業・サンプリングデータを同時に学習」の最良の計測の詳細を Table 4 に記載した。BERT, Albert の結果において、F 値が低いものは計算時に丸め処理でゼロ除算が発生するので割愛している。1～3 の学習方法を行った時点で Albert の成績が良くなかったため、4 の学習は Albert においては割愛した。
学習方式の中では、「3.手作業・サンプリングデータを同時に学習」が最も高かった。4.の「 サンプリングデータのみで学習の後、最良結果をロードし手作業のコーパスを学習」をしてしまうと、SpaCy では特に先に学習させたサンプリングのスコアが落ちる結果となった。エポックを 20 と高い数値で学習させたが、計測の数値を見たところ 10 エポックぐらいで最大 F 値の-0.02 辺りまで達し、その後も±0.02 辺りを上下しているので、過学習は起こしていないと思われる。学習の結果は BERT の F 値が一番高かったが、BERT を動作させるには GPU 搭載のマシンが必要となる。 SpaCy でも F 値 0.8 以上の結果が出ており、かつ安価なレンタルサーバー上でも動作させることが可能になることから、プログラムの汎用性を考えて SpaCy を採用し、システム構築を行った。開発したアプリケーションのデモは URL（https://youtu.be/2jt_GMUqrWQ）で閲覧可能である。本研究の成果は 2023 年 3 月の日本植物分類学会で発表し、現在論文投稿中である。
⑤我々が開発した標本撮影装置は、本研究期間の 2019―2023 年の間に、岩手大学附属博物館、東京大学総合研究博物館、東京大学附属小石川植物園、京都大学総合博物館、大阪市立自然史博物館に、それぞれの施設で使用可能な部屋面積に合わせた形にアレンジされ導入された。2022 年の講習会後、伊丹市立ミュージアム、神奈川県立生命の星・地球博物館、東京都立大学牧野標本館への導入も決まった。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Masato Shirai, Atsuko Takano, Takahide Kurosawa, Masahito Inoue, Shuichiro Tagane, Tomoya Tanimoto, Tohru Koganeyama, Hirayuki Sato, Tomohiko Terasawa, Takehito Horie, Isao Mandai, Takashi Akihiro	４．巻 -
２．論文標題 Development of a system for the automated identification of herbarium specimens with high accuracy.	５．発行年 2022年
３．雑誌名 Scientific Reports	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-022-11450-y	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 高野 温子、堀内 保彦、青木 滉太、藤本 悠、三橋 弘宗	４．巻 68
２．論文標題 植物標本デジタル画像化とOCRによるラベルデータ自動読みとり手法の開発	５．発行年 2020年
３．雑誌名 植物地理・分類研究	６．最初と最後の頁 103～119
掲載論文のDOI（デジタルオブジェクト識別子） 10.18942/chiribunrui.0682-05	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 髙野温子・秋廣高志	４．巻 81
２．論文標題 標本画像を用いた植物種のAI同定システム	５．発行年 2022年
３．雑誌名 バイオサイエンスとインダストリ	６．最初と最後の頁 19‐21
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 秋廣高志・白井匡人・高野温子・黒沢高秀・井上雅仁・田金秀一郎・谷本朋也・小金山透・佐藤平行・寺澤知彦・堀江岳人・萬代功.
２．発表標題 AI画像認識技術を使って約2200種の植物の名前を高精度(96%)に判定するシステムの開発.
３．学会等名 日本植物分類学会
４．発表年 2022年

１．発表者名
髙野温子・小長井元

２．発表標題
自然言語処理技術を用いた植物標本ラベルデータ自動抽出法の 開発

３．学会等名
日本植物分類学会

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
ひとはく研究員個人業績紹介
https://www.hitohaku.jp/researchers/takano.html
小・中規模植物標本庫に適用可能な、簡便・低予算で最低限の画質を担保した植物標本画像撮影方法の開発
https://www.hitohaku.jp/research/h-research/2019.html#2019-06-takano
ひとはく資料の管理と活用
https://www.hitohaku.jp/material/innovation.html
植物標本デジタル化と植物標本画像からのラベル自動読み取りシステムの開発
https://www.hitohaku.jp/exhibition/planning/2-4_2020-takano.pdf

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	三橋　弘宗 (Mitsuhashi Hiromune) (50311486)	兵庫県立大学・自然・環境科学研究所・講師 (24506)	
研究分担者	藤本　悠 (Fujimoto Yu) (50609534)	芸術文化観光専門職大学・芸術文化・観光学部・准教授 (24507)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関