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
ranglin@wisc.edu

Jason Best
Botanical Research Institute of Texas (BRIT)/Biodiversity Informatics
jbest@brit.org

Renato Figueiredo
University of Florida/iDigBio
renatof@ufl.edu

Edward Gilbert
North American Bryophyte and Lichen TCN/Symbiota
egbiodiversity@gmail.com

Nathan Gnanasambandam
Xerox Research Center Webster
nathan.gnanasambandam@xerox.com

Stephen Gottschalk
New York Botanical Garden
sgottschalk@nybg.org

Elspeth Haston
Royal Botanic Garden Edinburgh
e.haston@rbge.ac.uk

P. Bryan Heidorn
University of Arizona/School of Information Resources and Library Science
heidorn@email.arizona.edu

Daryl Lafferty
Arizona State University/SALIX
daryl@daryllafferty.com

Peter Lang
ABBYY USA
peter.lang@abbyyusa.com

Gil Nelson
Florida State University/Institute for Digital Information (iDigInfo)
gnelson@bio.fsu.edu

Deborah Paul
Florida State University/Institute for Digital Information (iDigInfo)
dpaul@fsu.edu

William Ulate
Missouri Botanical Garden/ Biodiversity Heritage Library
william.ulate@mobot.org

Kimberly Watson
New York Botanical Garden
kwatson@nybg.org

Qianjin Zhang
University of Arizona/School of Information Resources and Library Science
qianjinzhang@email.arizona.edu

Abstract

There are an estimated 2 – 3 billion museum specimens world – wide (OECD 1999, Ariño 2010). In an effort to increase the research value of their collections, institutions across the U. S. have been seeking new ways to cost effectively transcribe the label information associated with these specimen collections. Current digitization methods are still relatively slow, labor-intensive, and therefore expensive. New methods, such as optical character recognition (OCR), natural language processing, and human-in-the-loop assisted parsing are being explored to reduce these costs. The National Science Foundation (NSF), through the Advancing Digitization of Biodiversity Collections (ADBC) program, funded Integrated Digitized Biocollections (iDigBio) in 2011 to create a Home Uniting Biodiversity Collections (HUB) cyberinfrastructure to aggregate and collectively integrate specimen data and find ways to digitize specimen data faithfully and faster and disseminate the knowledge of how to achieve this. The iDigBio Augmenting OCR Working Group is part of this national effort.

Keywords: iDigBio, OCR, natural language, information analysis, machine language

Anglin, R., Best, J., Figueiredo, R., Gilbert, E., Gnanasambandam, N., Gottschalk, S., Haston, E., Heidorn, P. B., Lafferty, D., Lang, P., Nelson, G., Paul, D., Ulate, W., Watson, K., & Zhang, Q. (2013). 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. iConference 2013 Proceedings (pp.957-964).doi:10.9776/13493

Introduction

While optical character recognition (OCR) is currently utilized by some museums in their databasing workflows, better OCR strategies would increase the chances of meeting the following goals. Part of iDigBio's mission is to assist the biodiversity collections community in finding ways to:

- speed up the overall digitization process,
- lower the cost,
- improve overall efficiency,
- assure digitized data is fit-for-use (NIBA 2010, Chapman 2005), and
- provide the resulting digitized data records to researchers more quickly.

Some Projects and Challenges of the A-OCR Working Group

Those currently using OCR note there is also much room for improvement in issues including parsing of the output, autocorrection of text, recognition of text, recognition of handwriting and image segmentation. 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 (both within and outside of the biology digitization domain) and disseminate these tools, methods and workflows to the public. The A-OCR working group would like to integrate these tools, or seek funding for tool development.

Natural history museums contain a wealth of specimen data currently only accessible to those with the time, resources and permissions necessary to travel to the museums and walk through the research collections. Since most inventories are not accessible via the web, it is difficult for a researcher to ascertain where important specimens might exist. Collections vary in size from a few thousand specimens in research universities to many millions in the major natural history museums of the world. As part of a national, coordinated, multi-faceted effort to collate, integrate and expose this so-called "dark data" through a cyberinfrastructure hub, the National Science Foundation (NSF) started the Advancing Digitization of Biodiversity Collections (ADBC) program which then funded Integrated Digitized Biocollections, or iDigBio, to build this cloud-based database resource.

The data comes from NSF-funded Thematic Collection Networks (TCNs). The TCNs, made up of groups of museums, are funded to collect data from defined specimen groups in order to address specifically-proposed, timely research themes such as global warming and climate change, species discovery, and species-host-parasite relationships. Besides building an agile cloud-based system to facilitate synthesizing diverse museum collection data sets for research, iDigBio's goals include working with TCNs, natural history collections, and the broader community to look for ways to produce fit-for-research-use research data quicker and cheaper.

Since much of the to-be-captured data resides on museum specimen labels or in field notebooks as print, type-written text or hand-writing, OCR, algorithms for parsing OCR output, and efficient user interfaces for these tasks are natural targets for improvement in attempts to hasten data capture and insertion of that data into databases. The iDigBio Augmenting OCR Working Group (A-OCR) formed in March of 2012 and after outlining possible goals, held its first workshop on October 1 - 2, 2012 in Gainesville, Florida to:

- build a strategic plan for broader community engagement in our endeavors,
- combine our collective knowledge and experience with current OCR software and parsing strategies to produce website content at iDigBio for use by anyone seeking effective OCR practices when digitizing museum specimens,
- choose hackathon goals for our first iDigBio Augmenting OCR hackathon being held and hosted at the Botanical Research Institute of Texas (BRIT) concurrent with this 2013 iConference, and
- learn about recent developments in OCR, handwriting recognition, and OCR output parsing from the broader community and our working group members.

Each member of our working group brings knowledge and experience from unique uses of OCR and OCR output. As a group, we collected all the issues we would like to work on, for example: improving automated image segmentation. This 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 in figure 1 (Figure 1) 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.

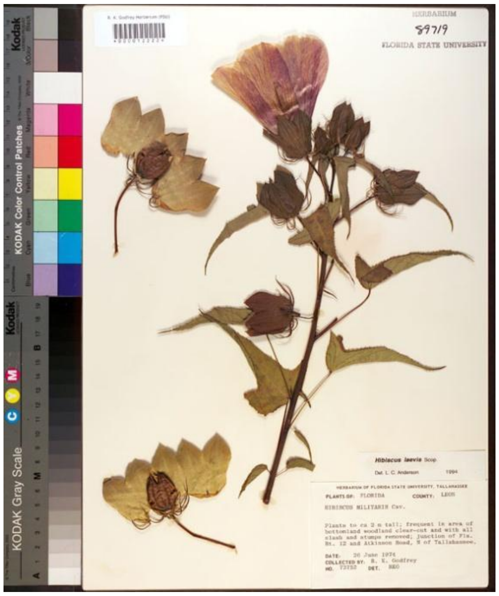
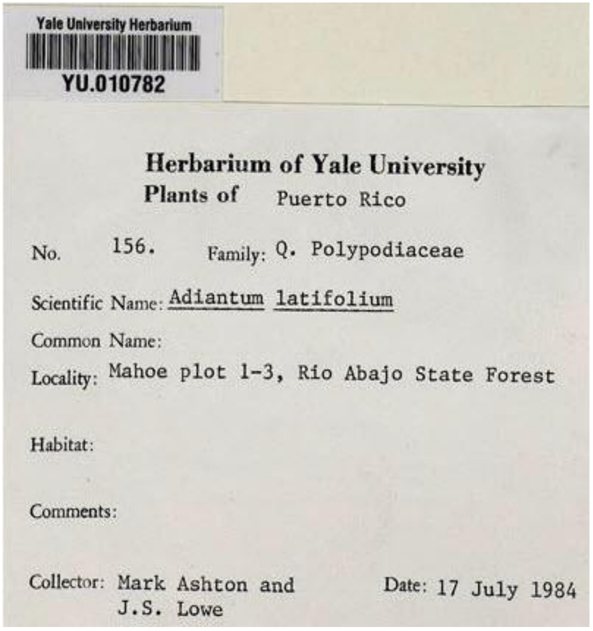

HERBARIUM
89719
FLORIDA STATE UNIVERSITY
KODAK Color Control Patches
KODAK Gray Scale
Hibiscus laevis Scop.
Det. L. C. Anderson 1994
HERBARIUM OF FLORIDA STATE UNIVERSITY, TALLAHASSEE
PLANTS OF: FLORIDA COUNTY: LEON
HIBISCUS MILITARIS Cav.
Plants to ca 2 m tall; frequent in area of bottomland woodland clear-cut and with all slash and stumps removed; junction of Fla. Rt. 12 and Atkinson Road, N of Tallahassee.
DATE: 26 June 1974
COLLECTED BY: R. K. Godfrey
NO. 73732 DET. RKG

Figure 1. Herbarium Sheet, Florida State University, Robert K. Godfrey Herbarium. Used with permission.

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 (Figure 2) and its parsed data.



Yale University Herbarium
YU.010782

Herbarium of Yale University
Plants of Puerto Rico

No. 156. Family: Q. Polypodiaceae

Scientific Name: Adiantum latifolium

Common Name:

Locality: Mahoe plot 1-3, Rio Abajo State Forest

Habitat:

Comments:

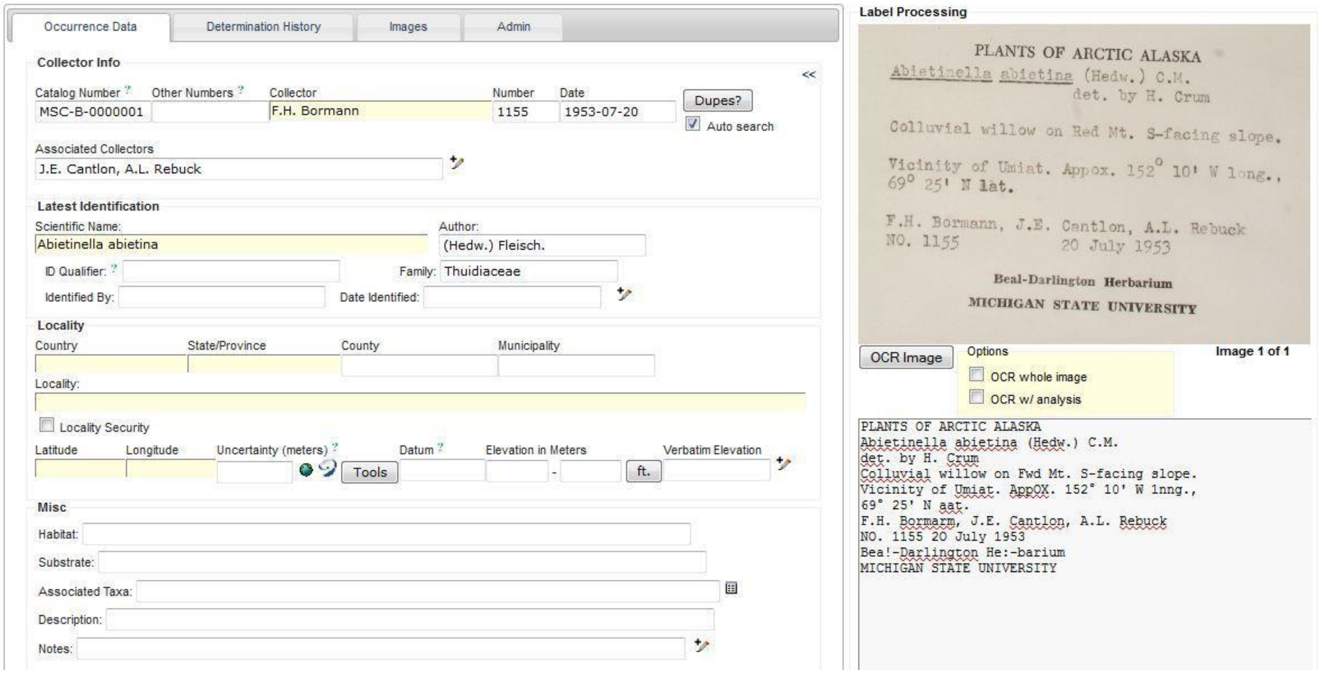
Collector: Mark Ashton and J.S. Lowe Date: 17 July 1984

Figure 2. Label suitable for effective OCR. Herbarium of Yale University. Used with permission.

Parsed formatted OCR output of label in figure 2 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 (Figure 3) that supports the digitization of specimen information directly from the images of the specimen labels (Figure 4).



Collector Info

Catalog Number: MSC-B-0000001
Other Numbers:
Collector: F.H. Bormann
Number: 1155
Date: 1953-07-20
☑ Auto search
Associated Collectors: J.E. Cantlon, A.L. Rebuck

Latest Identification

Scientific Name: Abietinella abietina
Author: (Hedw.) Fleisch.
ID Qualifier:
Family: Thuidiaceae
Identified By:
Date Identified:

Locality

Country:
State/Province:
County:
Municipality:
Locality:
□ Locality Security
Latitude:
Longitude:
Uncertainty (meters):
Datum:
Elevation in Meters:
Verbatim Elevation:

Misc

Habitat:
Substrate:
Associated Taxa:
Description:
Notes:

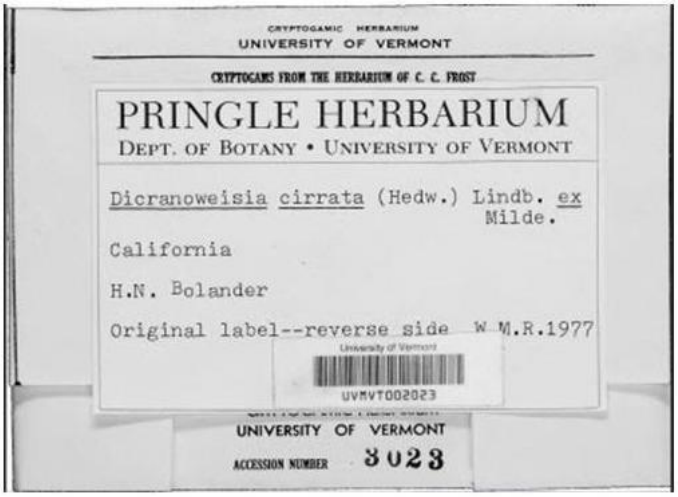
Label Processing

PLANTS OF ARCTIC ALASKA
Abietinella abietina (Hedw.) C.M.
det. by H. Crum
Colluvial willow on Red Mt. S-facing slope.
Vicinity of Umiat. Appox. 152° 10' W long.,
69° 25' N lat.
F.H. Bormann, J.E. Cantlon, A.L. Rebuck
NO. 1155 20 July 1953
Beal-Darlington Herbarium
MICHIGAN STATE UNIVERSITY

Options
□ OCR whole image
□ OCR w/ analysis
Image 1 of 1

```
PLANTS OF ARCTIC ALASKA
Abietinella abietina (Hedw.) C.M.
det. by H. Crum
Colluvial willow on Fwd Mt. S-facing slope.
Vicinity of Umiat. AppOX. 152° 10' W 1nng.,
69° 25' N aat.
F.H. Bormarm, J.E. Cantlon, A.L. Rebuck
NO. 1155 20 July 1953
Bea!-Darlington He:-barium
MICHIGAN STATE UNIVERSITY
```

Figure 3. Symbiota user interface. Note display of data record, image of label and ocr output.



CRYPTOGAMIC HERBARIUM
UNIVERSITY OF VERMONT

CRYPTOGAMS FROM THE HERBARIUM OF C. C. FROST

PRINGLE HERBARIUM
DEPT. OF BOTANY • UNIVERSITY OF VERMONT

Dicranoweisia cirrata (Hedw.) Lindb. ex Milde.

California

H.N. Bolander

Original label--reverse side W M.R.1977

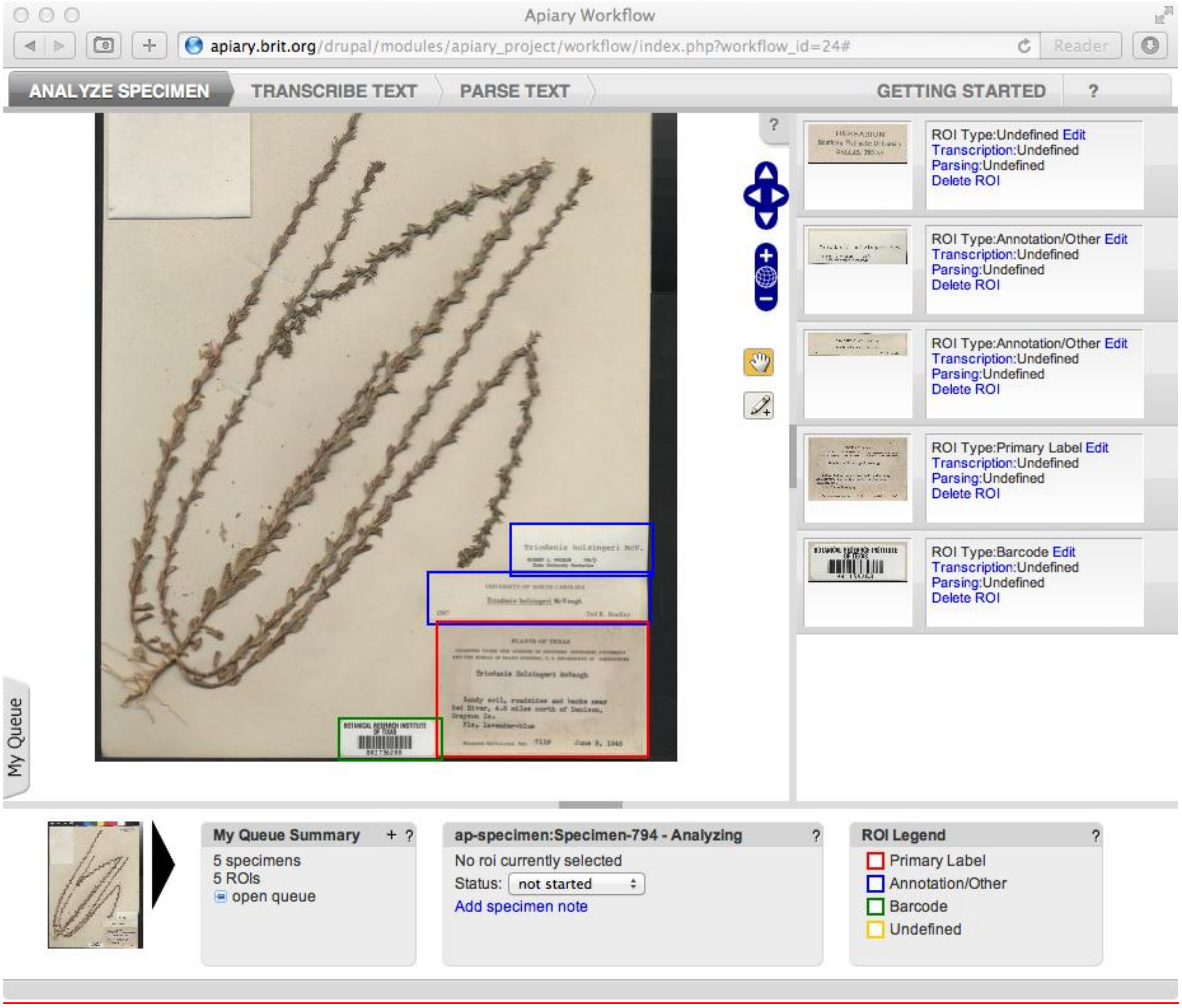
University of Vermont
UVMVT002023

UNIVERSITY OF VERMONT
ACCESSION NUMBER 3023

Figure 4. Bryophyte label typical in the LBCC project. University of Vermont, Pringle Herbarium. Used with permission.

While OCR, NLP, duplicate harvesting, and concepts of crowdsourcing have been integrated into the working model, the LBCC project continues to work on increasing efficiency and improving performance of these tools.

The Apiary Project (http://www.apiaryproject.org/) is a collaborative effort between the Botanical Research Institute of Texas and the Texas Center for Digital Knowledge (http://txcdk.unt.edu/) at the University of North Texas with the goal of providing a high-throughput workflow for computer-assisted human parsing of biological specimen label data. The Apiary workflow utilizes a three-stage process for extracting parsed text from digital images of herbarium specimens. This workflow provides a user interface through a web-based application. In the first stage, users view the full specimen image and delineate and classify image regions that contain textual content (Figure 5).



ROI Type:Undefined Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

ROI Type:Annotation/Other Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

ROI Type:Annotation/Other Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

ROI Type:Primary Label Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

ROI Type:Barcode Edit
Transcription:Undefined
Parsing:Undefined
Delete ROI

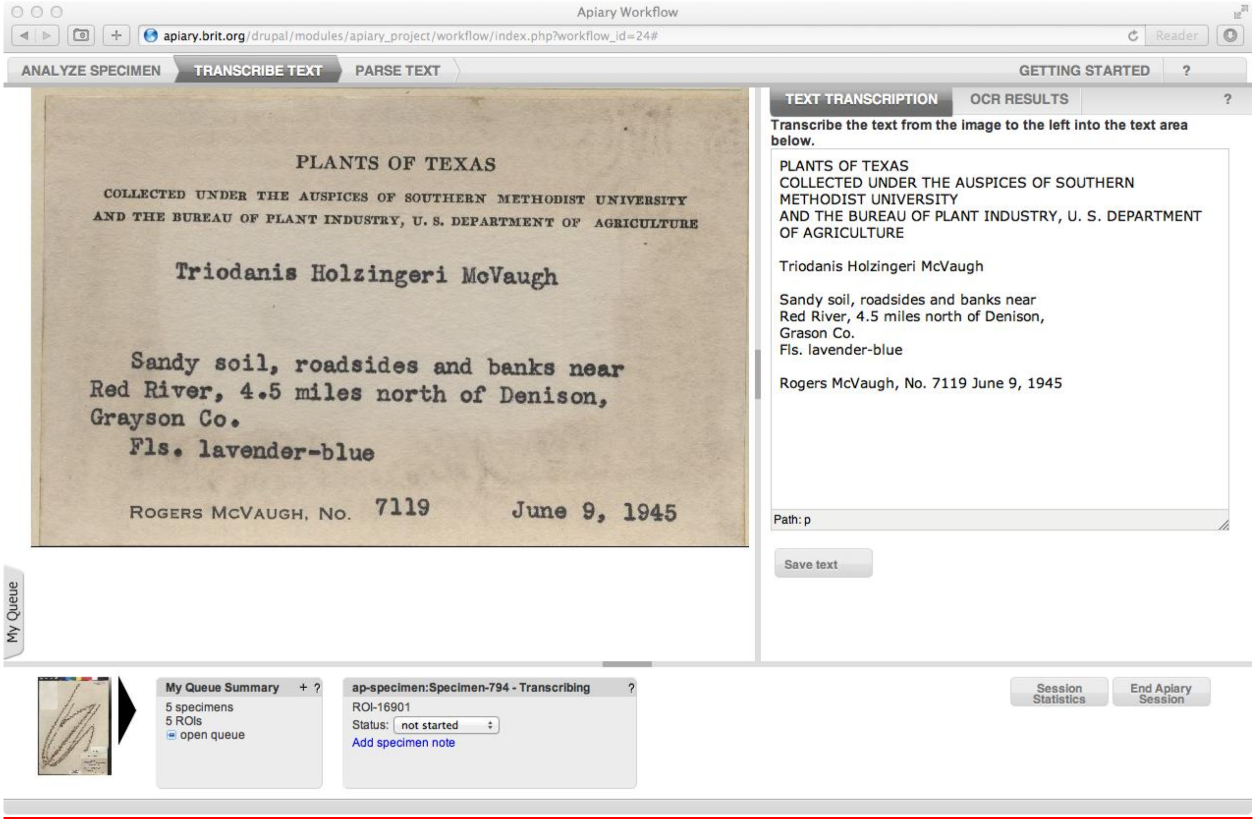
My Queue Summary + ?
5 specimens
5 ROIs
open queue

ap-specimen:Specimen-794 - Analyzing ?
No roi currently selected
Status: not started
Add specimen note

ROI Legend ?
🟥 Primary Label
🟦 Annotation/Other
🟩 Barcode
🟨 Undefined

Figure 5. Apiary interface to classify regions

In the next phase, these regions are processed by three OCR processes and the user is able to select the most accurate output. When the text output is not accurate, the user may make corrections or, as often is the case with handwritten labels, disregard the OCR output and transcribe the complete text of the region (Figure 6). Once the transcription is complete, the text is parsed into Darwin Core fields (Wieczorek et al., 2012) using controlled vocabularies and interface devices to help standardize and normalize the parsed record.



PLANTS OF TEXAS
COLLECTED UNDER THE AUSPICES OF SOUTHERN METHODIST UNIVERSITY
AND THE BUREAU OF PLANT INDUSTRY, U. S. DEPARTMENT OF AGRICULTURE

Triodanis Holzingeri McVaugh

Sandy soil, roadsides and banks near
Red River, 4.5 miles north of Denison,
Grayson Co.
Fls. lavender-blue

ROGERS MCVAUGH, NO. 7119 June 9, 1945

Transcribe the text from the image to the left into the text area below.

PLANTS OF TEXAS
COLLECTED UNDER THE AUSPICES OF SOUTHERN
METHODIST UNIVERSITY
AND THE BUREAU OF PLANT INDUSTRY, U. S. DEPARTMENT
OF AGRICULTURE

Triodanis Holzingeri McVaugh

Sandy soil, roadsides and banks near
Red River, 4.5 miles north of Denison,
Grason Co.
Fls. lavender-blue

Rogers McVaugh, No. 7119 June 9, 1945

Path: p

My Queue Summary + ?
5 specimens
5 ROIs
open queue

ap-specimen:Specimen-794 - Transcribing ?
ROI-16901
Status: not started
Add specimen note

Figure 6. Apiary transcription interface. Note label and transcribed output on the right.

Next, 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 pre-configured, 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.

Conclusion

We actively encourage you to contact any member of the iDigBio Augmenting OCR working group 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!

Acknowledgements: iDigBio is graciously funded by a grant from the National Science Foundation's Advancing Digitization of Biological Collections Program (#EF1115210). To each and every member of the iDigBio Augmenting Optical Character Recognition working group, many kind thanks. All of us in the working group would like to express our gratitude to the 2013 iConference for the education and outreach opportunities given to our working group at this gathering of the iSchools community.

References

Ariño, A. H. (2010). Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics,* 7, 81-92. Retrieved from https://journals.ku.edu/index.php/jbi/article/viewFile/3991/3805

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

OECD. (1999). *OECD Megascience Working Group - Biological Informatics - Final Report.* 74 pp. Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd.org/dataoecd/24/32/2105199.pdf

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

Wei, Q., Heidorn, P. B., Freeland, C. (2010). *Name Matters: Taxonomic Name Recognition (TNR) in Biodiversity Heritage Library (BHL.)* Retrieved from http://hdl.handle.net/2142/14919

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