Augmenting Optical Character Recognition (OCR) for Improved Digitization: Strategies to Access Scientific Data in Natural History Collections

Deborah L. Paul
Florida State University
iDigInfo, iDigBio
dpaul@fsu.edu

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

Abstract

The Augmenting OCR Working Group (A-OCR WG) at Integrated Digitized Biocollections (iDigBio) seeks to improve community OCR strategies and algorithms for faster, better parsing of OCR output derived from valuable data on natural history collection specimen labels. This task is exceedingly difficult because museum labels are often annotated, and vary in content, form and font. Under the National Science Foundation's (NSF) Advancing Digitization of Biological Collections (ADBC) program, iDigBio is building a cyberinfrastructure to aggregate quality data from museum specimens housed in collections across the United States for use by researchers, educators, environmentalists and the public. Since March of 2012, the A-OCR WG formed from community consensus to begin its role in this endeavor, defining reachable goals including setting up a hackathon concurrent with iConference 2013. This paper reports on the definition of some key problems identified by the A-OCR WG since these science problems will drive research and cyberinfrastructure development.

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

Introduction

iDigBio is a NSF Project under the Advancing Digitization of Biological Collections (ADBC) program. We at iDigBio are building a cloud-based cyberinfrastructure to aggregate United States vouchered specimen data across biological and paleontological collections. Natural history collections labels contain vital primary data about the specimens including for example, an assigned scientific name, location and date of collection, name of the collector, a collector identifier for the specimen, a museum identifier, and sometimes description of the specimen and the environment from which it was collected (NIBA, 2010) (http://digbiocol.files.wordpress.com/2010/08/niba_brochure.pdf).

This information has a broad range of scientific uses such as source data for ecological niche or historic species distribution models. The scope of this 10-year project requires innovation to succeed as it is estimated there are well over 2 - 3 billion specimens in the world (OECD, 1999; Ariño, 2010). NSF and the broader community recognizes the digitization processes currently in use to capture this data need to be faster and more efficient to meet difficult challenges facing science and society (Chapman, 2005; Blagoderov, 2012).

Acknowledgements: iDigBio and the efforts of the A-OCR WG are supported by NSF Award EF-1115210. For more about the origins of iDigBio and the ADBC program see A Strategic Plan for Establishing a Network Integrated Biocollections Alliance (NIBA, 2010). Many kind thanks to all the iDigBio Augmenting OCR Working Group members for contributions and hard work to date. Working Group members listed here in alphabetical order: Robert Anglin, Jason Best, Renato Figueiredo, Edward Gilbert, Nathan Gnanasambandam, Stephen Gottschalk, Elspeth Haston, Bryan Heidorn, Daryl Lafferty, Peter Lang, Gil Nelson, Deborah Paul, Nahil Sobh, William Ulate, Kimberly Watson, Qianjin Zhang. To the reviewers and iSchools organizers, we want to thank you for your encouragement, your interest in our topics and especially for your considered, thoughtful input.

Paul, D., & Heidorn, P. B. (2013). Augmenting optical character recognition (OCR) for improved digitization: Strategies to access scientific data in natural history collections. *iConference 2013 Proceedings* (pp. 514-518). doi:10.9776/13266

Looking For Insights

The iDigBio Augmenting OCR (A-OCR) Working Group, one of several at iDigBio, formed with initial member suggestions from the broader community present at the October 2011 Kickoff Summit for iDigBio. Our working group seeks participation, collaboration and collective knowledge from a wider audience to address key issues in OCR use and structuring and correction of OCR output obtained from museum specimen labels as outlined below. We believe input from and collaboration with Information Scientists, Computer Scientists and Data Analysts is essential if we are to succeed in using OCR effectively for speeding up digitization, ensuring better quality data and effectively disseminating the information where it is needed.

Natural history collections and in particular biological and paleontological collections are composed of samples from that natural world and are meant to represent the world in a number of ways. They are not just collections of objects such as fossils, plants, insects, fish, birds, mammals or microbes but also collections of rich information about those objects that may include such specific details as habitat, elevation, soil type, host species data, associated species present, and even species not present (i.e., absence data). While perhaps more common in the past, many collectors also continue to use a field notebook for recording exquisite details about each collecting event (Canfield, 2012).

Inaccessible data. One difficulty is that many of these specimens were collected long before there were computer systems to track any of this associated information about these objects. Without a digital record, only a few individuals may ever manage direct access to any of this hidden data by going to the source museum in person. It is estimated that less than 10% of United States museum specimens have any online accessible record (NIBA, 2010). The NSF ADBC program was designed to help address this problem.

The OCR process is just one element of a longer process. In a typical workflow, a team decides which part of a collection should be digitized and then pulls these specimens from their storage cabinets and carries them to an imaging center. Here workers image the specimen and any label information. Current state-of-the-art OCR is well-suited to some common museum label and data types but not so successful with other types (Figure 1) where handwriting,spotted and yellowing paper, uneven text and other non-text objects confuse OCR.

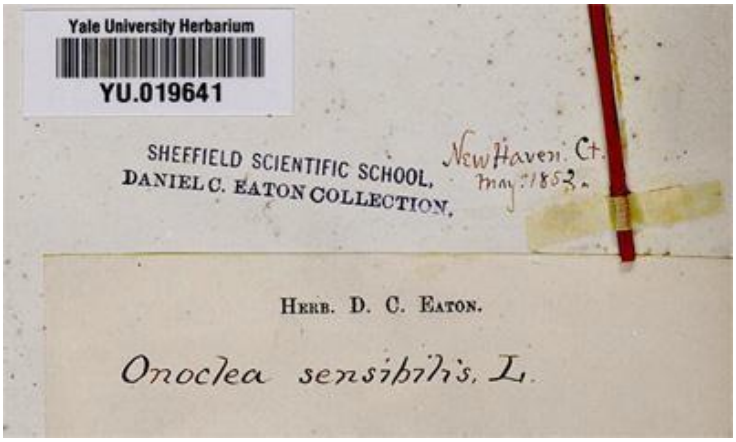

Figure 1. Museum Specimen Label not suitable for current OCR. Yale University Herbarium. Used with permission.

Parsing. Some types of specimens such as some insects may have only a serial number pinned under the specimen. With current technology, OCR will likely not be cost effective for these and project staff may simply type in the numbers or use voice recognition. But, additional information about the specimen is often found in field notebooks and grey literature that are potential targets for OCR. Herbarium specimens too, among other specimen types, frequently have a rich collection of information affixed to the mounting material for the specimen. It is logical too, that collections housing a greater percentage of more recently collected material will have a higher percentage of labels where OCR canto

produce usable output. OCR can be applied directly to the image of the specimen or the labels can be cropped and passed to OCR. The label in figure 2 exemplifies the type of information that can be found on one of these labels and shows the kind of label for which OCR can be effectively applied. Next, using the HERBIS / LABELX system (Heidorn, 2008), the resulting parsed and formatted OCR output from OCR of this label is shown.

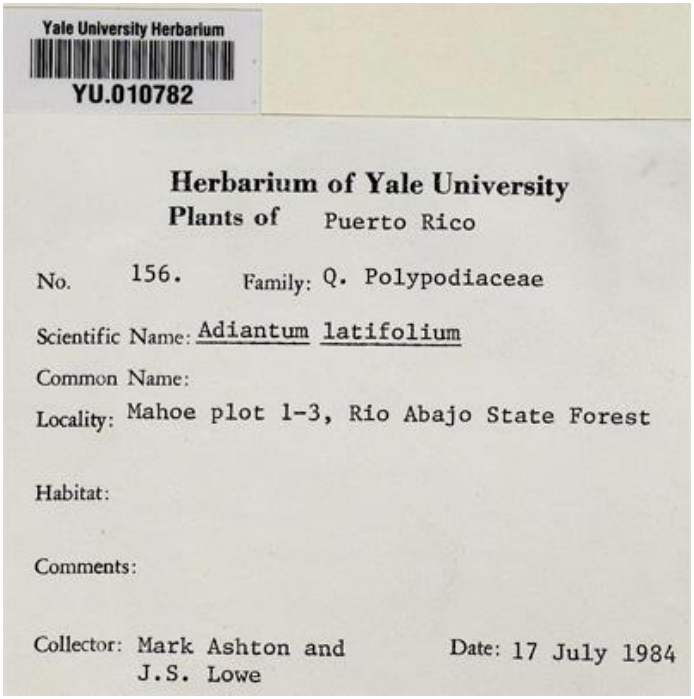

Figure 2. Label suitable for effective OCR. Yale University Herbarium. Used with permission.

Parsed formatted OCR output of label in figure 2 from HERBIS/LABELX system.

```
<?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 above is just one example of the processes our group seeks to improve in our efforts to get specimen data into databases faster. Unfortunately, OCR software is often not utilized to its fullest potential or the OCR output may be sub-par because the specimen label images are not in a format

appropriate for current OCR technology. For example, while handwriting recognition is an active field of research, the success rate on a generalized collection of labels with handwriting from a variety of collectors is near 0. In recent work, Steinke et al. (2010) have been able to digitize a high percentage of the handwriting of the famous scientist, Alexander Von Humboldt because the system is trained for just this one collector; this success is an exception rather than the rule for labels with handwritten content since there are many thousands of collectors. Note the development by Steinke et al. (2011) of better algorithms for recognizing non-text, non-handwriting elements may be of some use to algorithmically remove image data that does not contain text or print.

The situation with type-written labels is brighter. OCR is fairly successful on modern typefaces and well-aligned documents. Museum labels, however, are not in a standardized font or layout. Loose typewriter pinions seem to have been the norm for decades in museums meaning letters are not organized in straight lines, ribbon quality resulted in incomplete characters and age has led to paper and print fading. Consequently, the OCR quality on some collections of type-written labels can be marginal. The situation is exasperated by the fact that scientific vocabulary is not included in standard OCR dictionaries. All these issues contribute to the need for careful research and development in OCR.

The text strings that are the output of OCR need to be parsed into individual elements and placed into standardized formats for ingestion into databases or the semantic web. Much work is needed on parsing algorithms to facilitate getting OCR output mapped programmatically to current data standards for fast, automated information extraction and conversion into machine readable format (Heidorn, 2008; Ruiz et al., 2009; Wei, 2012). Some collections contain large numbers of similarly formatted labels so it is possible to write regular expression parsers to format the information. Most collections, however, have much more variable formats. These issues combined with inherent OCR errors require the application of more flexible approaches such as supervised or unsupervised machine learning approaches.

Broader Goals of the A-OCR WG. Members of our working group put together a current summary wish list of topics to work on, see: http://tinyurl.com/OCRHackathonWishList. The working group is collaborating to put together materials and consensus knowledge to help the community get more from their OCR strategies. No one approach will work for all labels because of the idiosyncratic nature of the collections but one goal of the A-OCR WG is to identify the methods which work best under different conditions. Topics we are gathering material on include:

- known effective practices for getting the most from any OCR software,
- known issues that hinder good (useful) OCR output,
- reporting findings after working with real image data and programmers to improve parsing of OCR output,
- lists of OCR software currently being utilized by the natural history collections community with contact information,
- training and evaluation procedures and data for comparing methods,
- compiling OCR resources such as natural history dictionaries with scientific names, collectors, location, institutions and other information,
- compiling relevant OCR related research in the iDigBio bibliography resource, and
- developing user-interfaces and workflows for human-in-the-loop participation in parsing.

In addition, we also seek to identify opportunities to find and leverage existing tools and technologies that are successful in and out of the biology digitization domain and find opportunities to integrate these tools, or to seek funding for tool development.

Our outreach strategy. The A-OCR WG held its first in-person meeting, October 1 - 2, 2012 at the University of Florida, home of iDigBio. The working group and invited guests initiated plans for our first OCR hackathon being held concurrently with this iConference2013: Scholarship in Action: Data - Innovation - Wisdom. Our first hackathon (http://tinyurl.com/aocrHack) concentrates on parsing and user-interfaces in an effort to establish a baseline of what is currently possible and find more partners outside our usual borders to move forward. In addition to this paper, we put together an iConference2013 Workshop titled, "Help iDigBio Reveal Hidden Data: iDigBio Augmenting OCR Working Group Needs You" to formally introduce the iSchools community to iDigBio, this working group, and the digitization efforts and challenges in the natural history museum collection world. In conjunction, on Friday, the A-OCR WG will present an iConference2013 Alternative Event to report back on the hackathon which is Wednesday and Thursday, February 13 – 14, at the Botanical Research Institute of Texas (BRIT).

Sustainability. As a community service and in order to begin creating a sustainable effort, plans are in place at iDigBio to set up a permanent virtual OCR sandbox available at any time to facilitate OCR

Engine experimentation, and machine language (ML) and Natural Language Processing (NLP) algorithm improvement for natural history collections. This effectively provides a virtual hackathon for anyone interested in trying out new algorithms on a standard set of images, for example.

Conclusion

We are confident that through these combined collaboration and outreach initiatives, new partnerships will evolve leading to improvements in OCR strategies that will positively impact the outcome of this national effort to create a new data resource for everyone. Current research supports this (Haston, 2012; Steinke, 2010) and so we are looking forward to the challenge.

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

Blagoderov, V., Smith, V. S. (2012). Bringing collections out of the dark. In Blagoderov, V., Smith, V. S. (Eds.), *No specimen left behind: mass digitization of natural history collections. ZooKeys* 209, 1–6. doi:10.3897/zookeys.209.3699

Canfield, M. R. (Ed.). (2011). *Field Notes on Science and Nature.* Cambridge, MA: Harvard University Press.

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

Haston, E., Cubey, R., Pullan, M., Atkins, H., Harris, D. J. (2012). Developing integrated workflows for the digitisation of herbarium specimens using a modular and scalable approach. In Blagoderov, V., Smith, V. S. (Eds.), *No specimen left behind: mass digitization of natural history collections. ZooKeys* 209, 93–102. doi:10.3897/zookeys.209.3121

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

Ruiz, M. E., Best, J., Heidorn, P. B., Neill, A., Moen, W. (2009). *Digital Libraries for Biodiversity and Natural History Collections.* Panel Discussion. Sponsors: SIG/DL Retrieved from http://www.asis.org/Conferences/AM09/panels/55.pdf

Steinke, K. H., Gehrke, M., Dzido, R. (2010). Recognition of Humboldt's Handwriting in Complex Surroundings. *Institute of Electrical and Electronics Engineers (IEEE) 12th International Conference on Frontiers in Handwriting Recognition (ICFHR).* Kolkata doi:10.1109/ICFHR.2010.91

Steinke, K. H., Gehrke, M., Dzido, R. (2011). Object Recognition in Herbarium Specimens. In Perner, P., (Ed.), *Advances in Data Mining - Poster and Industry Proceedings. 11th Industrial Conference, ICDM 2011, New York, USA.* (pp.1-16). IBaI Publishing

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