493

Improving the Character of Optical Character Recognition (OCR): iDigBio Augmenting OCR Working Group Seeks Collaborators and Strategies to Improve OCR Output and Parsing of OCR Output for Faster, More Efficient, Cheaper Natural History Collections Specimen Label Digitization

Robert Anglin
North American Bryophyte and Lichen TCN/Symbiota
Jason Best
Botanical Research Institute of Texas (BRIT)/Biodiversity Informatics
Renato Figueiredo
University of Florida/iDigBio
Edward Gilbert
North American Bryophyte and Lichen TCN/Symbiota
Nathan Gnanasambandam
Xerox Research Center Webster
Stephen Gottschalk
New York Botanical Garden
Elspeth Haston
Royal Botanic Garden Edinburgh
P. Bryan Heidorn
University of Arizona/School of Information Resources and Library Science
Daryl Lafferty
Arizona State University/SALIX
Peter Lang
ABBYY USA
Gil Nelson
Florida State University/Institute for Digital Information (iDigInfo)
Deborah Paul
Florida State University/Institute for Digital Information (iDigInfo)
William Ulate
Missouri Botanical Garden/ Biodiversity Heritage Library
Kimberly Watson
New York Botanical Garden
Qianjin Zhang
University of Arizona/School of Information Resources and Library Science

Museums across the U. S. have been seeking new ways to cost effectively transcribe the label information associated with their specimen collections in order to maximize everyone's access to valuable research data. Digitization methods are often relatively slow, labor-intensive, and expensive.

Improving the Character of Optical Character Recognition (OCR)

New methods, such as OCR, are being explored to reduce these costs. The National Science Foundation (NSF), through the Advancing Digitization of Biological Collections (ADBC) program, funded Integrated Digitized Biocollections (iDigBio) in 2011 to create a Home Uniting Biodiversity Collections (HUB) cyberinfrastructure to collectively integrate specimen data, in order to address specifically-proposed, timely research themes such as global warming and climate change, species discovery, and species-host-parasite relationships. Museums are funded by ADBC as Thematic Collection Networks (TCNs) to digitize and aggregate their collective specimen data addressing these themes for inclusion in the iDigBio database.

THE CHALLENGE
Since much of the to-be-captured data resides on museum specimen labels or in field notebooks as print, type written, or hand-written text, better OCR, image processing, machine language (ML) and natural language processing (NLP) strategies increase the chances of meeting our goals. There is room for improvement in parsing, autocorrection, text and handwriting recognition, and image segmentation.

iDigBio seeks to help the biodiversity collections community find ways to:

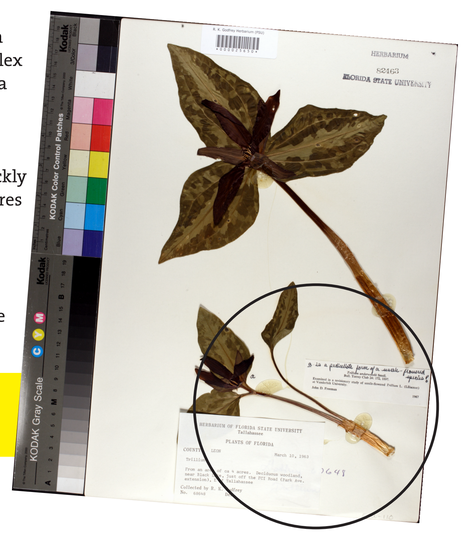
- speed up digitization
- lower the cost
- improve efficiency
- assure digitized data is fit-for-use*
- provide data to researchers more quickly

*NIBA 2010, Chapman 2005

The iDigBio Augmenting OCR (A-OCR) working group, formed in March of 2012, is actively engaged in identifying opportunities to leverage OCR tools and technologies that are successful and disseminate these tools to the public. The A-OCR working group would like to integrate these tools, or seek funding for tool development.

improving automated image segmentation

Improving automated image segmentation involves identifying the text block in complex images such as an herbarium specimen or a full tray image of insects. The sample herbarium sheet image here exemplifies the complexities of the task. Here the goal would be to develop an algorithm that quickly and correctly recognizes the label and ignores the plant. This would enable OCR of these objects to skip image-processing steps currently used like taking a separate image of just the label or using humans to crop the image by hand or indicate (segment) where the label is on a sheet.



Typical 11x17 Herbarium Sheet from Florida State University, Robert K. Godfrey Herbarium. (Mast, et al., 2012)

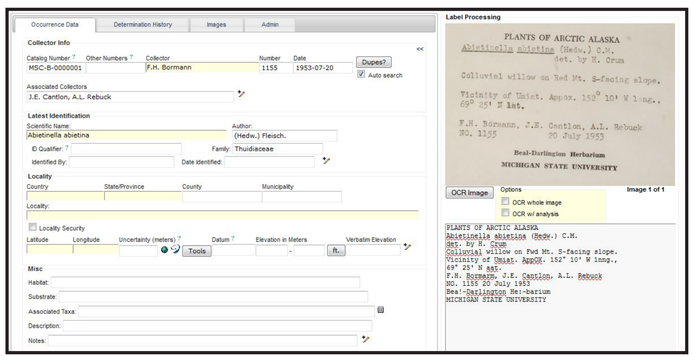
Another issue of interest involves developing algorithms that differentiate and classify image segments by successfully figuring out which section contains the primary label, the annotation label (if any), the herbarium stamp, the collecting event label (refers to insect specimens), or other text that may exist on the specimen. Once recognized, segmented OCR output is parsed into fields based on a data standard like Darwin Core for automated insertion into a database.

Only some label types, mainly those printed, and some typed, result in OCR output suitable for this type of parsing. Here's an example of such a label and its parsed data.

Label suitable for effective OCR from Herbarium of Yale University (Used with permission) and parsed formatted OCR output from HERBIS/LABELX system (Heidorn 2008)

```
<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="http://www3.isrl.uiuc.edu/~TeleNature/Herbis/semanticrelax.rng" type="xml"?>
<labeldata>
<bt>Yale University Herbarium</bt>
<bc>YU.010782</bc>
<in>Herbarium of Yale University</in>
<hdlc>Plants of Puerto Rico</hdlc><cnl>No. </cnl><cn cc="156.">156- </cn><fml>Family: </fml>
<fm cc="Q. Polypodiaceae">Q- Polypodiaceae</fm>
<in>Scientific isjamp- Adiantum latifolium</in>
<cml>Common Name:</cml>
<lcl>Locality: </lcl><lc>Mahoe plot 1-3, Rio Abajo State Forest</lc>
<hb>Habitat:</hb>
<ftl>Comments:</ftl>
<col>Collector: </col><co>Mark Ashton and</co>
<co>J.S. Lowe</co>
<cdl>Date: </cdl><cd>17 July 1934</cd>
</labeldata>
```

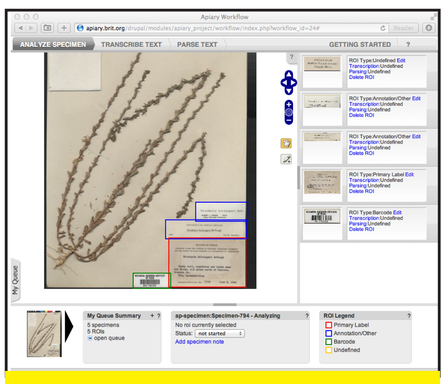
The *North American Bryophyte and Lichen* TCN (LBCC) has a goal of digitizing 2.3 million lichen and bryophyte specimens representing well over 90% of North American specimens. To achieve this goal, LBCC has integrated OCR and NLP capabilities into their processing workflows and their Symbiota web portals. Symbiota (*http://symbiota.org*) is open source software designed to aid biologists in establishing specimen-based public data portals. LBCC is making use of a suite of specimen management tools integrated into the basic user interface that supports the digitization of specimen information directly from the images of the specimen labels.



high-throughput workflow

The Apiary Project (*http://www.apiaryproject.org/*) is a collaborative effort between the *Botanical Research Institute of Texas* (*http://www.brit.org*) and the *Texas Center for Digital Knowledge* (*http://txcdk.unt.edu/*) at the *University of North Texas*, funded by the U. S. Institute of Museum and Library Services. The goal of the project is to provide a high-throughput workflow for *computer-assisted human parsing* of biological specimen label data. The web-based workflow consists of three primary stages:

- Image region delineation and categorization
- Text transcription (human and OCR)
- Text parsing (human and NLP) into Darwin Core elements (*Wieczorek et al., 2012*)



Apiary interface to classify regions

faster data capture & higher accuracy

The herbaria at the *Royal Botanic Garden Edinburgh* (RBGE) and *The New York Botanical Garden* (NY) use a more recently developed workflow:

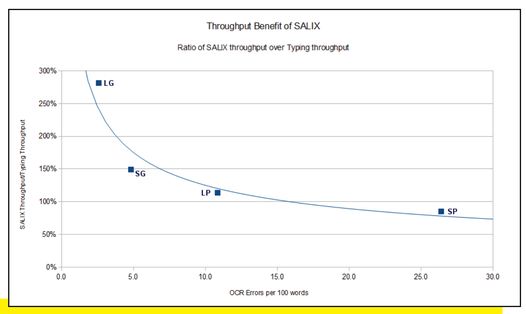
1. Capture minimal data for each specimen (e.g. barcode, geographic region, and the taxon name on the specimen folder)
2. Capture high quality digital images of each specimen
3. Process images with ABBYY® OCR software
4. Develop and/or use software tools to search OCR text output and sort the images/data based on principal data elements (e.g. by collector and country)
5. Sort image/data to enable *faster data capture & higher accuracy* by data transcriptionists, including duplicate record matching.

Both institutes aim to digitize their entire collections. NY holds 7 million specimens of which 680,000 have now been imaged. RBGE holds nearly 3 million specimens of which 200,000 have now been imaged. The *biggest challenge* remains the automated recognition and parsing of principal collection label data elements (e.g. collector, number, date, country) from the OCR text output.

30% faster than typing

Access to biodiversity data is limited by the time and resources required to digitize, but we have found that it is possible to do so at a faster rate. Data entry speed is dependent on user proficiency, label quality, and to a lesser degree, label length. The most recent project at the *Arizona State University Herbarium* involves the digitization - both the imaging and databasing - of approximately 55,000 vascular plant specimens from Latin America. SALIX (semi-automatic label information extraction) was developed as the central tool to handle automatic parsing, along with BarcodeRenamer (BCR) to automate image file renaming by barcode. These two developments, combined with some existing software, make up the SALIX Method.

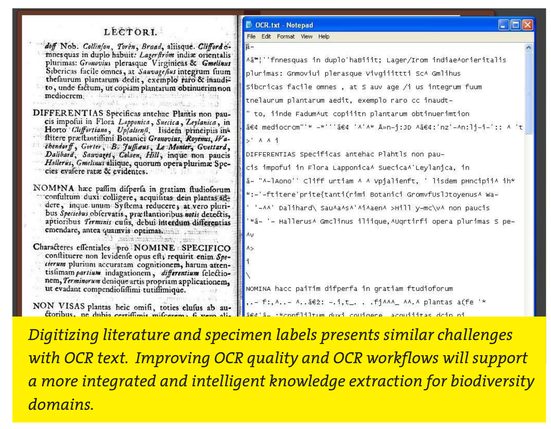
The SALIX Method provides a way to digitize herbarium specimens more efficiently than the traditional approach of typing data by hand. Using digital imaging, optical character recognition, and automatic parsing, ASU has found that the SALIX Method can process data at an average rate that is 30% faster than typing.



LG = long label, good text; SG: short label, good text; LP: long label, poor text; SP: short label, poor text

tagging and knowledge extraction

The Biodiversity Heritage Library (BHL) is a consortium of natural history and botanical libraries that cooperates with the international taxonomic community, rights holders, and other interested parties to digitize and make accessible the legacy literature of biodiversity held in their collections for open access and responsible use as a part of a global 'biodiversity commons. In partnership with the Internet Archive and through local digitization efforts, the BHL has digitized more than 40 million pages of biodiversity literature, representing tens of thousands of titles and over 110,000 volumes. As seen in the figure below, most challenges with label digitization, like font types, italics, bold, stains, spots, blurs, marks, image segmentation, handwritten notes and layouts, among others, are also common content from books, journals and field notebooks. Improving OCR workflow and reducing OCR errors will support a more integrated and intelligent knowledge extraction by allowing the creation of better parsing trees, controlled vocabularies and ontologies for biodiversity domains. At large, these tools will facilitate the conversion from legacy information to computable data useful for biological research activities like the generation of specimen identification keys, evolutionary trees and phenotype correlations.



Digitizing literature and specimen labels presents similar challenges with OCR text. Improving OCR quality and OCR workflows will support a more integrated and intelligent knowledge extraction for biodiversity domains.

Finally, a key aspect of the iDigBio cyberinfrastructure is the ability to provide cloud-oriented services to its users. In the context of OCR workflows, these services can include common *Web-based services* hosted by iDigBio and academic or commercial partners, as well as providing users and developers with the ability to develop, configure, package and disseminate new and experimental services by creating virtual appliances. Virtual appliances are preconfigured, ready-to-use "virtual machines" that include all the complex software and configuration needed for an OCR tool or workflow (operating systems, applications, libraries, scripts, etc) in a manner that allows the appliance to be instantiated by end users on their own computers, and/or hosted in the iDigBio cloud infrastructure.

The digitization of natural history museum collections data poses a serious challenge to researchers in image processing, OCR, text classification and segmentation, workflow methods and human computer interaction. The iDigBio Augmented OCR working group is actively seeking collaborators and new sustainable approaches to these problems. Contact any AOCR member to get involved. We need your collective energy and knowledge, from graduate students, programmers and professors to commercial companies ~ all are needed and welcome. Comments and collaboration anticipated and appreciated!

REFERENCES
Chapman, A. D. (2005). Uses of primary species-occurrence data, (version 1.0). 100 pp. Report for the Global Biodiversity Information Facility, Copenhagen. Retrieved from http://www.gbif.org/orc/?doc_id=1300

Heidorn, P. B., Wei, Q. (2008). Automatic Metadata Extraction from Museum Specimen Labels. In Greenberg, J., Klas, W. (Eds.), Proceedings of the International Conference on Dublin Core and Metadata Applications Berlin, 22-26 September 2008 DC 2008: Berlin, Germany. Retrieved from http://hdl.handle.net/2142/9138

Mast, A. R., Stuy, A., Nelson, G., Bugher, A., Weddington, N., Vega, J., Weismantel, K., Feller, D. S., Paul, D. (2004) onward (continuously updated). Database of Florida State University's Robert K. Godfrey Herbarium. Website http://herbarium.bio.fsu.edu/ [accessed 10 October 2012].

NIBA. (2010). A Strategic Plan for Establishing a Network Integrated Collections Alliance. Network Integrated Biocollections Alliance. Retrieved from http://digbiocol.files.wordpress.com/2010/08/niba_brochure.pdf

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., et al. (2012). Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. doi:10.1371/journal.pone.0029715

iDigBio is graciously funded by a grant from the National Science Foundation's Advancing Digitization of Biological Collections Program (#EF1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Many kind thanks to all members of the iDigBio Augmenting OCR working group for their efforts in putting this material together. A sincere thank you to all involved in iConference 2013 who have provided us with the opportunity to reach out to the Information Science community.

iDigBio Integrated Digitized Biocollections